BioMistral-Clinical: A Scalable Approach to Clinical LLMs via Incremental Learning and RAG

Ziwei Chen and **Bernhard Bermeitinger** and **Christina Niklaus**
University of St.Gallen
`ziwei.chen@student.unisg.ch`,
`{bernhard.bermeitinger, christina.niklaus}@unisg.ch`

Abstract

The integration of large language models (LLMs) into clinical medicine represents a major advancement in natural language processing (NLP). We introduce *BioMistral-Clinical 7B*, a clinical LLM built on *BioMistral-7B* (Labrak et al., 2024), designed to support continual learning from unstructured clinical notes for real-world tasks such as clinical decision support. Using the *augmented-clinical notes* dataset provided by Hugging Face (2024), we apply prompt engineering to transform unstructured text into structured JSON capturing key clinical information (symptoms, diagnoses, treatments, outcomes). We employ self-supervised continual learning (SPeCiaL) (Caccia and Pineau, 2021) to achieve efficient incremental training. Evaluation on MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) shows that BioMistral-Clinical 7B improves accuracy on MedMCQA by nearly 10 points (37.4% vs. 28.0%) over the base model, while maintaining comparable performance on MedQA (34.8% vs. 36.5%). Building on this, we propose the *BioMistral-Clinical System*, which integrates Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) to enrich responses with relevant clinical cases retrieved from a structured vector database. The full system enhances clinical reasoning by combining domain-specific adaptation with contextual retrieval.

1 Introduction

Medical natural language processing (NLP) plays a crucial role in improving clinical workflows and supporting healthcare decision-making. From early rule-based systems to modern machine learning approaches, the field has evolved significantly to better handle the complexity and variability of medical data (Fieschi et al., 2003; Sutton et al., 2020).

The emergence of Large Language Models (LLMs), particularly since GPT-3 (Brown et al., 2020b), has further transformed medical NLP by enabling the automation of clinical documentation, diagnostic support, and personalized care (Thirunavukarasu et al., 2023). However, the growing size of these models raises concerns about computational cost, deployment feasibility, and adaptability to clinical-specific language.

To address the limitations of general LLMs in clinical applications, specifically their limited capacity for domain adaptation and their inability to incorporate external knowledge, we propose the *BioMistral-Clinical System*, a lightweight framework based on *BioMistral-7B* (Labrak et al., 2024). We first built *BioMistral-Clinical 7B* through continual learning on structured JSON-formatted clinical records (Caccia and Pineau, 2021). In that way, we improve the domain specificity and clinical knowledge of the BioMistral model. To further improve context-aware response quality, we integrate retrieval-augmented generation (RAG) (Lewis et al., 2020). Thus, we enable real-time access to external medical knowledge, making its responses not only up-to-date but also transparent and verifiable (see Figure1).

Our contributions are as follows: (i) we propose a data preparation scheme that converts long unstructured clinical notes into structured, information-rich JSON representations using chain-of-thought prompting; (ii) we introduce a continual learning approach that updates only a targeted subset of model parameters, enabling efficient clinical knowledge integration while mitigating catastrophic forgetting; (iii) we include a retrieval-augmented layer that grounds model outputs in relevant patient cases, thereby improving response specificity without requiring additional fine-tuning; (iv) we validate the proposed methodology (incremental learning on lightweight models + RAG enhancement) for clinical LLMs through comprehensive qualitative and quantitative evaluations, confirming its utility in this domain; and (v)

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1171–1184
December 20-24, 2025 ©2025 Association for Computational Linguistics

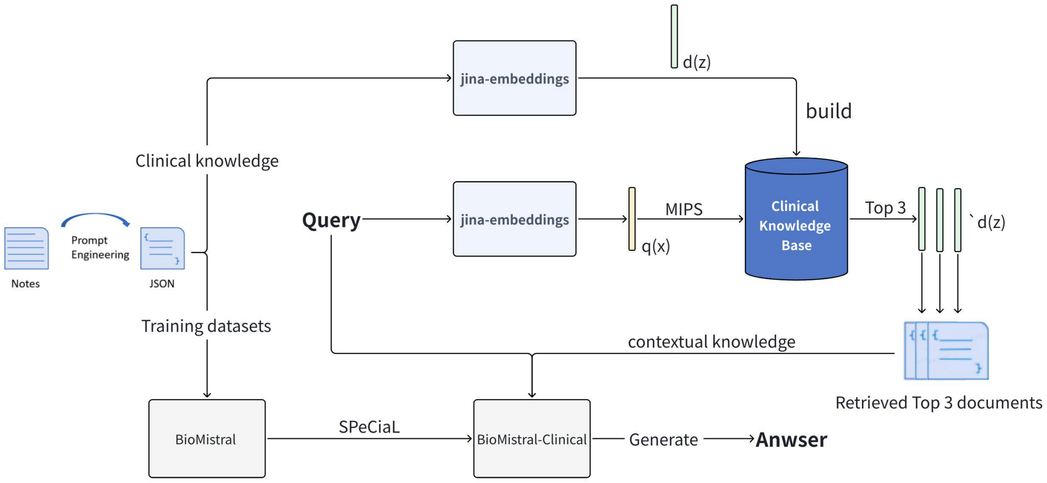


Figure 1: Overview of the *BioMistral-Clinical System*. Unstructured clinical notes are transformed into structured JSON using prompt engineering. These structured records are used to incrementally train the base model (*BioMistral*) via Self-Supervised Training for Continual Learning (SPeCiaL) , producing *BioMistral-Clinical*. The same data is embedded to construct a clinical knowledge base. At inference time, user queries retrieve the top-3 relevant documents via Maximum Inner Product Search (MIPS). These documents, together with the query, are passed to the *BioMistral-Clinical* model to generate the final answer.

we open-source the dataset[1] and model[2]on HuggingFace under the MIT License, to support clinical NLP research.

2 Related Work

2.1 Traditional Approaches for Medical Decision Support

Traditional approaches to prediction of medical outcomes have long struggled with a trade-off between rigid structure and practical usability. Early rule-based Medical Decision Support Systems (MDSS), for instance, offered interpretable if-then logic, but were inflexible and sensitive to data quality (Fieschi et al., 2003). Although subsequent generations, such as Clinical Decision Support Systems (CDSS), improved adherence to clinical guidelines, their widespread adoption was ultimately hampered by persistent usability issues (Sutton et al., 2020). This evolution highlights a critical gap for decision support tools that are clinically robust and highly intuitive, a niche that medical LLMs are poised to fill.

2.2 Development of Medical LLMs

Since the launch of GPT-3 (Brown et al., 2020b), general-purpose LLMs have been applied to clinical domains. Despite their strong language generation capabilities, their lack of domain-specific training raises safety concerns in clinical settings (Korngiebel and Mooney, 2021). These risks highlight the need for models developed specifically for the medical domain.

In response, specialized LLMs such as GatorTron (Yang et al., 2022) and PMC-LLaMA (Wu et al., 2024) were introduced. GatorTron, trained on over 90 billion words (including 82 billion de-identified clinical tokens), significantly improved performance on clinical NLP tasks. PMC-LLaMA incorporated biomedical textbooks and literature, outperforming ChatGPT on QA benchmarks. Later models like MEDITRON-70B (Chen et al., 2023) and OpenBioLLM-70B (Ankit Pal, 2024) further scaled parameters to achieve state-of-the-art performance. However, scaling introduces challenges: high computational cost, limited deployability, and diminishing returns. As an example, MEDITRON-70B improved only 5–8% over its 13B version despite a 4× increase in training expense (Hoffmann et al., 2022; Chen et al., 2023).

2.3 Lightweight Clinical LLMs

Recent research has shifted toward the development of lightweight medical LLMs to reduce computational demands and improve deployability. *BioMistral-7B* (Labrak et al., 2024) achieved 85% of OpenBioLLM-70B performance with only

---

[1] Available at: `https://huggingface.co/datasets/ZiweiChen/BioMistral-Clinical-Knowledge-Base`

[2] Available at: `https://huggingface.co/ZiweiChen/BioMistral-Clinical-7B`

1/10th of the parameters, supporting applications on edge devices. This highlights a trend toward efficiency and task-specific adaptability over pure scale, enabling broader clinical adoption without compromising reliability. However, *BioMistral-7B* was trained primarily on general biomedical corpora, which limits its grasp of real-world clinical language.

Building on this trend, our work integrates lightweight domain adaptation, self-supervised continual learning, and retrieval augmentation into a unified clinical language modeling framework.

3 Methodology

Our approach leverages prompt engineering to structure unannotated clinical data into JSON format, enabling domain-adaptive pretraining via the SPeCiaL framework. We further construct a clinical knowledge base to support real-time document retrieval during inference. Technically, we contribute a lightweight yet domain-specialized model based on *BioMistral-7B*, a training pipeline that supports continual learning, and a hybrid system that integrates retrieval and generation for improved clinical question answering (see Figure 1).

3.1 Datasets

This study utilizes the *Augmented Clinical Notes* dataset provided by Hugging Face (2024) The dataset comprises approximately 30,000 clinical notes sourced from a combination of real-world and synthetic data.

This dataset was originally developed to train MediNote-7B and MediNote-13B, a pair of fine-tuned clinical note generators from the MediTron (Chen et al., 2023) family of LLMs. In this study, we use this dataset for incremental training and build the knowledge base.

The average length of the note is approximately 3000 words (SD = 1,473), with the full range extending from 746 to more than 31,000 words. Each record contains diverse clinical components, such as symptoms, diagnoses, treatment, and patient outcomes. The complexity and extent of these narratives make the dataset highly suitable for building clinical-specific models.

Example Case Summary

> *A 67-year-old patient with metastatic renal cell carcinoma presented with shortness of breath, pleuritic chest pain, and left scapular discomfort. Imaging revealed a gastropleural fistula, multiple metastases, and atelectasis. Treatment involved gastrostomy and chest tubes, endoscopic suturing, and laparoscopic fistula repair. The patient recovered successfully and was discharged to rehabilitation, with complete tube removal after four months and no complications during follow-up.*

`Summarized by ChatGPT`

This example shows that this dataset is able to reflect complex clinical cases and diverse treatment trajectories. However, the notes lack structure because they are narrative texts filled with redundant or irrelevant information. It is inefficient to use such records directly for training, especially for lightweight models.

3.2 Model Selection: *BioMistral-7B*

In this study, we selected the publicly available *BioMistral-7B* (Labrak et al., 2024) model[3] as our base due to its demonstrated efficacy in processing complex biomedical and clinical texts. The model is built on *Mistral 7B* Instruct v0.1[4] and was designed to incorporate instructions and fine-tune across a range of tasks efficiently. It has been extensively pre-trained on the PubMed Central corpus (Jin et al., 2019), providing it with a strong foundation in the medical literature, which aligns well with the goals of this research in the medical field.

One of the key reasons for selecting *BioMistral-7B* is its lightweight architecture. In contrast to the growing trend toward large-scale LLMs, there is increasing interest in developing more efficient, lightweight models that can deliver similar performance benefits without requiring excessive computational power (Tian et al., 2024). With only 1/10th the parameters of OpenBioLLM-70B, it can reach 85% of its accuracy (Labrak et al., 2024). This lightweight design makes it an ideal candidate for further refinement and specialized clinical applications, especially when hardware resources are limited. However, there remains room for improvement, particularly in terms of its adaptability to real-world clinical settings.

3.3 Prompt Engineering

To build high-quality data for incremental pretraining, we transform unstructured clinical notes into a structured format. We employ prompt engineering (PE) with a general-purpose LLM to convert

---

[3] Available at: `https://huggingface.co/BioMistral/BioMistral-7B`

[4] Available at: `https://huggingface.co/mistralai/Mistral-7B-v0.1`

each note into a standardized JSON object. This process captures essential clinical elements (e.g., main complaint, history, findings, diagnosis, and treatment) within distinct subfields, thereby enhancing the data's granularity and relational clarity for model training.

We conducted initial experiments where we used Zero-Shot and Few-Shot Prompting (Brown et al., 2020a). Although Few-Shot prompting improved format consistency, we found that it often failed to capture fine-grained details across clinical subfields. To improve consistency, we adopted Chain-of-Thought (CoT) prompting (Wei et al., 2022), which guides the model to reason through subtasks step by step. The CoT template includes role definition, field explanations, rules, and multiple input-output examples (see Figure in the appendix A). GPT-4 was selected for large-scale annotation to balance performance and cost.

Annotating 30K notes with GPT-4 consumed ~100M tokens and took ~40 hours. The structured output averaged 1,300 tokens, significantly shorter than the original input length of approximately 3,000 tokens. This reduction also led to a decrease in variability, with the standard deviation dropping from 1,473 to 477. Figures 5 in the appendix B show the length distribution after transformation. In addition, an example of the JSON output is shown in the appendix G.

3.4 Incremental Learning

3.4.1 Data Preparation and Tokenization

Structured JSON data from Section 3.3 distilled through prompt engineering captures key patient information such as medical history, diagnoses, treatments, and results. We set a maximum input length of 1,024 tokens, covering 99.67% of all entries (see Figure 6).

Tokenization was performed using the original *BioMistral-7B* tokenizer to ensure vocabulary alignment. The data set was split 80/20 into training and validation sets, the latter being used to monitor generalization and prevent overfitting.

3.4.2 SPeCiaL: Self-Supervised Training for Continual Learning

We adopt the continual self-supervised learning framework proposed by Caccia and Pineau (2021), where a pretrained LLM is incrementally updated via autoregressive learning on new domain-specific data. This strategy enables knowledge integration without catastrophic forgetting, avoiding the need for full retraining.

Self-supervised learning predicts future tokens from the past context using causal masking, and unlabeled data to refine model representations. This is especially beneficial in clinical domains where labeled data is scarce.

Our approach uses *BioMistral-7B*, a 32-layer, 7.2B parameter decoder-only transformer. To retain basic biomedical knowledge while adapting to clinical notes, we freeze the bottom 20 layers of the model while fine-tuning the top 12 layers for efficient continuous adaptation.

3.4.3 Training Strategy

Training was conducted for 5 epochs using an autoregressive objective on our structured clinical inputs. Each batch contained 16 samples, fully utilizing an NVIDIA A800 80G GPU. We used Hugging Face Transformers to load the base model and tokenizer with default settings. The total training time was 37 hours.

Training loss steadily decreased, indicating successful learning. Validation loss initially dropped but began rising after 11,000 steps, signaling overfitting. We thus selected the 10,000-step checkpoint as the final model based on optimal validation performance.

The resulting model, *BioMistral-Clinical 7B*, inherits the general biomedical knowledge of *BioMistral-7B* while being specialized for structured clinical narratives. All metrics were tracked via Weights & Biases (wandb), as shown in Figures 7 and 8 in the appendix. All reported results are based on a single training run without multiple seed averaging.

3.5 Supervised Fine-Tuning

To adapt the model for multiple-choice clinical question-answering tasks, we perform supervised fine-tuning (SFT) using the low-rank adaptation method (LoRA) (Hu et al., 2022), which enables parameter-efficient learning by injecting trainable low-rank matrices into pretrained weights while keeping the original model frozen.

We trained on the multiple-choice datasets from MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022). A maximum sequence length of 256 tokens was selected to cover the entire question-answer pairs, and training was performed over 5 epochs. For this parameter-efficient tuning, we employed LoRA with a rank of $r = 8$ and a scaling factor of $\alpha = 16$. This resulted in approximately 20 million trainable parameters, which is only about 0.5% of the complete model.

Training was conducted using batch sizes suitable for single-GPU setups. On an NVIDIA A800 80G GPU, training completed in approximately 7 hours, compared to 17 hours on an A10 24G. We save four LoRA-adapted checkpoints for evaluation.

3.6 Retrieval-Augmented Generation

To address limitations in fixed-knowledge language models and improve response specificity, we integrate a Retrieval-Augmented Generation (RAG) framework (Lewis et al., 2020) into the *BioMistral-Clinical System*. This enables the model to dynamically retrieve relevant clinical cases from a structured corpus at inference time, providing real-time contextual grounding for each query.

As shown in Figure 1, our RAG implementation begins by constructing a vector knowledge base. Each clinical note z in our corpus is encoded into a 1024-dimensional embedding $d(z)$ using the lightweight `jinaai/jina-embeddings-v3` model (Sturua et al., 2024). At inference, a given query x is similarly embedded, and we use Maximum Inner Product Search (MIPS) to retrieve the top $K = 3$ most relevant documents $\{z_1, z_2, z_3\}$. These documents are then prepended to the original query to form a context-rich prompt for *BioMistral-Clinical*, enhancing response specificity and consistency without requiring model retraining.

4 Evaluation

We present a comprehensive evaluation of our approach via an ablation study on retrieval augmentation. We first evaluate the base model *BioMistral-Clinical 7B* with incremental integration of clinical knowledge (see Section 4.1), then evaluate the complete *BioMistral-Clinical System*, which extends the base model with a RAG mechanism for contextual retrieval in real time (see Section 4.2).

4.1 Evaluation of BioMistral-Clinical 7B

To obtain a comprehensive view of the behavior of our model, we evaluated *BioMistral-Clinical 7B* in two dimensions: (1) qualitative case comparisons that simulate real-world clinical scenarios (see Section 4.1.1 and Section 4.1.2), and (2) quantitative evaluation using generated clinical queries and public multiple-choice medical benchmarks (see Section 4.1.3 and Section 4.1.4).

4.1.1 Clinical Scenario Comparison

To qualitatively assess clinical reasoning ability, we generated a synthetic dataset of 100 basic clinical questions using GPT-4 (Achiam et al., 2023). The questions were created by prompting the model with: *"Please generate 100 common clinical questions."* These queries reflect general diagnostic and treatment scenarios commonly encountered in clinical practice. We used this dataset to evaluate and compare the responses of *BioMistral-7B* and *BioMistral-Clinical 7B*. Examples of these queries are provided in Appendix 2.

We selected a representative clinical scenario query. Below, we compare the responses generated by the original *BioMistral-7B* and our proposed *BioMistral-Clinical 7B*.

Example Query: *How to treat severe obesity?*

BioMistral-7B:
Severe obesity can be treated using 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:
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.*

As illustrated above, while both models outline general treatment strategies, *BioMistral-7B* provides only broad recommendations. In contrast, *BioMistral-Clinical 7B* generates a more structured and specific response, suggesting different types of interventions and team-based care. A more detailed clinical example is provided in Appendix D for further illustration.

4.1.2 Manual Content Analysis

Our manual analysis of the 100 generated clinical question-answer pairs revealed that *BioMistral-Clinical 7B* consistently generates qualitatively superior responses compared to baseline. Specifically, we observed the following patterns of improvement throughout the reviewed samples:

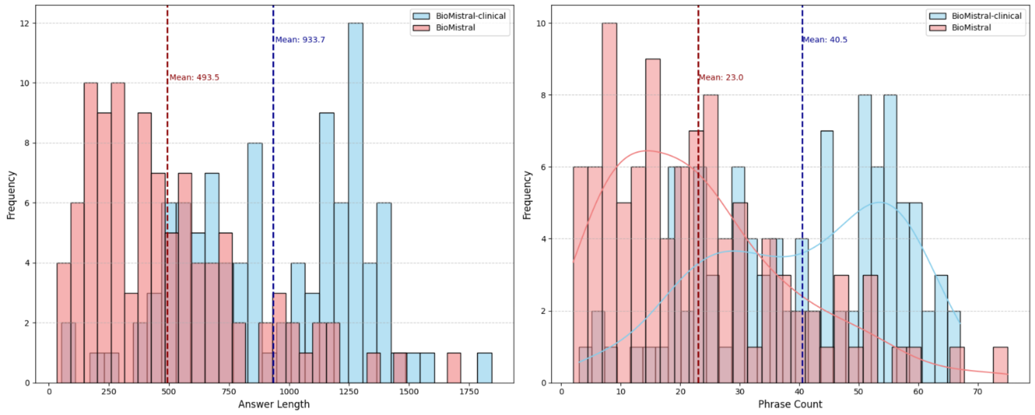


Figure 2: Comparison of answer length distribution and phrase count distribution between *BioMistral-7B* and *BioMistral-Clinical 7B* across 100 clinical queries. Left: histogram of answer lengths; Right: histogram of phrase unit counts.

1. It demonstrated significantly greater pharmacological specificity, providing concrete recommendations (e.g., “Amlodipine 5 mg once daily, along with...hydrochlorothiazide 25 mg...”) instead of generic categories.
2. It generated more comprehensive treatment plans by integrating diagnostics, lifestyle modifications, and multistep medication regimens.
3. It exhibited stronger contextual reasoning, tailoring recommendations to available patient data, and crucially, refusing to speculate when such information was absent.

Interestingly, our analysis revealed that *BioMistral-Clinical 7B* shows significant clinical caution. In our manual review of 100 samples, this behavior occurred in 9 cases where prompts lacked sufficient patient-specific details. In these instances, instead of generating a potentially unsafe recommendation, the model would refuse to answer and state its reasoning. For example, it would respond:

“The patient’s medical history, specific diagnostic findings, and diagnosis are not provided in the available medical information. Therefore, it is not possible to provide a comprehensive treatment plan or postoperative course.”

This behavior reflects an increased sensitivity to the clinical context, likely resulting from the domain-specific pretraining process, where rich contextual cues are often present. Notably, when such contextual information is provided, the model is able to generate more precise and individualized recommendations.

4.1.3 Quantitative Analysis

To complement the results of the qualitative analysis, we performed quantitative analyses of these 100 queries and the corresponding answers.

Answer Length Analysis We used answer length (in characters) as a proxy for response richness. As shown in Figure 2 (left), *BioMistral-Clinical 7B* produced significantly longer responses (mean: 933.69) than *BioMistral-7B* (mean: 493.46). This can demonstrate that the new model can suggest more in terms of response length.

Phrase Count Analysis To further evaluate the depth and granularity of the models’ responses, we performed a phrase-level analysis using spaCy’s `en_core_web_sm` model to extract noun phrases from each answer. These phrases, which often encapsulate key clinical concepts and entities, reveal the richness of information in the responses. On average, *BioMistral-Clinical 7B* generated responses containing 40.5 noun phrases, while *BioMistral-7B* produced an average of 23 noun phrases per response. These findings are visualized in Figure 2 (right).

Vocabulary Usage Analysis We also analyzed word frequencies, focusing on the top common terms from each model’s output, to assess the clinical importance and semantic depth of their output. As shown in Figure 3, *BioMistral-Clinical 7B* relies heavily on medical domain-specific terminology (“treatment,” “diagnosis,” “symptoms,” “medication”), reflecting a strong focus on clinical themes. In contrast, *BioMistral-7B* features frequent terms like “treatment,” “therapy,” and “patient” alongside more generic words (“avoid,” “help,” “cause”), in-

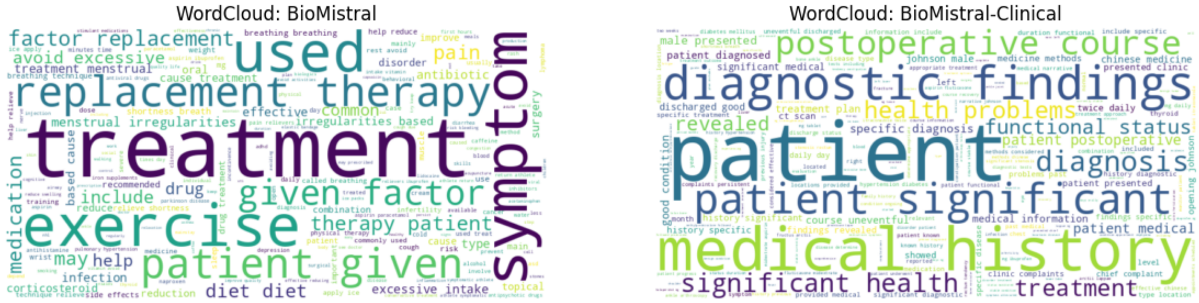


Figure 3: Word cloud graph in responses generated by *BioMistral-7B* and *BioMistral-Clinical 7B* across 100 clinical queries.

dicating a less concentrated emphasis on medical expertise.

4.1.4 Public Dataset Comparison

To assess the capabilities of *BioMistral-Clinical 7B*, we benchmark it against its base model, *BioMistral-7B*, and other prominent medical LLMs on two public multiple-choice question answering datasets: MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022). We evaluate the models in two settings: zero-shot and after supervised fine-tuning (SFT). All results, including comparisons to external models, are presented in Table 1. Example questions are available in Appendix C.1 and C.2.

Zero-Shot Performance We first evaluate the models without any task-specific fine-tuning. For MedQA, we use a 10% random sample (1,273 questions) of the English test set. For MedMCQA, we use a 1,000-question sample from its surgery subset to align with our model's clinical focus. The results reveal a clear trade-off: on the broader MedQA dataset, the base model slightly outperforms the clinical version (36.5% vs. 34.8%). In contrast, in the specialized MedMCQA surgery subset, *BioMistral-Clinical 7B* shows a substantial advantage over the base model, with a +9.4% improvement in accuracy (37. 4% vs 28. 0%). This highlights that continued pretraining on clinical data enhances performance on domain-specific tasks, albeit at the cost of a marginal drop in broader medical knowledge.

Post-SFT Performance Next, we fine-tune both models on the training splits of each dataset. For MedQA, we use its complete training set. For MedMCQA, to maintain focus on clinical specialization, we fine-tune both models exclusively on their surgery subset, which contains 16,862 questions. As shown in Table 1, SFT increases performance for both models. In MedQA, the performance gap narrows, and the base model maintains a slight edge (43. 5% vs 42. 3%). However, in the MedMCQA surgery subset, the advantage of *BioMistral-Clinical 7B* persists and remains significant, achieving 47. 7% precision compared to 41. 2% for the base model (+6.5%). This indicates that while SFT is broadly effective, the specialized pre-training of *BioMistral-Clinical 7B* provides a stronger foundation for adaptation to specific clinical domains.

Comparison with External Models When compared to other 7B parameter models, *BioMistral-Clinical 7B* shows competitive performance in the general MedQA dataset and establishes a clear advantage in its specialized clinical domain. Specifically, while performing on the same level as *MediTron-7B* (Chen et al., 2023) in MedQA (42.3% vs 41.6%), our model outperforms it in the MedMCQA surgery subset (47.7% vs 41.3%). It also substantially surpasses *PMC-LLaMA 7B* (Wu et al., 2024) on both benchmarks. Although there is a performance gap to the much larger *GPT-3.5 Turbo* (OpenAI, 2023), these results establish *BioMistral-Clinical 7B* as a leading open-source model for tasks that require deep knowledge of the clinical domain.

4.2 Evaluation of BioMistral-Clinical System

4.2.1 Retrieval Accuracy Evaluation

To quantify the quality of document retrieval, we conducted a systematic manual evaluation on a set of 100 queries sampled from our test scenarios (Section 4.1.1). We assessed the relevance of the top-3 retrieved documents for each query based on two criteria: (i) **Clinical Relevance:** Does the retrieved document address the same primary medical condition or clinical question as the query?

Model	MedQA	MedMCQA	MedQA-SFT	MedMCQA-SFT
BioMistral-Clinical 7B (Clinical)	34.8%	37.4%	42.3%	47.7%
BioMistral-7B (Base)	36.5%	28.0%	43.5%	41.2%
Improvement (Clinical - Base)	-1.7%	+9.4%	-1.2%	+6.5%
*MediTron-7B**	-	-	41.6%	41.3%
*PMC-LLaMA 7B**	-	-	25.5%	26.6%
*GPT-3.5 Turbo 1106**	57.71%	53.79%	-	-

Table 1: Performance comparison on public medical QA datasets. Results are accuracy (%). *Scores are from (Labrak et al., 2024). Positive improvements are in blue, negative in red.

(Binary: Yes/No); and (ii) **Usefulness:** Does the document contain specific, actionable information (e.g., diagnostic criteria, treatment options, symptoms) that could contribute to a comprehensive answer? (Binary: Yes/No)

Our analysis of the 300 retrieved documents (100 × 3 documents) revealed that a document was clinically relevant in 82% of cases. Furthermore, for 77.6% of the queries, at least one of the top-3 retrieved documents was rated as "useful," providing substantive information for the generation stage. All judgments were made using the network-connected Gemini 2.5 Pro model (Google DeepMind, 2025). Detailed examples of successful and failed retrievals are provided in Appendix 2.

4.2.2 Qualitative Evaluation of the RAG System

We conducted a qualitative analysis to assess the impact of Retrieval-Augmented Generation (RAG) on the final output. We compared the responses from the RAG-enhanced *BioMistral-Clinical 7B* against the baseline *BioMistral-7B* (without RAG) for the same 100 queries. Our manual assessment focused on identifying common patterns of improvement, primarily in three areas: (i) **Specificity:** Providing concrete details (e.g., naming specific drugs or therapies) versus generic statements; (ii) **Completeness:** Offering a more comprehensive overview that covers multiple aspects of the query (e.g., causes, symptoms, and treatments); and (iii) **Factual Grounding:** Reducing factual inaccuracies or hallucinations by grounding the response in retrieved evidence.

Across our analysis, we observed that RAG significantly enhanced the responses, particularly for complex or less common conditions. In approximately 70% of such cases, the RAG-enhanced model provided greater specificity and completeness. For instance, the example below illustrates how RAG enriches a response about a complex neurological disorder.

Query: *How to manage multiple sclerosis?*

BioMistral-7B (No RAG):

> There is **no cure** for multiple sclerosis (MS). Treatment includes **medications**, **lifestyle changes**, and **rehabilitation**.

BioMistral-Clinical (RAG-Enhanced):

> Multiple sclerosis (MS) is a **chronic autoimmune disorder** affecting the **central nervous system**, causing symptoms such as fatigue, vision problems, and cognitive issues. There is **no cure** that exists, the management involves **medications**, **physical therapy**, and supportive therapies tailored to **individual patient needs**.

This systematic analysis highlights that the RAG component is crucial for moving beyond the model's pre-trained knowledge, enabling it to generate more accurate, detailed, and clinically useful answers.

5 Conclusion

This paper presents the *BioMistral-Clinical System*, a lightweight clinical language model that combines incremental learning with RAG. We show that prompt-based annotation can efficiently structure unlabeled clinical notes, enabling self-supervised learning. A compact embedding-based knowledge base facilitates semantic retrieval and, when integrated via RAG, enhances response quality by grounding outputs in real-world clinical content. These results highlight the potential of lightweight, structured, and retrieval-enhanced approaches to build practical clinical LLMs with minimal annotation and computational cost.

6 Limitations

Although this study uses continual training to develop an improved model, *BioMistral-Clinical 7B*, which achieves a nearly 10% gain on MedMCQA compared to the base *BioMistral-7B*. In addition, it incorporates RAG to form the *BioMistral-Clinical System*, allowing dynamic access to external medical knowledge for more context-aware responses. Despite these advancements, it is important to recognize several limitations that frame the scope and generalizability of the findings.

First, although the study provides strong evidence for the feasibility of self-supervised incremental learning in structured clinical data, the underlying dataset itself is inherently imperfect. The JSON-formatted entries generated through PE used in training are still based on a limited corpus of clinical narratives. In particular, many of these narratives were synthetically generated rather than transcribed from actual patient-doctor interactions (Hugging Face, 2024). As a result, they may lack the linguistic diversity, contextual nuance, and clinical irregularities found in real-world settings. This constraint implies that certain specialties, rare conditions, or edge cases may be underrepresented, thereby limiting the breadth and balance of the knowledge captured by the model.

Secondly, the clinical knowledge base constructed for the RAG module, although it was designed systematically and empirically validated, must recognize that the scope remains narrow. The coverage of the knowledge base is still limited compared to the diversity of real-world clinical practice. The current findings validate the methodology for transforming structured data via embedding and retrieval, but do not yet reflect the behavior of the system at scale.

Third, a minor trade-off in general-domain performance was observed after incremental clinical learning, as evidenced by a slight decrease in MedQA accuracy (34.8% compared to 36.5% for the base model). Although this does not detract from the clinical improvements of the model, it does highlight the importance of maintaining domain balance during specialization. We still lack experimental proof of the same approach for other domains, such as finance or education, and it is not possible to draw generalizations.

Finally, due to the absence of publicly available benchmark datasets that map detailed patient symptoms to case-level retrieval outcomes, the evaluation of the RAG pipeline in this study relies in part on qualitative analysis. Qualitative and manual evaluations such as the ones conducted in this study are often considered the gold standard for assessing retrieval relevance. However, the absence of standardized, large-scale benchmarks limits the ability to perform consistent and reproducible quantitative comparisons. Therefore, the development of such benchmarks remains an important direction for future work in the field.

7 Ethical Considerations

In addition to the promise in clinical applications, ethical issues must be thoughtfully addressed. First and foremost, patient privacy and data protection are critical. Since the system handles sensitive clinical content, compliance with HIPAA (1996) and GDPR (2016) standards is essential to prevent misuse and maintain trust (Yadav et al., 2023).

Second, the system is designed as a clinical decision support tool and cannot substitute professional medical judgment. Generative models exhibit variability in their outputs in different runs and inputs (Zhu et al., 2024), and it is the responsibility of healthcare professionals to critically interpret the suggestions of the model. AI-generated recommendations should complement, not replace human expertise. The responsible incorporation of such systems necessitates transparency, protection features, and ongoing emphasis on human oversight.

All datasets used in this study, including MedQA and MedMCQA, are publicly released and licensed for research use. In accordance with their original terms, all derived artifacts, including structured JSON output, the trained BioMistral-Clinical 7B model, and the retrieval corpus, are used solely for academic research purposes. These artifacts are not intended for clinical deployment or commercial use.

Lastly, we disclose that AI assistants (e.g., ChatGPT and Deepseek) were used to support writing and code development. However, all scientific design decisions, data processing steps, and model training were conducted solely by the authors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Accountability Act. 1996. Health insurance portability and accountability act of 1996. *Public law*, 104:191.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020b. Language Models are Few-Shot Learners.

Lucas Caccia and Joelle Pineau. 2021. Special: Self-supervised pretraining for continual learning. In *International Workshop on Continual Semi-Supervised Learning*, pages 91–103. Springer.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

M Fieschi, J-C Dufour, P Staccini, J Gouvernet, and O Bouhaddou. 2003. Medical decision support systems: old dilemmas and new paradigms? *Methods of information in medicine*, 42(03):190–198.

Google DeepMind. 2025. Gemini 2.5 pro. https://deepmind.google/technologies/gemini/. Accessed: 2025-07-24.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hugging Face. 2024. Augmented-clinical-notes dataset. https://huggingface.co/datasets/augmented-clinical-notes.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Diane M. Korngiebel and Sean D. Mooney. 2021. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *npj Digital Medicine*, 4(1).

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-Tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. *Neural Information Processing Systems*, 33:9459–9474.

OpenAI. 2023. GPT-3.5 Turbo. https://platform.openai.com/docs/models/gpt-3-5. Model version: gpt-3.5-turbo-1106.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

General Data Protection Regulation. 2016. General data protection regulation (gdpr)–official legal text. *Gen Data Prot Regul*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V Chawla. 2024. Tinyllm: Learning a small student from multiple large language models. *arXiv e-prints*, pages arXiv–2402.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances* *in neural information processing systems*, 35:24824–24837.

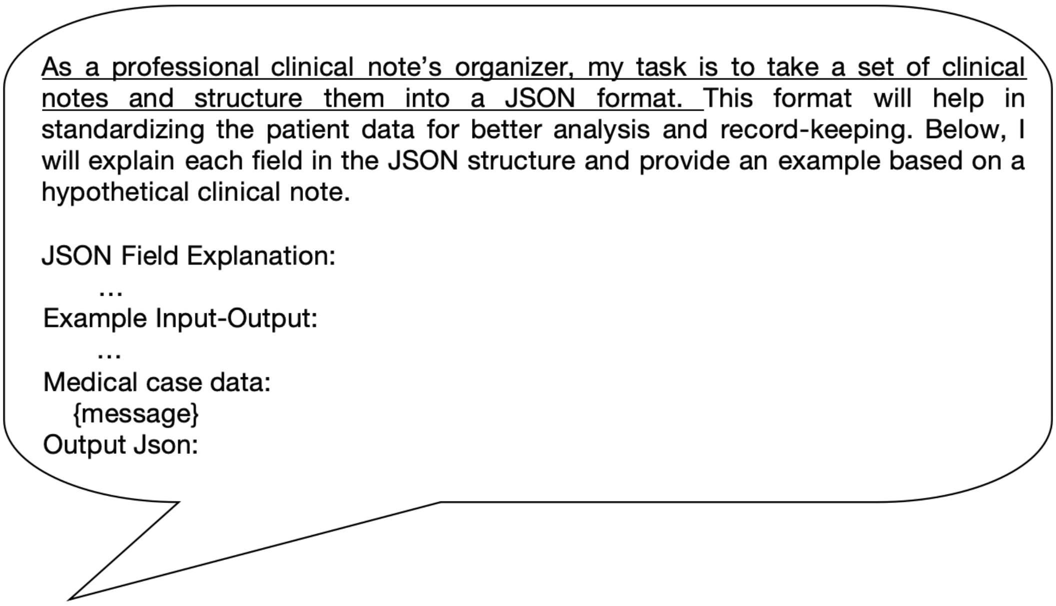
Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Neel Yadav, Saumya Pandey, Amit Gupta, Pankhuri Dudani, Somesh Gupta, and Krithika Rangarajan. 2023. Data privacy in healthcare: In the era of artificial intelligence. *Indian Dermatology Online Journal*, 14(6):788–792.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, and 1 others. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.

Lingxuan Zhu, Weiming Mou, Chenglin Hong, Tao Yang, Yancheng Lai, Chang Qi, Anqi Lin, Jian Zhang, and Peng Luo. 2024. The evaluation of generative ai should include repetition to assess stability. *JMIR mHealth and uHealth*, 12:e57978.

A The Chain-of-Thought prompt template.



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 Input-Output:
...
Medical case data:
{message}
Output Json:

Figure 4: The Chain-of-Thought prompt template.

B Data distribution

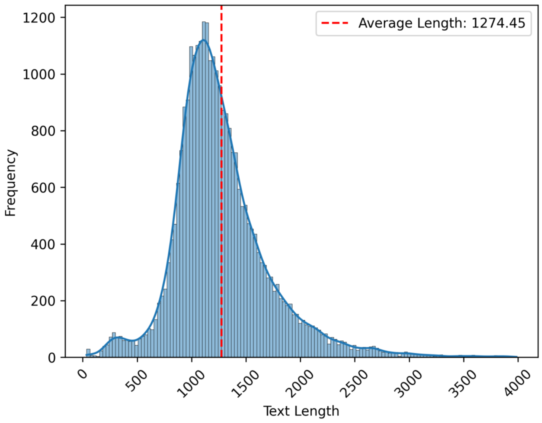


Figure 5: Text length distribution after PE.

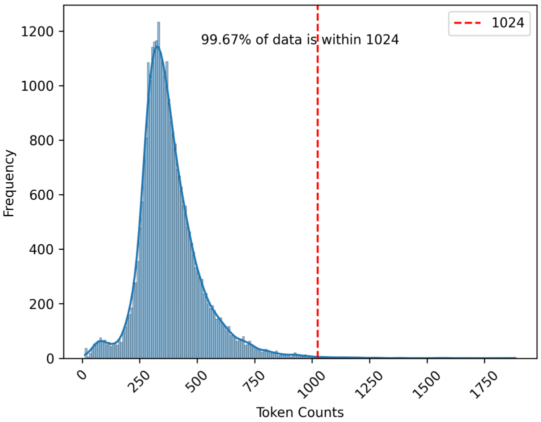


Figure 6: Token counts distribution after tokenization.

C Example Question from Public QA Benchmark

C.1 Example Question from MedQA

Question: A 3-month-old baby died suddenly at night while asleep. His mother noticed that he had died only after she awoke in the morning. No cause of death was determined based on the autopsy. Which of the following precautions could have prevented the death of the baby?

Options:

A. Placing the infant in a supine position on a firm mattress while sleeping

B. Routine postnatal electrocardiogram (ECG)

C. Keeping the infant covered and maintaining a high room temperature

D. Application of a device to maintain the sleeping position

E. Avoiding pacifier use during sleep

Correct Answer: A

C.2 Example Question from MedMCQA-Surgery

Question: Delayed union of fracture of a bone following a surgical treatment may be due to:

Options:

A. Infection

B. Inadequate circulation

C. Inadequate mobilization

D. All of the above

Correct Answer: D

D Example: Refined Scenario – Patient-Specific Case

Detailed Question: *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:

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:

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 dietitian*** *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 problems. The patient will be* ***closely monitored*** *with* ***regular follow-up appointments*** *to ensure the treatment is effective and to address any concerns.*

E Training and Validation loss over steps

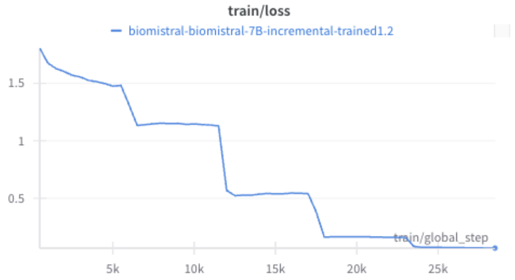


Figure 7: Training loss over steps.

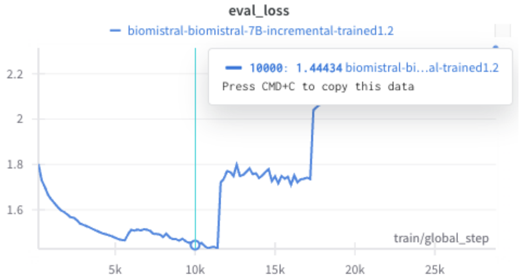


Figure 8: Validation loss over steps.

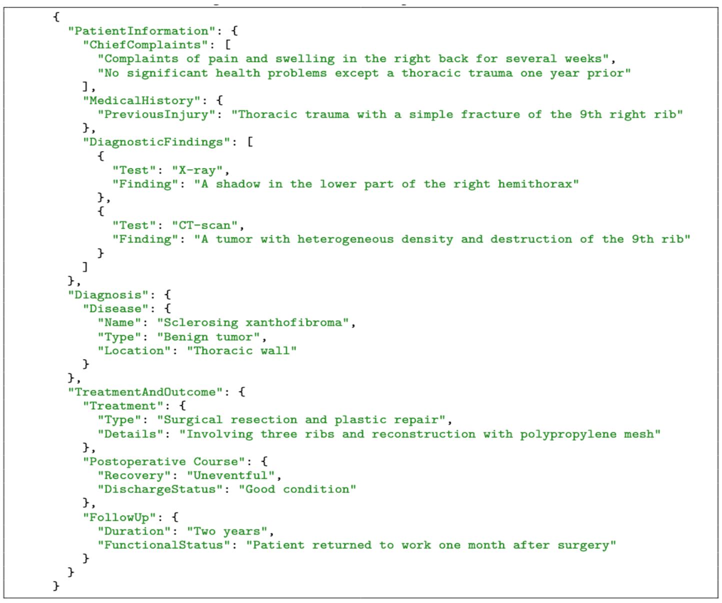
F Evaluation of Retrieval Examples Table

Table 2: Evaluation of the top-3 retrieved documents for 10 sample clinical queries. We assess each retrieved document's diagnosis based on **Clinical Relevance (R)** and **Usefulness (U)** for answering the query. **R**: Is the diagnosis clinically related to the query? (✓ Yes / ✗ No). **U**: Does the diagnosis provide specific, useful information for answering the query? (✓ Yes / ✗ No).

Clinical Query	**Top-1 Document**			**Top-2 Document**			**Top-3 Document**		
	Diagnosis	**R**	**U**	**Diagnosis**	**R**	**U**	**Diagnosis**	**R**	**U**
Manage type 2 diabetes?	Type 2 diabetes	✓	✓	Reactive Hypoglycemia	✓	✓	Type 2 diabetes	✓	✓
Treatment for hypertension?	Resistant hypertension	✓	✓	Renal artery stenosis	✓	✓	Renal artery stenosis	✓	✓
Treatments for asthma?	Uncontrolled asthma	✓	✓	Allergic bronchopulm. asp.	✓	✓	Severe asthma	✓	✓
Manage high cholesterol?	Hyperlipidemia	✓	✓	*None*	✗	✗	Thoracic aortic aneurysm	✗	✗
Treat a skin rash?	Mycosis Fungoides	✓	✓	*None*	✓	✓	Fixed drug eruption	✓	✓
Manage arthritis pain?	*None*	✓	✗	Rheumatoid Arthritis	✓	✓	Gouty arthritis	✓	✓
Deal with constipation?	Functional Constipation	✓	✓	Severe constipation	✓	✓	Rectal stenosis	✓	✓
Treatments for bronchitis?	Traction bronchiectasis	✓	✓	Chronic bronchiectasis	✓	✓	Bronchiectasis	✓	✓
Manage eczema?	Eczematous dermatitis	✓	✓	Mycosis Fungoides	✓	✗	Erythema nodosum	✗	✗
Manage varicose veins?	Symptomatic varices	✓	✓	Varicose Veins	✓	✓	Large gastric varix	✓	✓

Note: 'None' indicates a missing 'diagnosis' field in the retrieved document. For these cases, relevance (R) and usefulness (U) were assessed by manually inspecting the full document content (e.g., symptoms, treatment plan).

G Example 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"
    }
  }
}
```

Figure 9: Structured JSON output from the prompt-engineered dataset.