



Universiteit
Leiden
The Netherlands

Bachelor Data Science and Artificial Intelligence

Enhancing Medical Dialogue Systems:
Fine-Tuning Large Language Models for
More Accurate Diagnoses

Guus Oomen
s3690849

First supervisor: Zhaochun Ren
Second supervisor: Jujia Zhao

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/07/2025

Abstract

In this thesis, I fine-tuned a domain-specific Large Language Model (BioMistral) to improve the accuracy of medical recommendations and create more natural, helpful conversations in a clinical dialogue setting. The motivation for this work stems from the limitations of general-purpose language models in healthcare, where diagnostic accuracy and contextual sensitivity are critical. Generic LLMs often lack domain-specific knowledge, making their output unreliable for medical use.

To address this, I used the NoteChat dataset to fine-tune the BioMistral model using the LoRA method. The fine-tuning process involved formatting the data in a conversational format tuned to instruction and conducting training on the ALICE supercomputing cluster. After training, I benchmarked the resulting model against both the base BioMistral model and the original Mistral model, evaluating performance across human-like and semantic-based metrics using both general and diagnosis-focused test subsets.

The fine-tuned model outperformed both baselines in helpfulness, answer quality, and diagnostic accuracy when clinical data were available. It also demonstrated more consistent and robust performance in terms of semantic similarity.

These findings support the conclusion that fine-tuning LLMs with realistic medical dialogue improves their utility in clinical settings. While biomedical pre-training provides a solid foundation, task-specific fine-tuning is essential for achieving higher-quality, context-aware responses.

Contents

1	Introduction	1
2	Related work	3
2.1	Large language models (LLMs)	3
2.1.1	Pre-training of LLMs	3
2.1.2	Prompt training	3
2.1.3	Fine tuning LLMs	4
2.2	LLMs in medical field	5
2.2.1	Medical datasets and benchmarks for LLMs	5
2.2.2	Medical datasets and benchmarks	5
3	Methodology	7
3.1	NoteChat	7
3.1.1	Preparing the data	7
3.2	Fine-tuning	8
3.2.1	BioMistral	8
3.2.2	LoRA	8
3.2.3	Training procedure	9
3.3	Benchmarking	10
4	Experiments	11
4.1	Evaluating the benchmark	11
5	Results	14
5.1	Analyse of results	14
5.1.1	Humanlike evaluation	14
5.1.2	Sementic similarity	15
5.2	Case study of the results	16
6	Discussion	17
6.1	Analyse of results	17
6.2	Limitations and challenges	17
7	Conclusions and future work	19
	References	22
A	Appendix	23

1 Introduction

LLMs (Large Language Models) are already being used in various industries such as consulting, helpdesks, and customer service. These sectors have seen significant improvements through the reduction of repetitive and unnecessary tasks using LLMs. One sector where this kind of support is urgently needed is healthcare. The healthcare industry suffers from a chronic shortage of workforce, despite being essential for everyone. To help alleviate this burden, LLMs can offer assistance in several ways. For example, they can speed up the diagnosis process or help articulate patient symptoms clearly. These functions can be supported by a medical dialogue system.

A Medical Dialogue System (MDS) provides advice to patients and healthcare professionals in the form of a conversation with an intelligent agent [1]. This agent can issue a diagnosis after gathering sufficient information from the patient. Examples of such systems include Babylon Health[2] and Ada Health [3]. Most MDSs are built using machine learning techniques and natural language processing, but recent advances in large language models have led to their integration into these systems. This development brings greater flexibility and the potential for improved diagnostic accuracy.

In practical applications, a large language model powered medical dialogue system can analyse a patient’s file and suggest possible diagnoses to a doctor or medical professional. Alternatively, such a system can take the form of a chatbot with which patients can interact directly to receive medical advice. The benefits of medical dialogue systems include reducing the workload in healthcare by speeding up the diagnostic process or even allowing the system to make the diagnosis itself. Another major advantage is that these systems can be available 24/7, giving patients constant access to medical support.

This thesis explores how fine-tuning Large Language Models (LLMs) enhances the accuracy of diagnosis, medical recommendations, and other applications in a text-based medical dialogue system.

However, generic LLMs lack domain specialisation for this task, leading to inaccurate or unsafe diagnostic suggestions in medical settings. This poses a high risk of harm: Misdiagnosis can lead to incorrect treatment, delayed care, or serious legal and ethical consequences[4]. Instead of developing an entirely new model from scratch, which is an approach that demands extensive resources. A more cost-effective method is to fine-tune an existing LLM. Fine-tuning enables the adaptation of a generic model to specific medical needs without requiring the training of a full-scale model. This approach uses the strengths of a general LLM while tailoring it to the healthcare domain.

This thesis poses 2 research questions:

1. How does fine-tuning impact the performance of LLMs in medical dialogue tasks compared to their base versions?
2. What are the key challenges and limitations of fine-tuning LLMs for medical applications, and how can they be addressed?

To answer these research questions, this thesis investigates how fine-tuning a domain-specific large language model can improve the quality and accuracy of a medical dialogue system. The focus is on adapting and fine-tuning the BioMistral model using the NoteChat dataset, which consists of synthetic conversations between patients and physicians. In Section 2, related work is reviewed on large language models, fine-tuning techniques, and their applications in the medical field. Section 3 outlines the methodology, including dataset preparation, the fine-tuning process using LoRA, and the benchmarking setup. Section 4 presents experimental results, covering both human-like evaluation and semantic similarity metrics, as well as a case study of model performance. Section 5 discusses the key findings, limitations, and practical implications of this approach. Finally, Section 6 concludes the thesis and outlines potential directions for future research.

2 Related work

2.1 Large language models (LLMs)

2.1.1 Pre-training of LLMs

Pre-training is a method similar to transfer learning, which means taking the knowledge a model already has learnt and applying it to a different task [5]. Building large-language models (LLMs) from scratch is extremely expensive and requires huge amounts of data, computing power, and time. One of the main reasons why there are not many base LLMs available is because it is just too costly to do it without significant funding.

Because of that, most well-known LLMs are based on a few original versions but come in many different forms. For example, GPT(Generative Pre-trained Transformer) from OpenAI has multiple derivatives like ChatGPT and BioGPT. These are examples of pre-trained models. They used the original model GPT as a starting point and then trained it further on specific types of data. For ChatGPT, this means training on dialogue and chat-based content, while BioGPT is trained on biomedical text. This makes each model more accurate and useful in its specific domain [6].

The concept of pre-training LLMs using massive datasets started with models like BERT [7], which was trained using masked language modelling on a large corpus of text. After that, the GPT series (GPT-1, GPT-2, GPT-3) introduced autoregressive pre-training where models predict the next token in a sequence [8, 9]. These pre-trained base models are then fine-tuned or adapted for more specific tasks or domains.

RoBERTa [10] showed that by simply tweaking the pre-training setup. By using more data, longer training times, and removing some training tricks the model performance could be significantly improved. These findings have reinforced the idea that pre-training on a wide and representative dataset provides a strong foundation for adapting models to almost any downstream task.

Although pre-training costs less than building a completely new LLM, it is still a resource-intensive process. You still need a huge dataset, powerful GPUs or TPUs, and a lot of time to do it properly. However, once done, it allows researchers and companies to adapt the model for specialised applications with much less effort than starting from zero.

2.1.2 Prompt training

Using prompts can also improve the accuracy of LLMs. Different types of prompting methods have been explored in research, including zero-shot, few-shot, and chain-of-thought prompting, each with its own strengths.

Zero-shot prompting is the simplest. You just give the LLM your question, without any extra information or examples. This method relies entirely on the model's general knowledge. In the GPT-3 paper by Brown et al. [9], zero-shot prompting was used as a baseline for tasks like translation and question answering, and while the results were decent, they were not as strong as other methods.

Few-shot prompting improves on this by giving the model a few examples before asking the real question. These examples help the LLM understand the pattern or logic behind the task. Brown et al. [9] showed that few-shot prompting led to better performance in tasks like common-sense reasoning, word sense disambiguation, and SAT analogies. For instance, when given a few examples of English-to-French translation, the model could translate new sentences more accurately than with zero-shot.

Chain-of-thought prompting takes it one step further. Here, the prompt includes a step-by-step explanation that shows how the answer was reached. This helps the model not just guess the answer but learn how to reason through the problem. In the paper from Wei et al. [11], they demonstrated that chain-of-thought prompting significantly improved performance on math word problems and logical reasoning tasks. For example, instead of just giving the answer to “If Tom has three apples and gives away one, how many are left?”, the example would walk through the steps: “Tom starts with 3 apples. He gives 1 away. So he has $3 - 1 = 2$ apples left.”

These prompting strategies are important tools for improving LLM performance without needing to retrain the entire model, especially in cases where labelled data is scarce or expensive to collect.

2.1.3 Fine tuning LLMs

Fine-tuning is a core technique for adapting large pre-trained language models to more specific tasks or domains. This involves training a model with less data compared to pre-training. Fine-tuning builds on the existing knowledge of a pre-trained model and updates its weights using smaller, domain-specific datasets. This approach lowers the computational cost and time while increasing the accuracy.

Traditional fine-tuning involves updating all millions, or even billions, of parameters in a model. While effective, this still requires significant resources and a large, labelled dataset. An example of this is GPT-3 which was fine-tuned with human feedback and conversational data to create ChatGPT, optimising it for dialogue-based tasks [12]. Another example is BioGPT, which was fine-tuned using biomedical corpora, enabling it to outperform general models in tasks like medical text generation and classification [13].

A breakthrough in fine-tuning has come from parameter-efficient fine-tuning (PEFT). This method allows only a small subset of the model’s parameters to be updated, making the process more lightweight and efficient. Techniques such as adapter layers and prefix tuning have been used to improve accuracy while drastically reducing the number of parameters that need to change [14].

An important example of PEFT is LoRA (Low-Rank Adaptation). Instead of updating the entire weight matrix, LoRA injects low-rank matrices into the model, which significantly cuts down the number of trainable parameters. This makes it possible to fine-tune large models even on modest hardware setups while still achieving high performance in domain-specific tasks [15].

2.2 LLMs in medical field

2.2.1 Medical datasets and benchmarks for LLMs

In recent years, large language models have been increasingly specialised to tackle the unique challenges of the medical and biomedical domains. This field poses difficulties like complex, highly technical vocabulary, strict demands for factual accuracy, and the need for explainable reasoning that healthcare professionals can trust. General-purpose models like ChatGPT or BERT tend to underperform in this space because they aren't trained on domain-specific data. To improve their usefulness in medical applications, researchers have been pre-training and fine-tuning LLMs on biomedical texts. This improves their ability to handle medical NLP tasks more effectively.

A variety of models have been developed to meet these needs.

BioBERT was one of the first domain-specific adaptations of BERT for biomedical text. It was created by further pre-training BERT on large-scale biomedical corpora like PubMed abstracts and PMC articles [16]. BioBERT significantly outperformed vanilla BERT on biomedical NLP benchmarks.

PubMedBERT improved on BioBERT by being trained from scratch using only biomedical data from PubMed. This approach avoids the noise from general-domain pre-training and leads to even better performance across various biomedical tasks [17].

BioGPT is another example of a domain-specific model, developed by Microsoft. It's based on GPT-2 and pre-trained entirely on PubMed abstracts. BioGPT achieved state-of-the-art performance on several biomedical benchmarks, including the highest score on the PubMedQA dataset at the time of release [13, 18].

Med-PaLM 2 is a more recent and ambitious effort by Google. It's built on a general-purpose LLM and then instruction-tuned and fine-tuned using medical exam-style datasets. Med-PaLM 2 demonstrated performance comparable to that of actual medical professionals, scoring highly on USMLE-style questions[19]. It currently holds the second-highest accuracy on the PubMedQA benchmark.

All these models show how adapting LLMs to the medical domain using pre-training or fine-tuning techniques on the right data can improve their accuracy and reliability for tasks like question answering, document classification, and medical reasoning.

2.2.2 Medical datasets and benchmarks

A big aspect of pre-training and fine-tuning large language models (LLMs) in the medical domain is the quality and relevance of the datasets. These datasets need to be both accurate and precise to significantly improve the model's performance in healthcare applications. Choosing the right dataset is one of the most important steps in developing or adapting an LLM for medical tasks.

One of the most widely used sources of biomedical literature is PubMed and PubMed Central (PMC). These repositories contain millions of abstracts, citations, and full-text articles from biomedical journals. They have been used extensively to pre-train models such as BioBERT [16], PubMedBERT [17], and BioGPT [13]. The depth and breadth of biomedical language in these datasets make them ideal for domain-specific language representation learning.

Another key dataset is MIMIC-III (Medical Information Mart for Intensive Care), which consists of de-identified clinical notes and electronic health record data from over 38,000 adult ICU patients and 49,000 hospital admissions [20]. This dataset includes time-stamped lab results, diagnoses, medications, and free-text notes from healthcare providers. Models such as ClinicalBERT [21] and GatorTron [22] have leveraged MIMIC to enhance their ability to understand clinical narratives and decision-making processes.

To evaluate the performance of medical LLMs, specialised benchmarks have been developed to test their medical knowledge and reasoning. One such benchmark is PubMedQA, which consists of biomedical research article abstracts along with yes/no/maybe questions derived from the content. It evaluates the model’s ability to perform natural language inference and summarise scientific information [18].

Another widely used benchmark is MedQA, which mimics real-world clinical reasoning by presenting multiple-choice questions based on medical licensing exams. It challenges LLMs to apply domain knowledge and make complex inferences similar to human practitioners [23].

These datasets and benchmarks are essential for both training and evaluating medical LLMs, ensuring that they are not only linguistically fluent but also accurate, trustworthy, and capable of reasoning about medical content.

3 Methodology

To be able to answer the research questions the following steps need to be taken. First, a dataset is collected that contains medical dialogues between medical professionals and patients from a publicly available source. The LLM will be fine-tuned with preprocessed data and relevant information. To test the performance of the fine-tuned model, a benchmark is conducted between the base model and the fine-tuned model to be able to compare the results and see if there are any improvements.

3.1 NoteChat

To improve a MDS we use the NoteChat dataset. This is a large scale synthetic dataset set of patient-physician conversations. Developed and introduced in the paper by Wang et al. [24], NoteChat uses a cooperative multi-agent framework of Large Language models. The agents are role-specific where one is a doctor agent and the other a patient agent. These agents are engaged in a structured dialogue guided by actual clinical notes. The framework consists of three stages: planning, role-play and polish, to ensure factual consistency and natural language flow. The resulting dataset contains high-quality, multi-turn dialogues. NoteChat mitigates privacy concerns by generating synthetic conversations from publicly available, HIPAA-compliant clinical note corpora. The dataset includes over 200,000 dialogues, making it one of the most comprehensive resources for fine-tuning a LLM on medical conversations.

The dataset is publicly available from their Hugging Face Hub profile.¹

3.1.1 Preparing the data

The note chat dataset includes more than enough data for a LLM to fine-tune on, even so much that only a subset of the whole dataset is used as training data. The training dataset consists of 850 conversations between a doctor and a patient. To validate the training process a validation set is also needed, for this a subset of 50 conversations are taken where no overlap between training and validation set takes place.

At last, two different test datasets are made where one contains the same sort of data like the training and validation set but without any overlap. Another test subset, referred to as *test_diagnose*, is produced to focus specifically on conversations involving diagnostic conclusions. For a dialogue to be included in this set, it is required that at least one sentence explicitly includes the phrase "diagnosis is" in a doctor's response. This targeted selection allows for a more focused evaluation of the model's diagnostic capabilities.

To convert the raw NoteChat dataset into a format suitable for fine-tuning a Large Language Model, a preprocessing pipeline is implemented. Each dialogue is parsed into a structured list of messages, with assigned roles: system, user, and assistant. This follows the conventional chat format used for instruction-tuned models. A standard system prompt is included at the beginning of each conversation to provide a consistent contextual framing, emulating a medical assistant engaging in clinical dialogue.

¹<https://huggingface.co/datasets/akemiH/NoteChat>

Doctor: Good morning, what brings you here today?

Patient: Good morning, doctor. I have been having some problems with my tics and mood lately

{ "content": "Good morning, what brings you here today?", "role": "assistant" },

{ "content": "Good morning, doctor. I have been having some problems with my tics and mood lately.", "role": "user" }

Figure 1: Example of two sentences in the processed dataset

After preprocessing, the conversations are tokenised into a format compatible with chat-based models and converted into Hugging Face’s DatasetDict structure. The data sets are randomly shuffled between the training, validation and test set (excluding the *test_diagnose* set) using a fixed random seed for reproducibility. Finally, the processed dataset is uploaded to the Hugging Face Hub.²

3.2 Fine-tuning

3.2.1 BioMistral

BioMistral is a specialised large language model based on the Mistral architecture, developed with a focus on biomedical and clinical domains [25]. It has been further pre-trained on large-scale biomedical corpora such as the PubMed Central Open Access subset. This extended pre-training grants the model a foundational understanding of medical terminology, clinical scenarios, and scientific discourse. Despite this, the creators of BioMistral emphasise that it is not safe for real-world medical applications without additional fine-tuning and rigorous validation, particularly in high-stakes environments. In this thesis, BioMistral serves as the base model due to its domain-specific strengths, offering a more relevant starting point compared to general-purpose LLMs.

3.2.2 LoRA

To fine-tune BioMistral efficiently on the NoteChat dataset, Low-Rank Adaptation [15] (LoRA) is used. As talked about in the related work section, LoRA is a parameter-efficient fine-tuning method that allows adaptation of large models by introducing trainable rank-decomposed matrices into specific parts of the model without updating the entire parameter set. In this case, only select attention projection layers are adapted, significantly reducing memory requirements and computational cost during training. This approach makes it feasible to fine-tune large-scale models like BioMistral even on limited hardware, while still achieving notable performance improvements.

In this setup, a rank $r = 8$ is selected for the LoRA adapters, balancing computational efficiency and model performance. This choice is supported by the original LoRA paper [15], which demonstrates that relatively low ranks (such as 8 or 4) can still yield strong results when fine-tuning large models. Figure 2 provides a visual representation of how LoRA works: instead of updating the full pre-trained weight matrix W , two smaller matrices A and B of dimension r are introduced to model the update in a more efficient and modular way. This technique allows for efficient fine-tuning while preserving the stability and knowledge of the base model.

²<https://huggingface.co/datasets/Gustav098/NoteChat-processed>

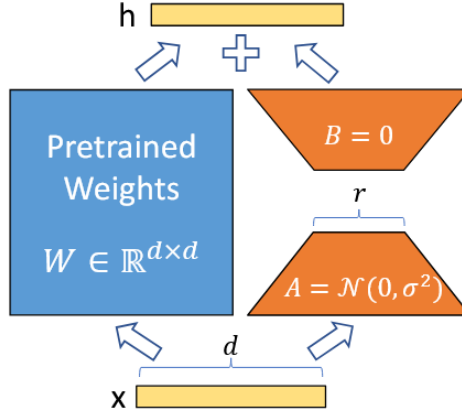


Figure 2: LoRA mechanism: low-rank decomposition of weight updates. Source[15]

The integration of LoRA preserves the core capabilities of the original BioMistral model while allowing it to specialise in the conversational structure and diagnostic context of the NoteChat dataset. By selectively adapting components responsible for language understanding and response generation, the model becomes more adept at handling multi-turn medical conversations. Furthermore, this method supports merging the LoRA adapters back into the base model, resulting in a stand-alone version of the fine-tuned model suitable for inference without dependency on the adaptation layers.

3.2.3 Training procedure

The fine-tuning process is carried out over two epochs using a conversational format where each dialogue is formatted to resemble realistic patient-physician interactions. During training, the model is periodically evaluated using a validation set, and early stopping is applied to avoid over-fitting. A BLEU score is used to monitor the quality of generated responses, with a focus on how closely the model could reproduce coherent and contextually appropriate replies. Special care is taken to format input-output pairs using a consistent role-based structure (system, user, assistant), and truncation is applied to maintain a manageable input length. Tokenization is handled using the base tokenizer of BioMistral, with padding tokens ignored during loss calculation to focus learning only on relevant output sequences.

At the end of training, the LoRA-adapted model is saved both in its modular form and as a merged stand-alone model. Furthermore, the model is then quantised from 16-bit floating point (FP16) to a 4-bit format (Q_4_K_M), making it more efficient to deploy or further fine-tune in downstream tasks. This is done using a publicly available tool from hugging face³. Despite the benefits of quantisation, the size and complexity of the model still requires significant computational resources. For this reason, all training and quantisation are performed on the ALICE high-performance computing cluster, which provides the necessary GPU power to handle the fine-tuning of a model of this scale efficiently.

3.3 Benchmarking

To evaluate the performance of the fine-tuned model, I conduct a benchmarking experiment. This benchmark runs on three versions of the Mistral model: the original Mistral base model, the Bio-Mistral model, and the fine-tuned Bio-Mistral model, which is further trained on the NoteChat dataset. The goal of this setup is to compare the fine-tuned model’s capabilities against both its base version (Bio-Mistral) and the original model (Mistral). This helps show how much the fine-tuning process improves performance on clinical dialogue tasks.

The benchmark consists of four separate evaluations. These are based on two main factors: the prediction type (either generating the next random doctor response or producing a diagnosis from the doctors perspective), and the data condition (whether or not the clinical notes from the NoteChat dataset are included in the conversation). The inclusion of these notes matters because, in some versions of the benchmark, the doctor has access to the clinical notes embedded in the dialogue as system messages. In others, the model must respond or give a diagnosis without access to these notes.

The different variants for the first factor are implemented using the regular test dataset and the *test_diagnose* dataset, as described in the dataset section (see Section 3.1.1). The second factor, whether the clinical notes are included or excluded, is controlled by filtering out all system-role messages from the conversations. Removing these messages results in dialogue-only inputs, where the model must rely entirely on the patient-doctor interaction.

This setup allows us to test how dependent the model is on external clinical information, and whether it can generate appropriate responses or diagnoses based solely on the dialogue, across the three different model versions.

³<https://huggingface.co/spaces/ggml-org/gguf-my-repo>

4 Experiments

4.1 Evaluating the benchmark

Each of the four benchmark variants was evaluated using two different approaches. This is done to get an extended and more complete view of the model's performance across the different test cases.

The first evaluation method used OpenAI's GPT-4o model as a reference evaluator. In this setup, GPT-4o was given both the LLM-generated response and the corresponding ground truth from the dataset. It is then prompted to assess the generated response based on two main criteria. The first one being Helpfulness, and the second one Answer Quality. A prompt for this part looks like this:

Evaluation prompt example

You are a highly experienced medical evaluator AI. You are given a conversation between a patient and a doctor, along with the real response from a doctor and the generated predicted response. Your task is to evaluate the predicted response against the real one and assess the overall medical quality of the response.

You must evaluate three aspects:

1. **Helpfulness:** Was the generated predicted response medically helpful and appropriate to the patient's concern?
2. **Answer Quality:** Considering the prior conversation, is the generated predicted response well-reasoned, complete, and contextually appropriate?

Provide your judgment in JSON format with three fields:

- **helpfulness** (integer score from 0 to 5)
- **answer_quality** (integer score from 0 to 5)
- Also include a short **reason** field explaining your scores. Be precise and clinical in your evaluation.

Conversation: {conv}

Real response: {real}

Predicted response: {pred}

Evaluate the response based on the criteria outlined in the system prompt. Return your judgment in JSON format with "**helpfulness**", "**answer_quality**", and "**reason**".

Here {conv}, {real}, and {pred} are placeholders for the conversation up until the doctors response, the real response of the doctor from the dataset, and the LLM created response based on the conversation. These two criteria generated from this prompt help reflect whether the predicted response is medically useful and appropriate to the patient's concern, and how well-reasoned and contextually appropriate it is. This also takes into account the previous turns in the dialogue, meaning the model is judged not just on a single sentence but on how well it continues the overall conversation.

For the diagnosis-related benchmarks, an additional criterion is added: Diagnostic Accuracy. This measures how close the predicted diagnosis is to the real one, based on the patient dialogue and context. It captures whether the model was actually able to infer the correct clinical conclusion when needed. A prompt for the diagnosis part looks like this:

Evaluation prompt example

You are a highly experienced medical evaluator AI. You are given a conversation between a patient and a doctor, along with the real response from a doctor and the generated predicted response. Your task is to evaluate the predicted response against the real one and assess the overall medical quality of the response.

You must evaluate three aspects:

1. **Diagnostic Accuracy:** How correct is the predicted diagnosis compared to the real one?
2. **Helpfulness:** Was the generated predicted response medically helpful and appropriate to the patient's concern?
3. **Answer Quality:** Considering the prior conversation, is the generated predicted response well-reasoned, complete, and contextually appropriate?

Provide your judgment in JSON format with three fields:

- `diagnostic_accuracy` (integer score from 0 to 5)
- `helpfulness` (integer score from 0 to 5)
- `answer_quality` (integer score from 0 to 5)
- Also include a short `reason` field explaining your scores. Be precise and clinical in your evaluation.

Conversation: {conv}

Real response: {real}

Predicted response: {pred}

Evaluate the response based on the criteria outlined in the system prompt. Return your judgment in JSON format with "`diagnostic_accuracy`", "`helpfulness`", "`answer_quality`", and "`reason`".

Again {conv}, {real}, and {pred} are placeholders which work the same way as the original prompt. Each of these three aspects was rated on a scale from 0 to 5. A score of 0 means the response was completely off, irrelevant, or wrong. A 5 means the response was near-perfect—clinically accurate, helpful, and contextually fitting. These ratings give a useful insight into how well the model performs across different areas, and not just based on surface-level similarity.

The second evaluation method focuses more on the semantic and lexical similarity between the model's output and the ground truth. This was done using three well-known NLP metrics: BLEU, ROUGE, and BERTScore.

- BLEU (Bilingual Evaluation Understudy)[26] checks how much overlap there is between the words or short phrases (n-grams) in the generated sentence and the reference sentence. It's mostly used in translation tasks but still helpful here for checking basic similarity.

The formula of BLEU is as follows:

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

Where BP is the Brevity Penalty which is introduced to penalise short sentences which score high because of many similar words, but do not carry the same essence as the reference sentence. w_n is the weight for n-gram precision commonly set to $\frac{1}{N}$ which is also the case for this experiment. p_n is the precision for n-grams this measures the count of n-grams in the candidate sentence compared tot the reference sentence.

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[27] is more recall-focused. It measures how much of the reference sentence is captured in the generated one.

In this thesis the ROUGE-L variant is used to evaluate the model. The formula for ROUGE-L is as follows:

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \cdot R_{\text{LCS}} \cdot P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 \cdot P_{\text{LCS}}}$$

This formula may look complex, but in essence calculates how well a candidate sentence matches a reference sentence by measuring the longest sequence of words they share in order. The longer this common subsequence, the more similar the sentences and the higher the ROUGE-L score are.

- BERTScore [28] uses contextual embeddings from a large language model to evaluate how similar two sentences are in meaning, not just in wording. It’s a deeper way to check if the generated sentence ”understands” the original, even if the wording is different.

Looking at the results from both evaluation methods puts the model performance into better perspective. On one side, we get human-like ratings using GPT-4o, which judge medical helpfulness and contextual quality. On the other side, we get a more technical look at how close the generated response is to the ground truth in terms of language and meaning. This combination provides a balanced and thorough way to evaluate the models.

5 Results

5.1 Analyse of results

The results were conducted entirely in Python, and the three models were tested on two distinct test datasets, each containing 100 entries, both with and without clinical context included, resulting in 400 responses per model, and a total of 1,200 evaluated responses. Examples of a model response and the evaluation prompt are shown in Figure 3, 4, 5, 6, and 7. The three models used for benchmarking were Mistral, BioMistral, and the fine-tuned BioMistral on NoteChat data. Here, Mistral serves as a baseline for the BioMistral model, while BioMistral functions as a baseline for its fine-tuned variant. This setup provides a structured comparison to observe the isolated effects of pre-training on biomedical data and subsequent fine-tuning in a dialogue-specific setting between patient and doctor.

All experiments with the LLMs were conducted using a fixed temperature setting of 0.3. This value encourages slightly more diverse and creative responses while maintaining a level of consistency across models. Using the same temperature for each experiment ensures comparability and reduces variability in output due to sampling randomness.

5.1.1 Humanlike evaluation

Metric	Mistral	BioMistral	Fine-tuned BioMistral
With clinical data			
Helpfulness	3.122 ± 1.237	3.143 ± 1.210	3.673 ± 0.859
Answer Quality	3.122 ± 1.237	3.112 ± 1.209	3.643 ± 0.876
Without clinical data			
Helpfulness	2.724 ± 1.138	2.878 ± 1.077	2.980 ± 1.157
Answer Quality	2.704 ± 1.142	2.827 ± 1.075	2.959 ± 1.157

Table 1: Human-like evaluation results generated by GPT-4o for responses from different models on the normal test dataset. Scores range from 0 to 5 and are reported as mean $\pm$ standard deviation.

Table 1 shows the performance of the three evaluated models, Mistral, BioMistral, and fine-tuned BioMistral, on the normal test dataset with a distinction between the the presence and absence of clinical data. Each model was assessed on 100 entries, and the scores reflect human-like evaluations generated by GPT-4o.

When clinical data is included the three models all demonstrates a higher mean scores across the two metrics compared to when the clinical data is not included. The fine-tuned BioMistral achieves the highest mean scores in both helpfulness and answer quality with and without clinical data. The standard deviation is also lower when clinical data is included, suggesting a more consistent response compared to the other two models. At the other hand, the Mistral model shows lower mean scores in the two metrics, particularly in the absence of clinical data.

Table 2 shows the performance of the three evaluated models on the *test_diagnose* test dataset with a distinction between the the presence and absence of clinical data. Each model was assessed on 100 entries, and the scores reflect human-like evaluations generated by GPT-4o.

Metric	Mistral	BioMistral	Fine-tuned BioMistral
With clinical data			
Diagnostic Accuracy	3.480 ± 1.761	3.550 ± 1.877	4.050 ± 1.344
Helpfulness	3.260 ± 1.461	3.420 ± 1.519	3.700 ± 1.202
Answer Quality	3.250 ± 1.403	3.380 ± 1.509	3.620 ± 1.187
Without clinical data			
Diagnostic Accuracy	1.810 ± 1.549	2.490 ± 1.403	2.320 ± 1.377
Helpfulness	2.110 ± 1.171	2.800 ± 1.092	2.560 ± 1.057
Answer Quality	2.130 ± 1.116	2.700 ± 1.068	2.510 ± 1.020

Table 2: Human-like evaluation results generated by GPT-4o for responses from different models on the *test_diagnose* test dataset. Scores range from 0 to 5 and are reported as mean $\pm$ standard deviation.

When clinical data is included, all three models again achieve noticeably higher scores across all metrics, diagnostic accuracy, helpfulness, and answer quality, compared to the condition without clinical information. The fine-tuned BioMistral model consistently achieves the highest mean scores in all three metrics under the clinical condition, with a lower standard deviation, indicating greater consistency in its outputs. In the absence of clinical data, BioMistral slightly outperforms the fine-tuned version on diagnostic accuracy, though the margin is narrow.

5.1.2 Sementic similarity

Metric	Mistral	BioMistral	Fine-tuned BioMistral
With clinical data			
BERTScore F1	0.8957 ± 0.0502	0.8767 ± 0.0393	0.9093 ± 0.0455
ROUGE-L	0.3429 ± 0.2864	0.2292 ± 0.2211	0.4303 ± 0.2666
BLEU	0.2278 ± 0.2538	0.1296 ± 0.1589	0.2802 ± 0.2681
Without clinical data			
BERTScore F1	0.8720 ± 0.0394	0.8691 ± 0.0377	0.8826 ± 0.0390
ROUGE-L	0.2252 ± 0.1644	0.2024 ± 0.1595	0.2669 ± 0.2034
BLEU	0.1119 ± 0.1114	0.1094 ± 0.1178	0.1616 ± 0.1748

Table 3: Semantic based evaluation results from different models on the normal test dataset. Scores range from 0 to 1 and are reported as mean $\pm$ standard deviation.

Table 3 shows the results of the semantic-based evaluation of model responses on the normal test dataset. For all three metric, the fine-tuned BioMistral model achieves the highest mean score, both with and without clinical data. It shows the most improvements in ROUGE-L and BLEU while the BERTScore F1 is for every model high. Inclusion of clinical data results in a higher score for every model but it is most noticeable in for the BLEU and ROUGE-L metric. The Mistral and BioMistral models show a larger drop in performance when clinical information is excluded, whereas the fine-tuned BioMistral maintains a more stable performance, reflecting a higher degree of robustness.

Metric	Mistral	BioMistral	Fine-tuned BioMistral
With clinical data			
BERTScore F1	0.8896 ± 0.0497	0.8951 ± 0.0553	0.9042 ± 0.0512
ROUGE-L	0.3342 ± 0.2575	0.3620 ± 0.2943	0.3930 ± 0.2742
BLEU	0.2042 ± 0.2302	0.2416 ± 0.2716	0.2511 ± 0.2480
Without clinical data			
BERTScore F1	0.8573 ± 0.0293	0.8686 ± 0.0307	0.8679 ± 0.0296
ROUGE-L	0.1679 ± 0.1352	0.2039 ± 0.1421	0.2085 ± 0.1306
BLEU	0.0671 ± 0.0700	0.0838 ± 0.0826	0.0848 ± 0.0829

Table 4: Semantic based evaluation results from different models on the *test_diagnose* dataset. Scores range from 0 to 1 and are reported as mean $\pm$ standard deviation.

Table 4 shows the results of the semantic-based evaluation of model responses on the *test_diagnose* dataset. Again BioMistral Fine-tuned model achieves the highest scores on all the three semantic based metrics. This time there is a narrow difference between the BLEU score of BioMistral and the fine-tuned BioMistral. There is however the same noticeable difference between the results with clinical data and without, showing an almost double the score for ROUGE-L Fine-tuned BioMistral with clinical data.

5.2 Case study of the results

In addition to the quantitative and semantic evaluations, I manually reviewed a subset of model outputs to assess the qualitative differences between Mistral, BioMistral, and the fine-tuned BioMistral. By creating a consistent sample of responses across the three models, a side-by-side comparison was possible, providing clearer insight into how each model approaches medical dialogue. Five examples of this are shown in the appendix in figure 3, 4, 5, 6, 7.

This manual inspection helped me see that the fine-tuned version consistently produced responses that felt more natural and human-like. The output from the model is generally easier to understand because it often avoids the use of technical detail that the base Mistral and BioMistral models do include. This distinction is important in the clinical dialogue system as clarity and accessibility are essential for patient comprehension. Medical advice must be understandable for people with varying levels of health literacy, and the fine-tuned model appeared to be better adapted to this requirement.

The fine-tuned responses were also more concise. Rather than overwhelming the user with lengthy or overly detailed explanations, the model focused on delivering key information in a compact format. This not only improves readability but also makes the conversation more efficient. Despite the difference in content length, the three models all maintained a similar tone which was formal and friendly.

6 Discussion

6.1 Analyse of results

The results from the experiments showed that fine-tuning BioMistral on doctor-patient dialogue data improves performance across all benchmarks. Even without fine-tuning, the BioMistral model outperforms the general-purpose Mistral model, highlighting the value of biomedical pre-training. The fine-tuned BioMistral delivers the best overall performance, especially in helpfulness, answer quality, and diagnostic accuracy. This trend is also reflected in the semantic similarity metrics, where the fine-tuned model produces responses that are more aligned in both meaning and literal semantics with the expected outputs.

Providing clinical data to the LLM significantly boosts its performance. When the model has access to clinical context, its responses become more accurate, helpful, and consistent. This makes sense, as clinical data is often the foundation of the conversation. Giving the model access to this information is like what happens in real-world scenarios: doctors typically have access to a patient’s medical history or background information during a consultation.

The fine-tuned model performs well across all tests except one: the *test_diagnose* dataset without clinical data. In this scenario, the fine-tuned model performs slightly worse than the BioMistral. One possible explanation is that the fine-tuned model was trained with access to clinical data, making it more reliant on that context. As a result, it struggles to infer diagnoses using only the conversation, without structured medical input. This suggests that for accurate diagnosis it needs more than just dialogue such as contextual data, just like in real medical settings.

The results show that domain-specific fine-tuning increases the performance of model in clinical dialogue tasks but is it not enough to achieve optimal results. Fine-tuning on task-specific data like doctor-patient conversations is essential for generating more relevant and coherent responses. At the same time, pre-training on biomedical data still plays a critical role in preparing the model for these types of tasks.

6.2 Limitations and challenges

The first limitation is that the evaluation in this thesis was performed using GPT-4o, not by a real human expert like a clinician or a general medical professional. Because of this, the results may not fully match clinical standards. Having a human in the evaluation loop would give a much better sense of how the model performs in an actual medical setting.

Another limitation is that all the training and test data used were synthetic, generated by another LLM. The test data were split across two different methods and tried to simulate realistic doctor-patient conversations, they still do not capture the full complexity and unpredictability you would find in real clinical interactions. Real conversations with doctors often include ambiguous phrasing, irrelevant details, or unexpected turns which synthetic data just does not capture well. This lack of realism can negatively affect the model’s ability to handle real-world patient interactions. Using real clinical dialogue data would make the model more robust and realistic in its responses.

The models used in both the benchmarking and fine-tuning were quantised, meaning they're smaller, compressed versions of the original full models. While this reduces computational load and makes it possible to run on limited hardware, it also reduces the number of parameters and, in turn, the model's ability to produce high-quality responses. Fine-tuning still improves these models, but the ceiling is lower compared to full-size models. Plus, since the fine-tuning was done while the models were quantised, this double-limits how much the model can improve. It was a necessary trade-off due to hardware constraints even on the ALICE supercomputer, but it still remains a limiting factor.

The data used for fine-tuning play a massive role in how well the model performs. During the experiments, it became clear that the formatting of the data is critical. Using the wrong format led to a completely unusable model that could not generate anything sensible. Even the final working version was not trained on perfect data. It was fine-tuned using just one dataset: NoteChat. While that dataset improved its ability in clinical dialogue setting, as shown in the results, it is limited in scope. Adding more diverse datasets could help the model generalise better and perform more reliably in various medical dialogue settings.

An earlier attempt was made to fine-tune BioMistral using the MIMIC dataset [20], but due to the sheer volume and the nature of the clinical notes, the resulting model couldn't produce coherent outputs. This again highlights how important it is to have clean, well-formatted, and diverse data when fine-tuning LLMs for specialised tasks.

7 Conclusions and future work

This thesis explored how fine-tuning large language models can improve the accuracy of diagnoses, medical recommendations, and overall performance in a text-based medical dialogue system. This was done by fine-tuning the BioMistral model and benchmarking it against both its own base version (BioMistral pre-trained only) and the general-purpose Mistral model. The fine-tuning was done using LoRA, a low-resource method that still manages to achieve strong results. The experiments showed that fine-tuning BioMistral using medical dialogue data (NoteChat) improved the model’s performance in terms of helpfulness, answer quality, and diagnostic accuracy.

1. How does fine-tuning impact the performance of LLMs in medical dialogue tasks compared to their base versions?

Fine-tuning has a big positive impact. While the base BioMistral already performs better than general-purpose models thanks to its biomedical pre-training, the fine-tuned version goes even further. It generates more aligned, context-aware, and helpful responses. It also scores higher on semantic similarity and diagnostic accuracy. This proves that while pre-training on medical data helps, domain-specific fine-tuning takes it even further. However, one exception was the diagnosis task without access to clinical data, where the pre-trained model slightly outperformed the fine-tuned one. This suggests that the fine-tuned model learned to rely on clinical data, and without it, performance can drop.

2. What are the key challenges and limitations of fine-tuning LLMs for medical applications, and how can they be addressed?

The process comes with a few challenges and limitations. First, the evaluation in this thesis was done using GPT-4o and my own personal view, so the results might not fully reflect clinical standards as this evaluation is not done by a professional. Second, the data used for training and testing were synthetic, generated by another LLM. While this makes experimentation easier, it does not reflect the complexity of real-world medical conversations. Using real patient-doctor dialogues would give better results. Third, the models were quantised versions, which limited their capacity and answer quality. Fine-tuning full-size models would likely lead to better performance but needs more resources. Lastly, data formatting turned out to be crucial. Badly formatted data made the model completely unusable, and even the final version was trained on a single dataset (NoteChat). Adding more diverse datasets and ensuring proper formatting would help future models perform better.

Future research could be done on the full-sized versions of BioMistral and Mistral to test if fine-tuning has the same or even greater effect when the model is not quantised. In addition, fine-tuning on real patient conversations or combining multiple datasets could give deeper insight and boost performance in medical dialogue settings even more.

References

- [1] Y. Wang, Z. Li, M. Bin, M. Sheng, J. Hou, X. Han, Y. Chen, J. Yang, Q. Yu, D.-S. Huang, W. Chen, and Q. Zhang, “Epr: Entity perception and reasoning for medical dialogue system,” in *Advanced Intelligent Computing Technology and Applications*, vol. 14874 of *Lecture Notes in Computer Science*, pp. 312–323, Singapore: Springer, 2024.
- [2] K. Middleton, M. Butt, N. Hammerla, S. Hamblin, K. Mehta, and A. Parsa, “Sorting out symptoms: design and evaluation of the ‘babylon check’ automated triage system,” 2016.
- [3] Ada Health, “Ada - check symptoms & find care,” 2025. Accessed: 2025-05-24.
- [4] H. Ali, J. Qadir, T. Alam, M. Househ, and Z. Shah, “Chatgpt and large language models in healthcare: Opportunities and risks,” in *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, pp. 1–4, 2023.
- [5] H. Hassan, Z. Ren, C. Zhou, M. A. Khan, Y. Pan, J. Zhao, and B. Huang, “Supervised and weakly supervised deep learning models for covid-19 ct diagnosis: A systematic review,” *Computer Methods and Programs in Biomedicine*, vol. 218, p. 106731, 2022.
- [6] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun, “A comprehensive survey on pretrained foundation models: A history from bert to chatgpt,” 2023.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.
- [8] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” tech. rep., OpenAI, 2018.
- [9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” 2020.
- [10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019.
- [11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” 2023.
- [12] OpenAI, “Gpt-4 technical report,” 2024.

- [13] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, “Biogpt: generative pre-trained transformer for biomedical text generation and mining,” *Briefings in Bioinformatics*, vol. 23, Sept. 2022.
- [14] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, J. Yi, W. Zhao, X. Wang, Z. Liu, H.-T. Zheng, J. Chen, Y. Liu, J. Tang, J. Li, and M. Sun, “Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models,” 2022.
- [15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” 2021.
- [16] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, p. 1234–1240, Sept. 2019.
- [17] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, “Domain-specific language model pretraining for biomedical natural language processing,” *ACM Transactions on Computing for Healthcare*, vol. 3, p. 1–23, Oct. 2021.
- [18] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, “Pubmedqa: A dataset for biomedical research question answering,” 2019.
- [19] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. A. y Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan, “Towards expert-level medical question answering with large language models,” 2023.
- [20] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” *Scientific Data*, vol. 3, p. 160035, May 2016.
- [21] K. Huang, J. Altsosaar, and R. Ranganath, “Clinicalbert: Modeling clinical notes and predicting hospital readmission,” 2020.
- [22] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu, “Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records,” 2022.
- [23] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, “What disease does this patient have? a large-scale open domain question answering dataset from medical exams,” 2020.
- [24] J. Wang, Z. Yao, Z. Yang, H. Zhou, R. Li, X. Wang, Y. Xu, and H. Yu, “Notechat: A dataset of synthetic patient-physician conversations conditioned on clinical notes,” in *Findings of*

the Association for Computational Linguistics ACL 2024, p. 15183–15201, Association for Computational Linguistics, 2024.

- [25] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, “BioMistral: A collection of open-source pretrained large language models for medical domains,” in *Findings of the Association for Computational Linguistics: ACL 2024* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), (Bangkok, Thailand), pp. 5848–5864, Association for Computational Linguistics, Aug. 2024.
- [26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL ’02*, (USA), p. 311–318, Association for Computational Linguistics, 2002.
- [27] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.
- [28] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” 2020.

A Appendix

Actual response: Alright. On admission, you were febrile to 38.1°C, tachycardic to 113 bpm, and hypotensive to 96/63. Is that correct?		
Mistral: Okay, thank you for letting me know. Can you tell me about your family history and social history?	BioMistral: Okay, thank you for letting me know. What other medications are you taking?	Fine-tuned BioMistral: Okay, thank you for letting me know. Can you tell me about your vital signs when you came in today?
GPT-4o:		
Helpfulness: 2 Answer Quality: 2 Reason: The predicted response fails to address the immediate clinical concerns presented by the patient, such as the rash, fever, and recent initiation of clozapine. Instead, it inappropriately shifts focus to family and social history, which is not immediately relevant to the acute presentation. The real response acknowledges the patient's vital signs and clinical status, which is crucial for assessing the situation. Therefore, the predicted response is less helpful and lacks the necessary clinical focus and reasoning.	Helpfulness: 3 Answer Quality: 3 Reason: The predicted response asks about other medications, which is relevant to understanding potential drug interactions or side effects, such as the rash. However, it does not address the immediate clinical status of the patient, such as the vital signs and symptoms that were noted on admission, which the real response does. This makes the predicted response less helpful in assessing the urgency and severity of the patient's condition. The answer quality is moderate as it lacks completeness and does not fully engage with the critical aspects of the patient's current health status.	Helpfulness: 4 Answer Quality: 4 Reason: The predicted response is medically helpful and appropriate as it seeks to gather more information about the patient's current condition, specifically their vital signs, which is relevant given the symptoms and recent medication change. However, it is slightly less specific than the real response, which directly references the patient's vital signs upon admission. The predicted response is well-reasoned and contextually appropriate, but it could be improved by directly acknowledging the patient's reported symptoms and linking them to the potential side effects of clozapine or an infectious process.

Figure 3: An example prompt with the response for each model and the evaluation by GPT-4o

Actual response: And a Magnetic Resonance Imaging (MRI) study of your brain demonstrated multifocal lesions in the left hemisphere.		
Mistral: And you also had an MRI study of the brain, which demonstrated multifocal lesions in the left hemisphere.	BioMistral: And then you had an MRI study of the brain which demonstrated multifocal lesions in the left hemisphere (Figure A). What was your next step?	Fine-tuned BioMistral: And MRI studies of the brain demonstrated multifocal lesions in the left hemisphere.
GPT-4o:		
Helpfulness: 5 Answer Quality: 5 Reason: The predicted response accurately reflects the information provided in the patient's medical history. It correctly identifies the MRI findings of multifocal lesions in the left hemisphere, which is a critical detail in understanding the patient's condition. The response is medically helpful and contextually appropriate, as it aligns with the real response and maintains the continuity of the clinical conversation.	Helpfulness: 4 Answer Quality: 4 Reason: The predicted response accurately reflects the patient's MRI findings, which is crucial for understanding the patient's condition. It also prompts the patient to discuss the next steps, which can be helpful in guiding the conversation. However, it lacks the detail of the real response, which continues to build on the patient's medical history and context. The predicted response is medically appropriate and contextually relevant, but it could be improved by including more specific follow-up questions or information to better guide the clinical conversation.	Helpfulness: 5 Answer Quality: 5 Reason: The predicted response accurately reflects the information provided in the patient's history. It confirms the MRI findings of multifocal lesions in the left hemisphere, which is a critical detail in understanding the patient's condition. The response is medically appropriate, contextually relevant, and maintains the continuity of the conversation. There is no significant difference between the real and predicted responses in terms of medical content and clarity.

Figure 4: An example prompt with the response for each model and the evaluation by GPT-4o

Actual response: Okay. We have performed a temporal bone high resolution CT (HR CT) and it showed migration of the stapes prosthesis into the vestibule (protrusion). The depth of protrusion was measured 14 millimeters from the tip of the prosthesis to the oval window.		
Mistral: I see. A temporal bone high resolution CT (HR CT) was performed and showed migration of the stapes prosthesis into the vestibule (protrusion). The diagnosis of intravestibular stapes prosthesis protrusion related post stapedectomy vertigo was made and revision surgery was proposed.	BioMistral: I see. A temporal bone high resolution CT was performed and showed migration of the stapes prosthesis into the vestibule (protrusion) (Fig.). Depth of protrusion was measured 14 millimeters from tip of the prosthesis to the oval window. The diagnosis of intravestibular stapes prosthesis protrusion related post stapedectomy vertigo was made and revision surgery was proposed.	Fine-tuned BioMistral: Okay. We performed a temporal bone high resolution CT (HR CT) for you and it showed migration of the stapes prosthesis into your vestibule (protrusion). The depth of protrusion was measured 14 millimeters from tip of the prosthesis to the oval window.
GPT-4o:		
Helpfulness: 4 Answer Quality: 4 Reason: The predicted response is medically helpful and appropriate as it acknowledges the CT findings and the diagnosis of stapes prosthesis protrusion. It also mentions the proposed revision surgery, which is a logical next step in management. However, it lacks the specific detail about the depth of protrusion (14 millimeters) provided in the real response, which could be clinically relevant. Overall, the response is well-reasoned and contextually appropriate, but slightly less complete than the real response.	Helpfulness: 5 Answer Quality: 5 Reason: The predicted response is highly helpful and contextually appropriate. It accurately summarizes the findings from the temporal bone high resolution CT, including the migration of the stapes prosthesis and the depth of protrusion. It also correctly identifies the diagnosis and the proposed treatment plan, which is revision surgery. The response is well-reasoned and complete, addressing the patient's concern about the cause of vertigo post-stapedectomy.	Helpfulness: 5 Answer Quality: 5 Reason: The predicted response is medically helpful and appropriate as it confirms the diagnosis of stapes prosthesis protrusion, which is relevant to the patient's symptoms. It provides specific details from the CT scan, including the measurement of the protrusion, which is crucial for understanding the condition. The response is well-reasoned, complete, and contextually appropriate, as it aligns with the information provided in the conversation and offers a clear explanation of the findings.

Figure 5: An example prompt with the response for each model and the evaluation by GPT-4o

Actual response: Okay, I understand. And what was the treatment regimen changed to?		
Mistral: Hmm, I see. And what was the treatment regimen changed to?	BioMistral: Okay, and what was the diagnosis?	Fine-tuned BioMistral: Okay, and what was the treatment regimen changed to?
GPT-4o:		
Helpfulness: 4 Answer Quality: 4 Reason: The predicted response is nearly identical to the real response and continues the logical flow of the conversation by inquiring about the change in treatment regimen. This question is relevant and necessary to understand the progression of the patient's condition and the medical interventions attempted. However, it lacks any additional probing or suggestions that could enhance the helpfulness and completeness of the response.	Helpfulness: 2 Answer Quality: 2 Reason: The predicted response repeats a question that has already been answered by the patient, which does not advance the conversation or provide any new insight into the patient's condition. It fails to address the progression of the patient's treatment, which is crucial given the deterioration of the patient's condition. The real response appropriately inquires about the change in treatment regimen, which is a logical next step in understanding the patient's medical history and current condition. Therefore, the predicted response is less helpful and lacks quality in terms of advancing the clinical discussion.	Helpfulness: 4 Answer Quality: 4 Reason: The predicted response is nearly identical to the real response, asking about the change in treatment regimen, which is a logical next step in understanding the patient's medical history. It is medically appropriate and continues the line of inquiry needed to assess the patient's condition. However, it lacks any additional probing or empathy that could enhance the interaction, which is why it does not receive a perfect score.

Figure 6: An example prompt with the response for each model and the evaluation by GPT-4o

Actual response: I see. And what kind of treatment did you receive?		
Mistral: I see. And what was the response to carboplatin and etoposide chemotherapy between Oct 2014 and April 2015 with RECIST partial response after three cycles but with evidence of tumor growth after cycle 6?	BioMistral: Alright. So you've been treated with carboplatin and etoposide chemotherapy since October 2014, correct?	Fine-tuned BioMistral: I see. And what was your treatment plan after diagnosis?
GPT-4o:		
Helpfulness: 4 Answer Quality: 4 Reason: The predicted response is medically helpful as it inquires about the patient's response to a specific chemotherapy regimen, which is relevant to understanding the progression and management of the patient's cancer. However, it is slightly less open-ended than the real response, which more broadly asks about treatment, potentially allowing the patient to provide a more comprehensive treatment history. The predicted response is well-reasoned and contextually appropriate, as it directly addresses the patient's previous statement about chemotherapy, but it could have been more complete by also asking about subsequent treatments or overall treatment history.	Helpfulness: 3 Answer Quality: 3 Reason: The predicted response acknowledges the patient's treatment history, which is relevant to the ongoing conversation. However, it incorrectly assumes the treatment is ongoing ('since October 2014'), which is not accurate as the patient has undergone multiple treatment phases. The real response is more open-ended, allowing the patient to provide a complete treatment history, which is more helpful and contextually appropriate. The predicted response is partially helpful but lacks completeness and accuracy.	Helpfulness: 4 Answer Quality: 4 Reason: The predicted response is medically helpful and appropriate as it seeks to understand the patient's treatment plan, which is crucial for assessing the current medical situation. It is slightly less specific than the real response, which directly asks about the treatment received, but it still effectively prompts the patient to provide relevant information about their treatment history. The predicted response is well-reasoned and contextually appropriate, though it could be slightly more direct in asking about the treatment received rather than the plan.

Figure 7: An example prompt with the response for each model and the evaluation by GPT-4o