BioMistral-Clinical System: Enhancing Clinical Knowledge in Large Language Models through Incremental Learning Methods and Retrieval-Augmented Generation

Anonymous ACL submission

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, we apply prompt engineering to transform unstructured text into structured JSON capturing key clinical information (symptoms, diagnoses, treatments, outcomes). This enables efficient incremental training via self-supervised continual learning (SPeCiaL) (Caccia and Pineau, 2021). 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). (See Figure 1)

The emergence of Large Language Models (LLMs), particularly since GPT-3 (Brown et al., 2020b), has further transformed medical NLP by enabling 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, which are namely limited domain adaptation and inability to incorporate external knowledge. We propose the BioMistral-Clinical system, a lightweight framework based on BioMistral-7B (Labrak et al., 2024). We first obtain BioMistral-Clinical 7B through continual learning on structured JSON-formatted clinical records (Caccia and Pineau, 2021), enhancing domain specificity and clinical reasoning. To further improve context-aware response quality, we integrate retrieval-augmented generation (RAG) (Lewis et al., 2020), enabling real-time access to external medical knowledge.

This system offers a practical approach to deploying LLMs in clinical environments, addressing key barriers to real-world applicability and supporting real-time decision-making in healthcare.

2 Related Work

2.1 Traditional Rule-Based and Probabilistic Methods for Medical Decision Support

Traditionally, the prediction of medical outcomes relied on manual analysis and early rule-based Medical Decision Support Systems (MDSS) (Fieschi et al., 2003), which applied expert-defined if-then rules. Although these systems were interpretable, they lacked flexibility and were sensitive to data quality. To improve diagnostic accuracy, probabilistic models such as Bayesian networks were introduced to capture uncertainty and encode expert knowledge (Magrini et al., 2018). Before LLMs, Clinical Decision Support Systems (CDSS) im-

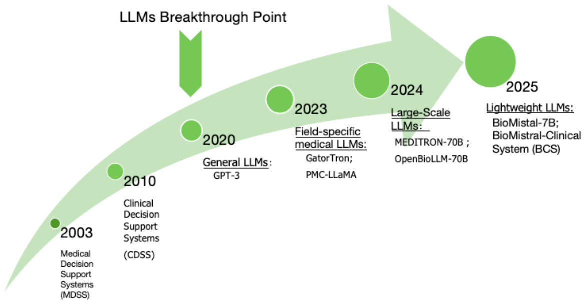

Figure 1: The development of NLP methods in the medical field.

proved care quality and guideline adherence (Sutton et al., 2020), but adoption was limited by usability issues, highlighting the need for more intuitive tools like medical LLMs.

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 Current Research Directions: Lightweight Clinical LLM

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

Current large language models (LLMs) often struggle to adapt to clinical-specific contexts and cannot dynamically incorporate up-to-date external knowledge. To address this gap, we propose the BioMistral-Clinical System, which combines self-supervised continual learning on structured clinical notes with RAG to enhance clinical reasoning and response specificity (Figure 2). 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.

3.1 Datasets

This study utilizes the *Augmented Clinical Notes* dataset curated by Hugging Face (2024)[1]. The dataset comprises approximately 30,000 triplets of clinical notes sourced from a combination of real-world and synthetic data.

This dataset was originally developed to train MediNote-7B[2] and MediNote-13B[3], 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 to train and construct the BioMistral-Clinical System's knowledge base.

[1]Available at: https://huggingface.co/datasets/AGBonnet/augmented-clinical-notes
[2]https://huggingface.co/AGBonnet/medinote-7b
[3]https://huggingface.co/AGBonnet/medinote-13b

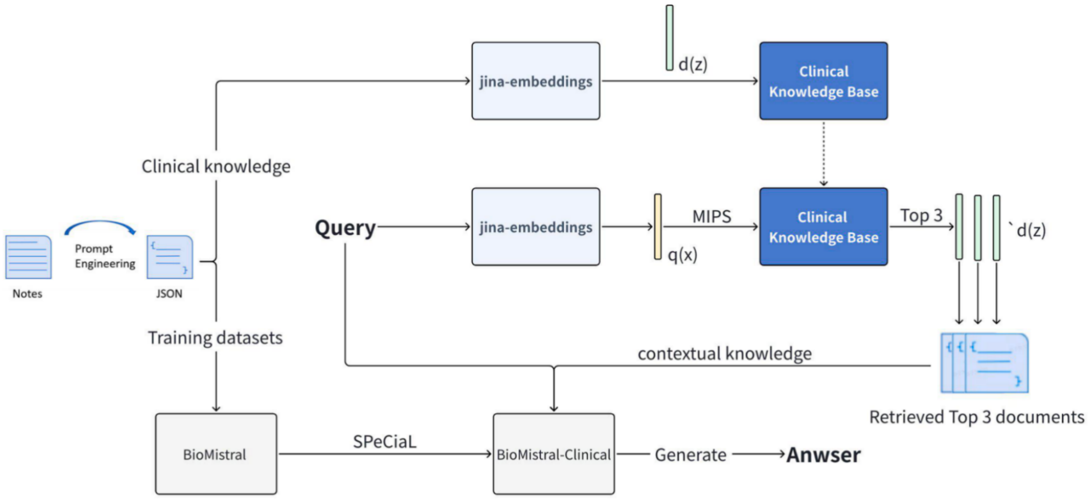

Figure 2: Overview of the BioMistral-Clinical System. Unstructured clinical notes are transformed into structured JSON using prompt engineering. These structured records are used to incrementally train the base model (BioMistral) via Self-Supervised Training for Continual Learning (SPeCiaL) , producing 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.

The average note length is approximately 3,000 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, the BioMistral-7B (Labrak et al., 2024) model was selected as the base model 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 efficiently incorporate instructions and fine-tune across a range of tasks. It has been extensively pretrained 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.

Although BioMistral-7B excels in its general medical knowledge, evidenced by its strong performance on 10 established English medical question-answering tasks (Labrak et al., 2024), there remains room for improvement, particularly in terms of its adaptability to real-world clinical settings.

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

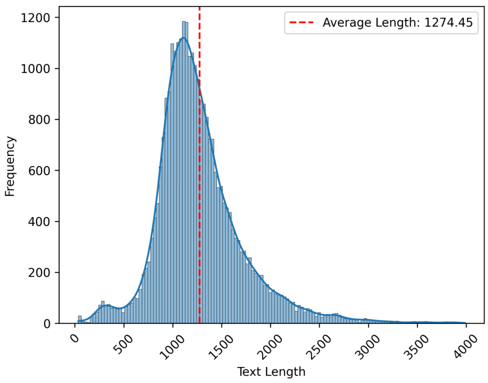

Figure 3: Text length distribution after PE.

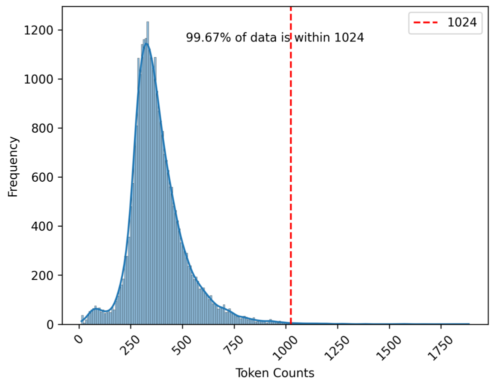

Figure 4: Token counts distribution after tokenization.

3.3 Prompt Engineering

After selecting the dataset and base model, the next key step is to build high-quality training data for incremental learning. This study aims to transform unstructured clinical notes into a structured format to enhance granularity and relational clarity.

To address this, we propose using prompt engineering (PE) with general-purpose LLMs to convert notes into a standardized JSON format. This structured representation captures essential elements, such as the main complaint, history, findings, diagnosis, treatment, and outcome, organized into clear subfields. As shown in Zhou et al. (2023), general models such as GPT-3.5 and GPT-4 are increasingly being used to generate training data, especially when manual labeling is costly.

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 A in the appendix). GPT-3.5 Turbo was selected for large-scale annotation to balance performance and cost.

Annotating 30K notes with GPT-3.5 Turbo consumed 100 million 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 3 show the length distribution after transformation. In addition, an example of the JSON output is shown in the appendix E.

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 4).

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 (see Figure 5 for

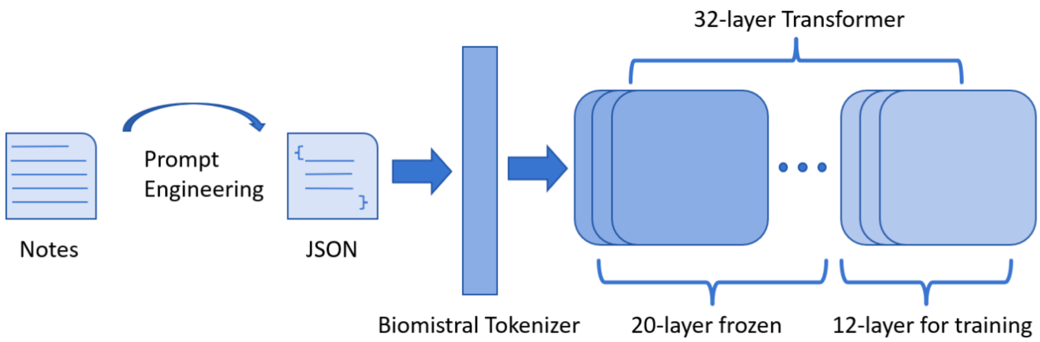

Figure 5: The SPeCiaL training pipeline used in BioMistral. Unstructured clinical notes are first converted into structured JSON via prompt engineering. The inputs are then tokenized using the BioMistral tokenizer and fed into a 32-layer decoder-only transformer. To preserve core biomedical knowledge while adapting to clinical-specific data, the bottom 20 layers are frozen and only the top 12 layers are fine-tuned, resulting in approximately 2 billion trainable parameters.

architecture).

3.5 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*[5], 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 10 and 11 in the appendix. All reported results are based on a single training run without multiple seed averaging.

3.6 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.

[5] URL removed for anonymity.

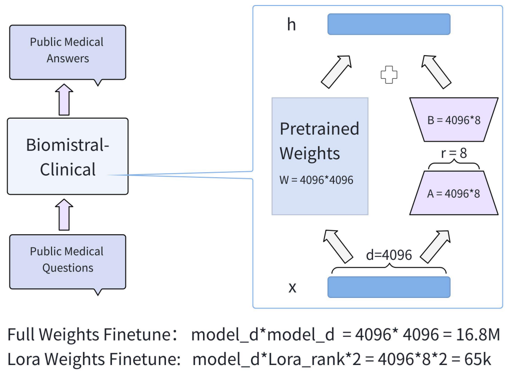

Figure 6: The LoRA structure setting for BioMistral-Clinical

3.6.1 Low-Rank Adaptation

We configure LoRA with rank $r = 8$ and scaling factor $\alpha = 16$, introducing approximately 20M trainable parameters, about 0.5% of the full model (which for BioMistral-Clinical 7B would require updating a matrix $W \in \mathbb{R}^{4096\times4096}$). Specifically, each weight update is represented by two matrices:

$$A \in \mathbb{R}^{r\times d}, B \in \mathbb{R}^{d\times r}, \text{with } d = 4096, r = 8$$

These amount to a total of $2 \times 4096 \times 8 = 65{,}536$ parameters per injection point (see Figure 6). This structure maintains adaptation capacity while significantly reducing computational overhead, enabling efficient fine-tuning for downstream tasks.

3.6.2 Training Specifications

SFT is conducted on multiple-choice datasets for evaluation. Given the structure of MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) (see Section 4.1.2 for details), a maximum sequence length of 256 tokens was selected to cover the entire question-answer pairs. Training is performed over 5 epochs using batch sizes suitable for single-GPU setups. On an NVIDIA A800 80G GPU, training completes in approximately 7 hours, compared to 17 hours on an A10 24G. We save four LoRA-adapted checkpoints for evaluation. This parameter-efficient tuning approach facilitates rapid specialization with minimal resource demands.

3.7 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. In our use case, RAG 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 2, we first construct a clinical knowledge base by embedding structured notes using the lightweight `jinaai/jina-embeddings-v3` model (Sturua et al., 2024). Each document z is encoded into a dense vector $d(z) \in \mathbb{R}^{1024}$, forming the retrieval index.

Algorithm 1 Retrieve Top-3 Relevant Cases

Input: Clinical query q; embedding model E; knowledge base $K = \{(z_i, e_i)\}_{i=1}^{N}$

Output: Top-3 retrieved case texts $Z = \{z_{i_1}, z_{i_2}, z_{i_3}\}$

Compute query embedding: $e_q \leftarrow E(q)$

foreach (z_i, e_i) *in* K **do**

  Compute similarity score: $s_i \leftarrow$ Similarity(e_q, e_i)

Select top-3 indices: $i_1, i_2, i_3 \leftarrow$ argsort$(\{s_i\})[:3]$

return $\{z_1, z_2, z_3\}$

When a new query x is submitted, this is also embedded using the same embedding model to get a query vector $q(x) \in \mathbb{R}^{1024}$ with the same dimensions. The system then performs a Maximum Inner Product Search (MIPS) between $q(x)$ and the document vectors $d(z)$ to retrieve the most appropriate clinical entries. Next, the top $K = 3$ documents are retrieved. The algorithm for retrieving relevant top-3 cases is shown in the pseudocode 1.

These top-3 documents $\{z_1, z_2, z_3\}$ are then concatenated with the original query into a unified prompt, which is passed into the BioMistral-Clinical model. This setup enhances answer specificity and consistency without the need for model retraining, making it suitable for real-world clinical applications.

4 Evaluation

We present a comprehensive evaluation of our proposed approach. The evaluation is structured as an ablation study to assess the impact of retrieval augmentation. First, we evaluate the base model, *BioMistral-Clinical 7B*, after incremental clinical knowledge integration. Then, we assess the full *BioMistral-Clinical System*, which extends the base model with a RAG mechanism for real-time contextual retrieval.

4.1 Evaluation of BioMistral-Clinical 7B

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

4.1.1 Clinical Scenario Comparison

To qualitatively assess clinical reasoning ability, 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.*

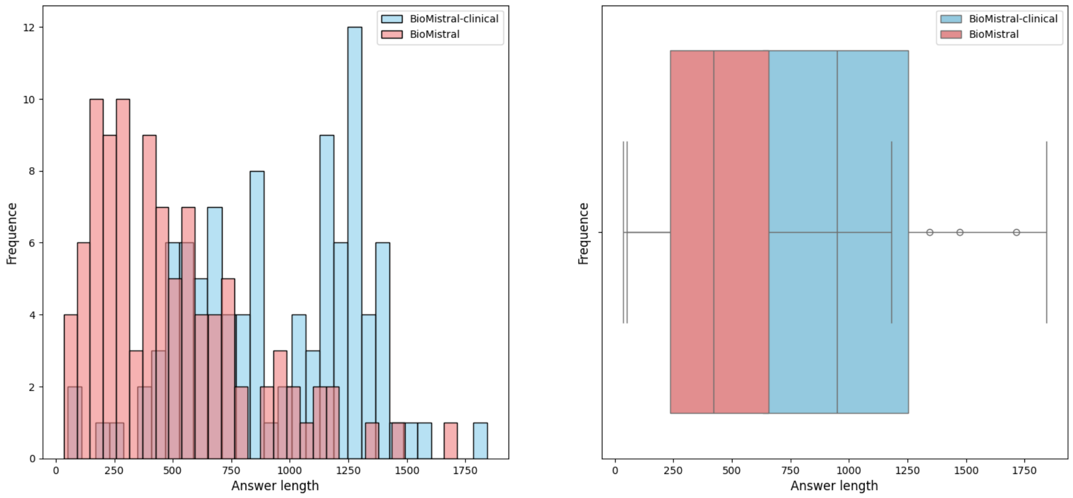

Figure 7: Comparison of answer length distribution between BioMistral-7B and BioMistral-Clinical 7B across 100 clinical queries. Left: histogram; Right: box plot.

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 C for further illustration.

To complement our qualitative findings, 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 and their corresponding retrieval outputs are provided in Appendix D.

We used answer length (in characters) as a proxy for response richness. As shown in Figure 7, BioMistral-Clinical 7B produced significantly longer responses (mean: 933.69) than BioMistral-7B (mean: 493.46). This suggests enhanced informativeness following clinical fine-tuning.

4.1.2 Quantitative Analysis

To complement qualitative assessments, we quantitatively benchmarked both models using two publicly available medical multiple-choice datasets: MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022).

MedQA The MedQA dataset contains 12,723 multiple-choice questions in English. We randomly selected 10% of the data (1,273 questions) as a test set to evaluate both models. BioMistral-7B achieved an accuracy of 36.5%, while BioMistral-Clinical 7B achieved 34.8%. This minor performance drop suggests a trade-off between clinical specialization and general-domain medical reasoning, though the difference is marginal. An example of a test question is provided in Appendix B.1.

MedMCQA-Surgery The MedMCQA dataset contains 194,000 multiple-choice questions spanning various medical domains. For this study, we extracted the surgery-related subset—*MedMCQA-Surgery*—which includes 16,862 questions. A sample of 1,000 questions was used for evaluation. BioMistral-Clinical 7B significantly outperformed the base model (37.4% vs. 28.0%), indicating that continued training on clinical data improves domain-specific reasoning in surgical contexts. An example of an evaluation question is provided in Appendix B.2.

Post-SFT Performance After SFT on the full training sets, both models exhibited improved accuracy. On MedQA, the performance gap narrowed: BioMistral-7B reached 43.5%, while BioMistral-Clinical 7B closely followed with 42.3%. In contrast, on the MedMCQA-Surgery subset, BioMistral-Clinical 7B showed a more pronounced gain, achieving 47.7% compared to 41.2% for the base model. These results indicate that SFT enhances both general and domain-specific performance, with BioMistral-Clinical 7B benefit-

Model	MedQA	MedMCQA	MedQA-SFT	MedMCQA-SFT
BioMistral-7B	36.5%	28.0%	43.5%	41.2%
BioMistral-Clinical 7B	34.8%	37.4%	42.3%	47.7%
Improvement (Clinical - Base)	-1.7%	+9.4%	-1.2%	+6.5%

Table 1: Performance comparison of BioMistral Clinical and original BioMistral models on public medical question answering datasets. Results are presented as accuracy percentages.

ing more in specialized clinical reasoning.

Summary Across multiple experimental setups, BioMistral-Clinical 7B demonstrates stronger clinical reasoning, improved task performance, and richer responses. These findings confirm that incremental learning of domain-specific data improves medical LLM capabilities without substantially compromising general domain knowledge.

4.2 Evaluation of BioMistral-Clinical System

4.2.1 Retrieval Accuracy Evaluation

To evaluate the quality of document retrieval in the *BioMistral-Clinical System*, we first analyzed whether the top-3 documents retrieved for each query are clinically relevant and contextually aligned with the user question. The retrieval module uses the `jinaai / Jina-embeddings-v3` model, a lightweight and efficient embedding model optimized for semantic search tasks. The embeddings were used to construct a vector-based clinical knowledge base, now publicly available[6].

We evaluated the clinical relevance of diagnoses retrieved for 100 queries (from section 4.1.1) using manual inspection. Results showed that the top-3 retrieved cases had high semantic and clinical alignment with query intent, indicating the retrieval module's effectiveness. Detailed examples are provided in Table D in the Appendix.

4.2.2 Qualitative Evaluation with RAG System

We qualitatively evaluated the generation stage of the *BioMistral-Clinical System*, combining top-3 retrieved cases with clinical queries into context-rich prompts (Wei et al., 2022). Manual assessment indicated that the RAG-augmented system consistently outperformed the baseline *BioMistral-7B*, particularly for complex conditions lacking sufficient pre-trained knowledge.

The following example illustrates the practical impact of RAG augmentation when addressing 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**.

Evaluation also highlighted system dependence on knowledge-base coverage. Queries enriched with additional clinical details (symptoms, lab results) significantly improved retrieval and generation quality.

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 unlabelled 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] URL removed for anonymity.

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.

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.

Alessandro Magrini, Davide Luciani, and Federico M Stefanini. 2018. A probabilistic network for the diagnosis of acute cardiopulmonary diseases. *Biometrical journal*, 60(1):174–195.

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*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

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.

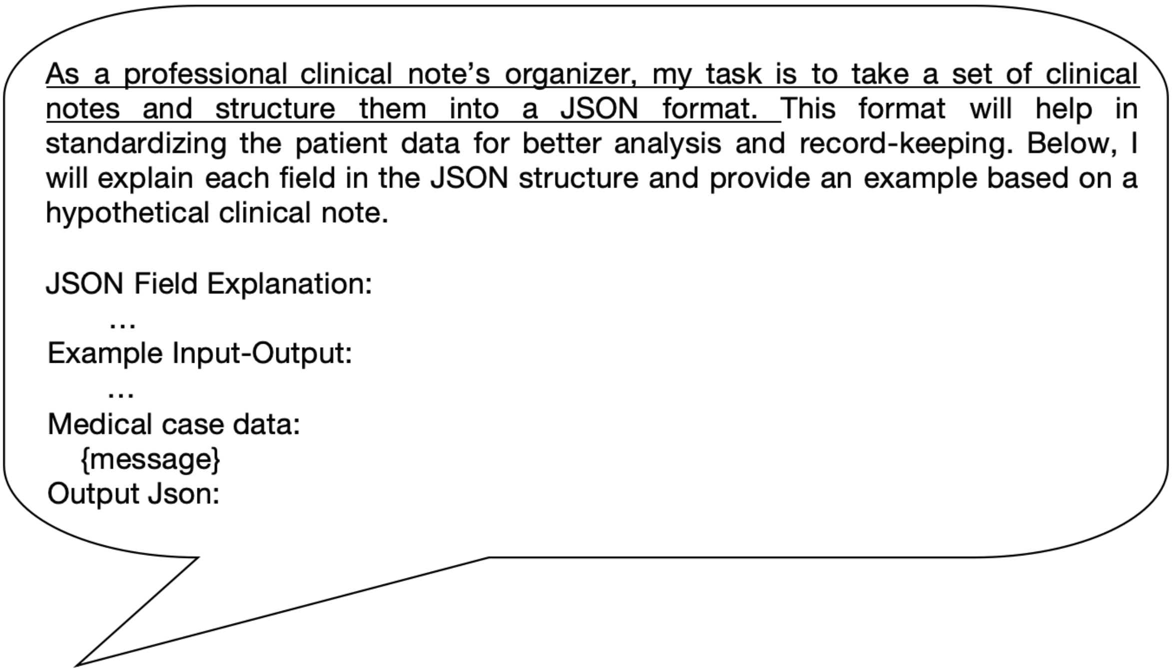

Figure 8: The Chain-of-Thought prompt template.

B Example Question from Public QA Benchmark

B.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

B.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

C 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.*

D Retrieval Examples Table

Clinical Query	**Top-3 Retrieved Diagnoses**
How to manage type 2 diabetes?	Type 2 diabetes, Reactive Hypoglycemia, Type 2 diabetes
What's the treatment for hypertension?	Resistant hypertension, Renal artery stenosis, Renal artery stenosis
What are the treatments for asthma?	Uncontrolled asthma, Allergic bronchopulmonary aspergillosis, Severe asthma
How to manage high cholesterol?	Hyperlipidemia, None, Thoracic aortic aneurysm with ARSA
How to treat a skin rash?	Mycosis Fungoides, None, Fixed drug eruption (FDE)
How to manage arthritis pain?	None, Rheumatoid Arthritis, Gouty arthritis with symptomatic meniscal tears
How to deal with constipation?	Functional Constipation, Severe constipation, Rectal stenosis
What are the treatments for bronchitis?	Traction bronchiectasis, Chronic bronchiectasis, Bronchiectasis
How to manage eczema?	Eczematous dermatitis, Mycosis Fungoides, Erythema nodosum migrans
How to manage varicose veins?	Symptomatic varices, Varicose Veins, Large gastric varix

Note: The value 'None' indicates that the diagnosis field was missing in the retrieved case note, not that the document itself was irrelevant. Other sections of the same document (e.g., symptoms, treatment) may still be contextually aligned with the query.

Table 2: Sample results of diagnosis sections retrieved for 10 clinical questions. Each row shows the top-3 diagnoses retrieved by the RAG system for the given query.

E Example JSON Output


```json
{
  "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.

F Training and Validation loss over steps

Figure 10: Training loss over steps.

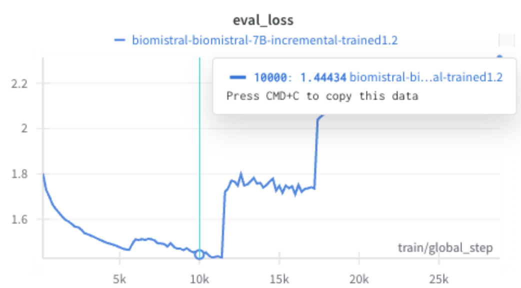

Figure 11: Validation loss over steps.