Doc Bot: The Medical LLM Fine-tuned on LLaMA 3 8B Using LoRA and Insights from the Medical Field

Abdulmalik Habaebi

mhabaebi@gmail.com

International Islamic University Malaysia

Akeem Olowolayemo

International Islamic University Malaysia

Sharyar Wani

International Islamic University Malaysia

Research Article

Keywords: Large Language Models (LLMs), Medical Chatbot, Low-Rank Adaptation (LoRA), Fine-Tuning, LLaMa 3, Parameter-Efficient Fine-Tuning (PEFT)

Posted Date: September 4th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7515191/v1

Additional Declarations: No competing interests reported.

Doc Bot: The Medical LLM Fine-tuned on LLaMA 3 8B Using LoRA and Insights from the Medical Field

Abdulmalik Mohamed Hadi Habaebi, Akeem Olowolayemo, Sharyar Wani

mhabaebi@gmail.com akeem@iium.edu.my sharyarwani@iium.edu.my

Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Selangor, Malaysia.

Abstract

Purpose

General-purpose large language models (LLMs) often lack the specialized accuracy required for the medical domain. This research aims to address this gap by developing and evaluating DocBot, a medical LLM, to demonstrate that domain-specific fine-tuning can significantly enhance performance even with constrained computational resources.

Methods

We fine-tuned the Meta LLaMa 3.1 8B model using Low-Rank Adaptation (LoRA), a parameter-efficient technique. The model was trained on a curated dataset of 2,000 patient-doctor dialogues sourced from ClinicalTrials, EMEA, and PubMed, using a single Tesla T4 GPU. Performance was evaluated against the base LLaMa model using BERTScore, BLEU, and ROUGE metrics, with responses from verified medical professionals serving as the reference.

Results

DocBot demonstrated significant improvements over the base LLaMa 3.1 8B model across all evaluation metrics. Specifically, DocBot achieved a higher BERTScore F1-score (83.56% vs. 81.47%), indicating enhanced semantic accuracy, fluency, and alignment with expert-generated text. The gains in precision and recall further confirm the model's superior ability to generate relevant and comprehensive medical information.

Conclusion

The successful development of DocBot showcases the feasibility and impact of creating domain-optimized LLMs efficiently. The results highlight the potential for specialized models to serve as reliable tools for augmenting clinical decision-making and delivering accessible medical support, particularly in resource-limited environments, paving the way for further innovation in specialized AI applications.

Keywords

Large Language Models (LLMs), Medical Chatbot, Low-Rank Adaptation (LoRA), Fine-Tuning, LLaMa 3, Parameter-Efficient Fine-Tuning (PEFT)

Introduction

Related Work

Medical chatbots have garnered significant attention with the advent of large language models (LLMs), which have demonstrated remarkable capabilities in natural language understanding and generation. Prior studies, such as ChatDoctor (Li et al., 2023), have highlighted the effectiveness of fine-tuning LLMs on domain-specific datasets in the medical field. ChatDoctor fine-tuned the LLaMA model on a dataset comprising 100,000 patient-doctor dialogues, achieving impressive results in diagnostic accuracy and response reliability. However, the approach faced scalability challenges due to its reliance on full-parameter tuning, which requires substantial computational resources and time. This limitation underscores the need for more efficient fine-tuning methods, especially in resource-constrained environments.

The challenges of LLMs in the medical domain extend beyond scalability. A critical issue is hallucination, where models generate plausible but factually incorrect information. For instance, GPT-4 has been reported to hallucinate in approximately 21% of medical answers (Salvagno et al., 2023). Such inaccuracies are unacceptable in medical scenarios, where reliability and precision are paramount. To address this, researchers have emphasized the importance of using high-quality, peer-reviewed datasets during both pre-training and fine-tuning phases. For example, Med-PaLM (Singhal et al., 2022) incorporated curated medical literature from sources like PubMed and clinical guidelines, significantly reducing hallucination rates and improving response accuracy.

Another challenge is the generalized nature of pre-trained LLMs, which lack domain-specific knowledge. While models like GPT-4 excel in general-purpose tasks, they often struggle with specialized medical queries, particularly those involving rare diseases or complex pharmacodynamics. This limitation has spurred the development of domain-specific LLMs, such as ClinicalGPT (Wang et al., 2023), which leverages medical datasets like MIMIC-III and ClinicalTrials.gov to enhance its diagnostic capabilities. However, these models often require extensive computational resources, making them inaccessible to smaller research teams or institutions with limited infrastructure.

To mitigate these challenges, parameter-efficient fine-tuning techniques have gained traction. Among these, Low-Rank Adaptation (LoRA) (Hu et al., 2021) has emerged as a promising solution. LoRA fine-tunes only 1–10% of a model's parameters by injecting low-rank matrices into the architecture, drastically reducing computational overhead. This approach has been successfully applied in medical LLMs, such as BioGPT (Luo et al., 2022), which used LoRA to achieve state-of-

the-art performance on biomedical text generation tasks while maintaining computational feasibility. The efficiency of LoRA makes it particularly suitable for resource-constrained environments, enabling faster experimentation and iteration.

Dataset quality is another critical factor influencing the performance of medical LLMs. While widely used datasets like PubMed and WebMD provide high-quality, peer-reviewed information, they primarily consist of static text, such as research articles and clinical guidelines. This lack of interactive dialogues limits their utility for training conversational agents. To address this, researchers have developed datasets like the Ruslan MV Medical Dialogue Dataset (Ruslan Magana Vsevolodovna, 2023), which contains 21 million patient-doctor interactions scraped from sources like ClinicalTrials and PubMed. Such datasets enable models to learn the nuances of medical conversations, improving their ability to handle real-world queries.

Despite these advancements, several gaps remain in the literature. First, there is a lack of focus on rare diseases and complex medical conditions, which are often underrepresented in existing datasets. Second, the ethical implications of using patient data for training LLMs, such as privacy concerns and data anonymization, require further exploration. Finally, the scalability of fine-tuning methods like LoRA in larger models (e.g., LLaMA 65B) remains an open question, particularly in terms of balancing performance and computational cost.

Materials And Methods

Architecture

Since the model used in this study was Meta's LLaMa 3.1 8B parameter version, it uses a standard transformer model with the added various optimisations to the model for improvements in efficiency and stability. The model sizes range from 7B to 65B parameter versions. For this model's 8B parameter size it has 32 layers and 32 attention heads. These variables increase as the model size increases. LLaMA uses a feedforward size of $2\frac{2}{3}d$(where d is the hidden dimension) which is for smaller models such as 8B. This gives the model an edge in terms of using less memory and compute for training and fine-tuning reducing the compute cost by 33%.

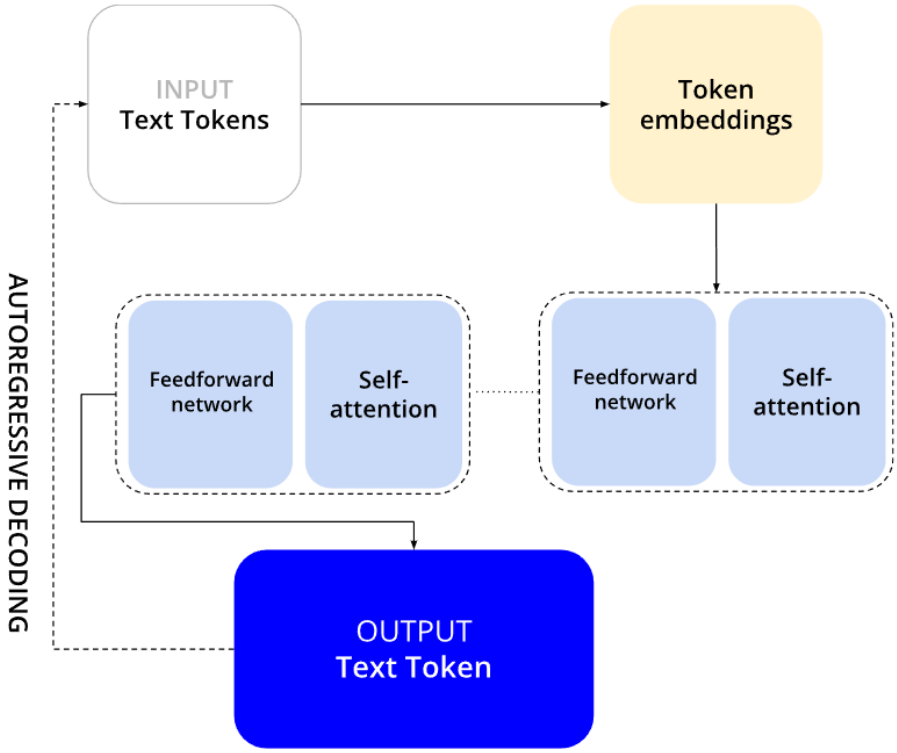

Figure 1: Illustration of the architecture and training of LLaMA 3. (Meta, 2024)

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Equation 1: Scaled Dot-Product Attention Equation (Vaswani et al., 2017)

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_h)W^o \quad (2)$$

Where:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (2.1)$$

Equation 2: Multi-Head Attention Equation (Vaswani et al., 2017)

$$FFN(x) = (0, xW_1 + b_1)\,W_2 + b_2 \quad (3)$$

Equation 3: Feed-Forward Network Equation (Vaswani et al., 2017)

$$L_{fine-tune} = -\frac{1}{N}\sum_{i=1}^{N} \quad log\ log\ P(y_i|x_i;\theta) \quad (4)$$

Equation 4: Fine-Tuning Loss Function Equation (Brown et al., 2020)

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_t = \theta_{t-1} - \eta\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} + \lambda\theta_{t-1}\right) \quad (5)$$

Equation 5: AdamW Optimization Function Equation (Loshchilov & Hutter, 2017)

These formulas collectively define the underlying mechanics of the Transformer architecture, enabling modern language models like LLaMA 3.1 to understand and generate text. The process begins with the scaled dot-product attention mechanism, which determines how much focus each token should allocate to others within a sequence. By computing similarity scores between tokens and normalizing them, this mechanism ensures that the model captures meaningful relationships and dependencies, allowing it to build contextual representations.

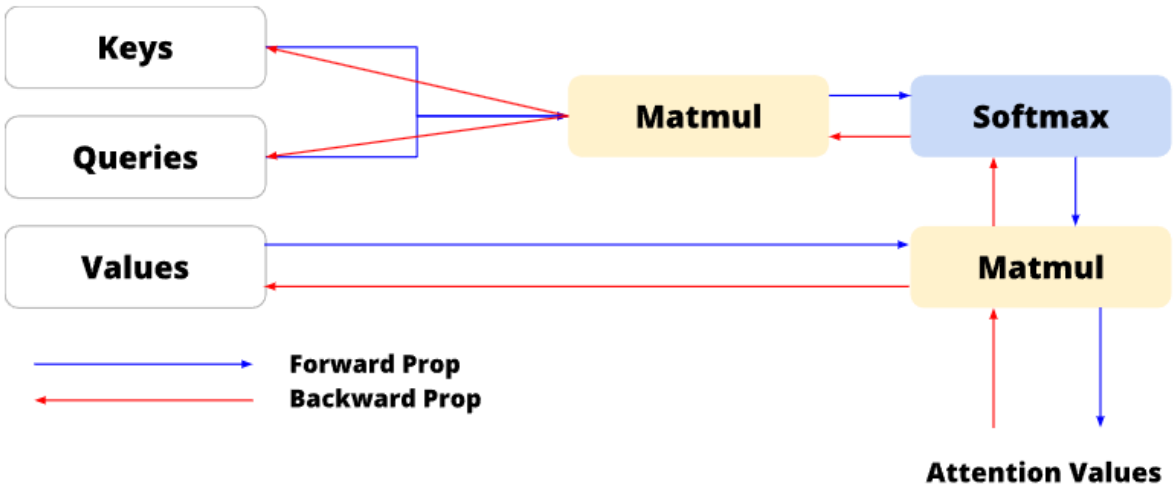

Figure 2: Illustration of the flow of computations in the scaled dot product attention mechanism in the attention block.

To further enhance this capability, multi-head attention extends the single-head attention mechanism by enabling multiple perspectives on the input. Each attention head processes the input sequence in different subspaces, capturing diverse contextual relationships. By concatenating and transforming the outputs from multiple attention heads, the model achieves a richer and more nuanced understanding of textual dependencies, making it more adept at tasks like translation and text generation.

Once the attention mechanism has established meaningful token relationships, the output is passed through a feed-forward network. This layer introduces non-linearity into the model, allowing it to learn complex hierarchical patterns that a purely linear transformation could not capture. By applying activation functions and weighted transformations, the feed-forward network refines token representations before passing them to subsequent layers, ensuring that the model can recognize intricate linguistic structures.

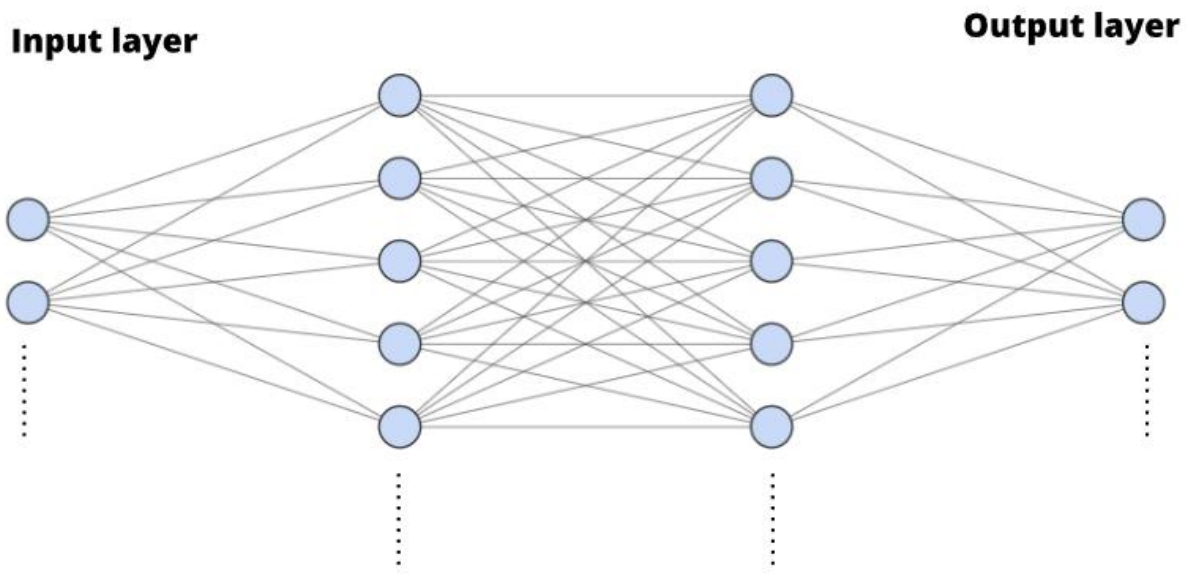

Figure 3: Illustration of the feedforward network

Throughout the training process, the model optimizes its predictions using a fine-tuning loss function. This function evaluates how well the model predicts the next token in a sequence, adjusting its parameters accordingly to minimize errors. By penalizing incorrect predictions and reinforcing correct ones, the model gradually improves its ability to generate coherent and contextually appropriate text. This function serves as the guiding metric for training, ensuring that the model aligns closely with the patterns observed in the training data.

The optimization of model parameters is facilitated by an adaptive optimization algorithm that leverages gradient-based updates. By maintaining moving averages of the gradients and applying weight decay, this method ensures stable and efficient learning. The adaptive learning rates prevent drastic parameter updates, helping the model converge more effectively while avoiding overfitting. The inclusion of weight decay further regularizes the model, reducing the likelihood of over-reliance on specific patterns in the training data.

Together, these components create a structured pipeline for text processing. The attention mechanism extracts contextual information, multi-head attention broadens the scope of understanding, the feed-forward network refines token representations, and the fine-tuning loss function guides learning. The optimization algorithm ties these elements together by adjusting parameters to enhance performance over time. Layered in depth, this iterative process results in a powerful language model capable of producing coherent, contextually rich, and grammatically sound text.

By integrating these formulas, the Transformer architecture achieves state-of-the-art performance in natural language processing. The combination of attention mechanisms, non-linear transformations, and adaptive optimization forms a highly efficient and scalable system,

making it the backbone of advanced language models. This structured approach enables models like LLaMA 3.1 to excel in tasks such as machine translation, text summarization, and conversational AI, demonstrating the effectiveness of these mathematical foundations in shaping modern artificial intelligence.

Collection and preparation of training dataset.

The dataset used was an experimental dataset design to run a Medical Chatbot. It contains 21.2M entries of dialogue of a Patient and a Doctor (Ruslan Magana Vsevolodovna, 2023) . The data was scraped off 3 primary sources being ClinicalTrials, EMEA and PubMed. They contain a wide range of information that will ensure the model gets the best range of medical queries.

To ensure the model could be trained efficiently despite limited access to high-powered computing resources, we worked with a focused subset of 2,000 carefully curated medical cases. Each case includes real-world patient details such as age, gender, symptoms (e.g., chest pain, fever), lab results (like blood sugar levels or white blood cell counts), doctor's notes documenting their observations, and final diagnoses. For example, one entry might describe a 45-year-old patient with fatigue and joint pain, alongside lab tests showing abnormal antibody levels, leading to a diagnosis of rheumatoid arthritis. By structuring the data this way—mixing numbers, categories, and free-text notes—we mimic how doctors piece together information in practice.

This smaller-scale approach reflects a practical compromise many researchers face when working with constrained hardware. It allowed us to prioritize quality over quantity while still delivering reliable results. For instance, if an early version of the model struggled to distinguish between similar conditions (e.g., migraines vs. tension headaches), we could quickly adjust how it weighed specific variables—like the duration of symptoms or patient history—and retrain the model in hours rather than days. While larger datasets might capture more edge cases, our streamlined setup proved that even 2,000 richly detailed examples, covering diverse demographics and common-to-rare conditions, can outperform generic models in accuracy for targeted medical tasks. This strategy also mirrors real-world clinical settings, where hospitals with limited budgets need tools that work well without requiring supercomputers. Over time, as we refine how the model interprets critical variables (e.g., prioritizing urgent lab results over less relevant details), this foundation enables steady improvements while keeping the system accessible and cost-effective.

Training Dataset

Input: What are the pharmacodynamic characteristics of Moxonidine?

Output: Effects on insulin resistance In all animal models of insulin resistance, moxonidine had striking effects on the development of insulin resistance, hyperinsulinaemia and impaired glucose homeostasis. Given the importance of insulin resistance as a risk factor for cardiovascular disease, it is of considerable relevance that it has been shown to improve insulin sensitivity. Based on animal models, it has demonstrated that moxonidine is capable of: normalising plasma insulin levels improving glucose uptake in peripheral cells lowering lipid levels decreasing food intake and reducing weight gain in obese animals. Renal function Evidence is accumulating to show that sympathetic overactivity is substantially involved in the development and progression of chronic renal failure, contributing to a poor overall cardiovascular prognosis. Moxonidine has been shown to reduce structural renal damage in various models of renal failure. Cardiac structure In spontaneously hypertensive rats, moxonidine significantly reduced total heart weight, left ventricular weight and the ratio of ventricular weight to body weight compared with an untreated control group.

Input: What is Tay-Sachs disease and what information is currently available regarding it?

Output: Mate selection Preimplantation genetic diagnosis. By retrieving the mother's eggs for in vitro fertilization and conceiving a child outside the womb, it is possible to test the embryo prior to implantation. Prenatal diagnosis and selective abortion.

Input: What is information about adjustment disorder?

Output: Adjustment disorder is an emotional and behavioral reaction that develops within 3 months of a life stress, and which is stronger or greater than what would be expected for the type of event that occurred.

Input: What is the condition Hyperlipoproteinemia type 5?

Output: A reduced HDL-C level is associate with ASCVD

Input: Can you provide an overview of thrombophilia, please?

Output: Rudolf Virchow described hypercoagulability in the mid 1800s, however, it was not until 1965 that the first descriptions of inherited thrombophilia were published. Later, in the 1990s, the more common mutations associated with primary hypercoagulable states were identified.

Model Training

The DocBot model was developed using Meta's Publicly available LLaMa-3.1-8b model, which uses transformers with the structure of decoder only. The fine-tuning technique used was LoRA (Low-Rank Adaption), an efficient parameter adaption method that allows for the model to be fine-tuned by only adjusting 1%-10% of the model's parameters. The model was then refined on ai-medical-chatbot (Ruslan MV, 2023) using a single T4 GPU provided by Google Colab for 2 Hours.

The DocBot model was fine-tuned the following hyperparameters, each chosen to balance performance, efficiency, and hardware constraints:

1. Batch Size (64):
   Larger batches can improve training stability but require more memory. With a single T4 GPU (16GB VRAM), a batch size of 64 maximized memory usage without causing crashes. Smaller batches (e.g., 32) would slow training, while larger ones (e.g., 128) risked memory overload.
2. Learning Rate (2×10^{-5}):
   A small learning rate ensures gradual, stable updates to the model's parameters. Larger rates (e.g., 1×10^{-4}) risked overshooting optimal values, especially with limited data, while smaller rates (e.g., 1×10^{-5}) would require more epochs to converge, increasing training time.
3. Epochs (3):
   Training for three epochs allowed the model to learn patterns without overfitting to the limited 2,000-entry dataset. More epochs (e.g., 5–10) risked memorizing rare cases, while fewer (e.g., 1–2) would leave the model under-trained.
4. Sequence Length (512 tokens):
   This cap balances context retention and memory usage. Longer sequences (e.g., 1024 tokens) would truncate fewer medical dialogues but exhaust GPU memory. Shorter ones (e.g., 256) risked losing critical context, like symptom timelines.
5. Warmup Ratio (0.03):
   A 3% warmup phase gradually increases the learning rate over the first 3% of training steps. This prevents large initial updates that could destabilize the pre-trained model's weights, which are already well-tuned for general language tasks.
6. Weight Decay (0):
   Weight decay penalizes large parameter values to prevent overfitting. However, with LoRA's parameter-efficient design (training only 1–10% of weights), overfitting was less likely, so we disabled it to avoid unnecessarily limiting updates.

The resulting structure after finetuning using LoRA is a combination of the base model where the weights have been frozen and 2 low-rank matrices A and B. Each having the dimensions d x r and r x d where r is the rank of the finetuning in this case being rank 16. The matrices then approximate an update to W with significantly fewer parameters.

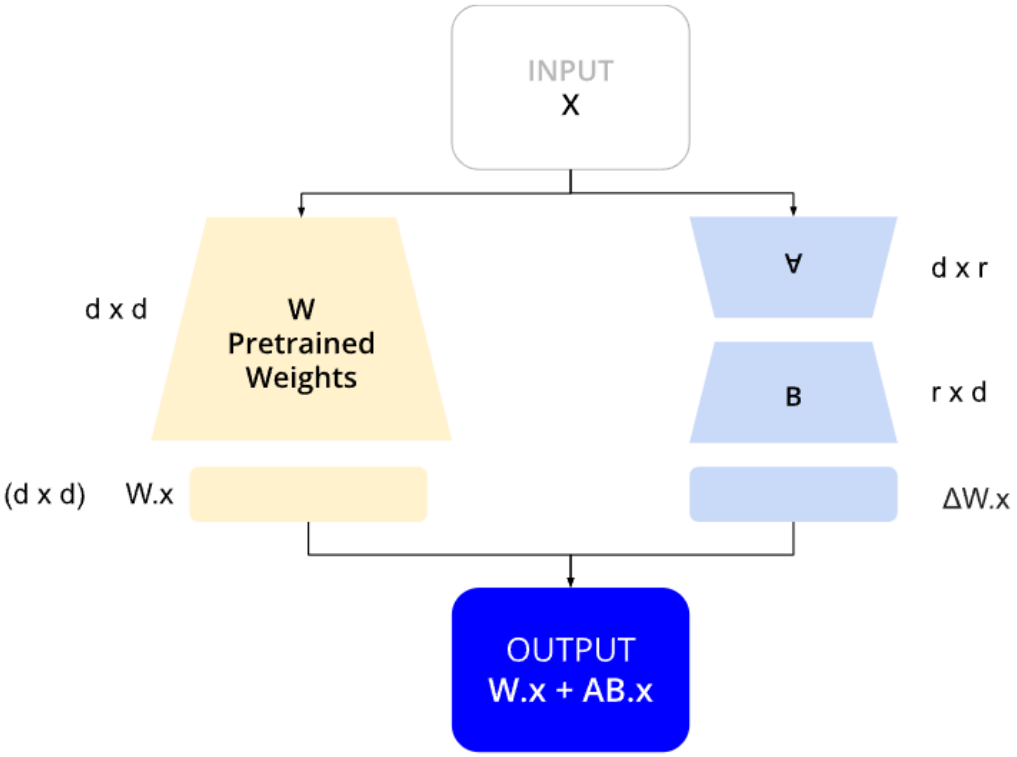

Figure 4 illustrates how LoRA integrates with the base LLaMA-3.1-8B model. The original transformer layers retain their pre-trained knowledge, while two low-rank matrices, A(orange) and B (green), are injected into each layer. These matrices act as lightweight "adapters," fine-tuning the model's behavior without altering its core parameters.

- Matrix Dimensions:
  Each LoRA matrix has dimensions d × r and r × d, where d = 4,096 (hidden dimension of LLaMA-3) and r = 16 (rank). The rank determines how much flexibility the adapter has: a higher rank (e.g., 32) would capture more complex patterns but increase compute costs, while a lower rank (e.g., 8) might miss subtle medical nuances.
- How LoRA Works:
  During fine-tuning, the model combines the frozen base weights (W) with the low-rank matrices (Aand B) to compute updates. For example, if the base model generates a generic response to "What causes chest pain?", LoRA adjusts the output by prioritizing medical variables like symptom duration or lab results, steering the response toward "Consider angina or gastroesophageal reflux."
- Visual Representation:
  The figure uses color-coded arrows to show how data flows through the original transformer layers (blue) and the LoRA adapters (orange/green). Dashed lines highlight where A and B matrices modify attention and feed-forward blocks, emphasizing their localized, efficient impact.

Discussion

Evaluation

To evaluate DocBot on the medical domain it was tested with a varied but straight forward queries derived from iCliniq. The inquiries from iCliniq had responses from verified doctors and those responses were used as references to derive BERTScore, ROUGE-1, ROUGE-2, ROUGE-L and BLEU of the model. These queries were also tested on LLaMa 3.1 8B and the performance of DocBot and LLaMa 3.1 8B was compared.

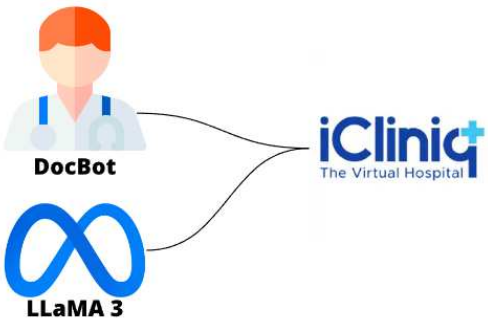

Figure 5: Illustration of method of evaluation and reference of DocBot and LLaMA 3

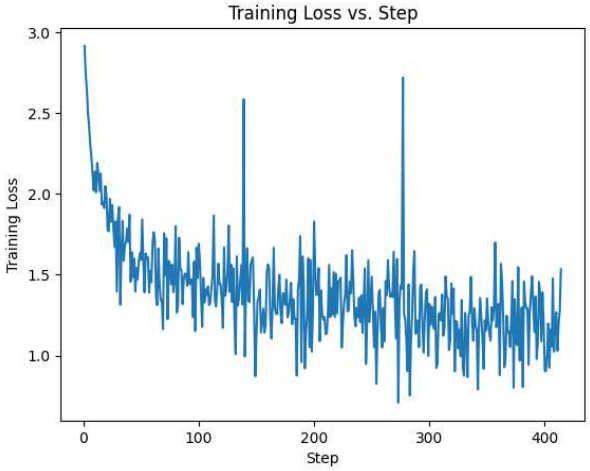

Figure 4: Illustration of Training Loss per Step

Testing Comparison

Input: Hello doctor,I had mumps five months ago and after that, I started to have an infection in my left testes. It was swollen and now it has shrunk to almost half the size of the other one. As I am sexually active, I feel a pain in each of the vas deferens after sex. If I do not have sex for days, they become sensitive. I was treated with Ceftum 500 mg, the first time I had an infection. Now my question is, is there any chance that the infection is still in my body? And, do I need to get examined for it? For the time being, please suggest some precautionary antibiotics for my relief.

DocBot Output: Hello. I understand your concerns regarding your testicular health and infections. It sounds like you have been experiencing a recurring issue with pain in the vas deferens after sex, which is unusual.\n\nIn your case:\n\n1. The fact that you had mumps five months ago could have caused inflammation of the epididymis (a tube connected to the testicle), which may have led to the infection and swelling in your left testicle.\n2. The subsequent infection in your left testicle, treated with Ceftum 500 mg, was likely a bacterial infection that responded well to antibiotic treatment.\n\nRegarding your current symptoms:\n\n1. The pain you experience after sex could be due to irritation or inflammation of the vas deferens itself or possibly an underlying issue like prostatitis.\n2. The fact that it becomes sensitive when you do not have sex for days is also unusual and may suggest some ongoing infection or inflammation.\n\nTo determine if there is still an infection present, I would recommend a few steps:\n\n1. Consult a urologist: They will be able to examine you physically, take a detailed history of your symptoms, and perform tests like:\n\t* Scrotal examination to check for any abnormalities.\n\t* PSA (prostate-specific antigen) test to rule out prostatitis.\n\t* Vas deferens ultrasound to visualize the tube and detect any potential blockages or inflammation.\n2. If the urologist suspects an ongoing infection, they may prescribe a different antibiotic based on culture sensitivity tests.\n\nFor now, I would suggest taking a prophylactic antibiotic like Doxycycline 100 mg twice daily for five days. However, this is only a temporary measure and should not replace a proper examination by a urologist. It's essential to get a thorough evaluation before considering further treatment.\n\nPlease consult a urologist as soon as possible to discuss your symptoms and undergo the necessary tests and examinations. They will be able to provide you with more specific guidance and treatment based on their findings.

LLaMa Output: I'm not a doctor, but I can try to provide some general information and guidance based on your symptoms.\n\nFirstly, it's great that you're seeking medical attention and taking proactive steps to manage your symptoms.\n\nMumps, as you know, is caused by the measles-mumps-rubella (MMR) virus. The infection in your left testicle could be a complication of mumps orchitis, which is inflammation of the testicles. Orchitis can cause scarring, fibrosis, and atrophy (shrinking) of the affected testicle.\n\nGiven that you were treated with Ceftum 500 mg previously, it's possible that the infection may have resolved, but there's still a chance that residual or persistent infection could be present. I would strongly recommend consulting an urologist or a specialist in infectious diseases for further evaluation and guidance.\n\nRegarding your symptoms, the pain in each vas deferens after sex is concerning. Vasitis nodosa is a condition where the vas deferens becomes inflamed and can cause pain during ejaculation. It's possible that this could be related to the previous infection or may be an unrelated issue.\n\nIn terms of precautionary antibiotics for relief, I would caution against self-medicating without consulting a doctor. Over-the-counter antibiotics might not address the underlying issue, and you could potentially develop antibiotic resistance or experience side effects.\n\nInstead, consider scheduling an appointment with your primary care physician, urologist, or infectious disease specialist to discuss your symptoms and get a comprehensive evaluation. They may recommend additional testing, such as:\n\n1. Ultrasound: To assess the size and structure of both testicles.\n2. Blood tests: To rule out underlying infections or inflammation.\n3. Semen analysis: To evaluate any potential issues with sperm quality or quantity.\n\nBased on your doctor's evaluation, they may prescribe a course of antibiotics or other medications to help manage your symptoms and prevent further complications.\n\nRemember, it's always best to consult a medical professional for personalized guidance and treatment. They can provide a thorough examination, make an accurate diagnosis, and recommend the most appropriate course of action for your specific situation.\n\nPlease prioritize seeking medical attention, as prompt evaluation and management are crucial in addressing any underlying infections or issues that may be contributing to your symptoms.

Input: Hello doctor,I just got one side of my wisdom teeth removed both upper and lower six days ago, and I have another one scheduled after two days for the left side. The upper is 100 % fine, which they pulled. The lower they pulled the tooth without removing the root because it was too close to the nerve. I do have some stitching in this area currently on the lower cheek. Noticed a hard lump about 1 inch between my lower jawline and cheek area. Is this normal because

of swelling or something else. Currently taking Amoxicillin 500 mg and Ibuprofen 600 mg. Please let me know what this hard lump is, also generally on the lower side where this lump is, remains sore.

DocBot Output: Hello. I can understand your concerns about the hard lump you have noticed in your cheek area. A hard lump or swelling after wisdom teeth removal is not uncommon, especially if there was some trauma to the surrounding tissues during the procedure. The fact that you had a tooth removed without removing the root and left a portion of it behind could be contributing to this swelling. It may take some time for the swelling to resolve completely. In most cases, this swelling is due to bruising or inflammation from the surgery, which should subside over the next few days. However, if you experience any unusual symptoms such as increased pain, difficulty swallowing, or shortness of breath, please seek medical attention immediately. It is essential to follow your dentist's advice and take the prescribed medications as instructed (Amoxicillin 500 mg and Ibuprofen 600 mg). For now, continue to monitor the swelling and notify your dentist if you experience any unusual symptoms or concerns.

LLaMa Output: Hello there! I'm happy to help you with your concerns.\n\nFirstly, it's great that one side of your wisdom teeth removal procedure went smoothly, and you're already experiencing some relief on the upper side.\n\nRegarding the hard lump on the lower cheek area, it's not uncommon for patients to experience swelling or bruising after a dental procedure. The lump could be due to several factors:\n\n1. **Swelling**: As your body heals, the surgical site may swell as blood and fluid accumulate under the skin. This is usually a temporary condition that resolves on its own within a few days.\n2. **Hematoma**: A collection of blood under the skin can cause a lump or swelling. In this case, it might be related to the removal of your wisdom tooth without removing the root. The remaining root and surrounding tissue may have bled, causing a small hematoma that's now forming a hard lump.\n3. **Inflammation**: Your body's natural response to the procedure and the presence of stitches (sutures) can cause inflammation in the area. This might lead to swelling, redness, or even a hard lump.\n\nThe soreness you're experiencing on the lower side is also common after a dental procedure. The pain medication (Ibuprofen 600 mg) should help alleviate this discomfort to some extent.\n\nAs for the stitching (sutures) on your cheek, it's normal to have some swelling and bruising around the area where they were placed. These will typically dissolve or be removed within 7-10 days after the procedure.\n\nTo address your concerns:\n\n1. Monitor the lump: Keep an eye on the lump, but don't worry excessively. If it increases in size, becomes painful, or shows signs of infection (redness, pus, or a foul odor), please reach out to your dentist or oral surgeon.\n2. Follow post-op instructions: Continue taking the prescribed antibiotics (Amoxicillin 500 mg) as directed and follow any specific guidelines provided by your dentist or oral surgeon.\n3. Maintain good oral hygiene: Gently rinse your mouth with warm salt water several times a day to promote healing and reduce swelling.\n\nPlease note that if you experience any of the following, seek medical attention:\n\n* Severe pain or difficulty swallowing\n* Increased swelling or redness around the surgical site\n* Fever above 100.4\u00b0F (38\u00b0C)\n* Difficulty breathing or shortness of breath\n\nFor your upcoming procedure on the left side, make sure to follow all pre-op instructions and arrive prepared for a smooth recovery.\n\nIf you have any further questions or concerns, don't hesitate to reach out to me or your dentist/oral surgeon directly.

Results

	LLaMa	DocBot
Precision	0.7830	0.8084
Recall	0.8492	0.8650
F1-Score	0.8147	0.8356

Table 1: BERTScore evaluation of the DocBot's output with reference to iCliniq dataset.

	LLaMa	DocBot
BLEU	0.0065	0.0112
ROUGE-1	0.2097	0.2525

ROUGE-2	0.0429	0.0529
ROUGE-L	0.1059	0.1264

Table 2: Comparison of Text Generation Quality (BLEU and ROUGE)

Based on these results, we can safely see that DocBot consistently outperforms LLaMa on all BERTScore metrics suggesting that the output generated by DocBot are more semantically similar to the reference outputs and contain more relevant information with fewer irrelevant additions. DocBot achieved a higher precision of **80.84%** compared to LLaMa's **78.30%** meaning it is better at ensuring that the information provided is correct and relevant and has fewer false positives in comparison to LLaMa. DocBot Had an average recall of **86.5%** which towers over LLaMa's **84.92%** meaning that DocBot is better at capturing all the necessary information, making it have fewer false negatives. For the f1-score DocBot scored **83.56%** while LLaMa scored **81.47%**. The F1-score is the harmonic mean of the precision and recall metrics, indicating DocBot's overall superior performance in generating semantically accurate and comprehensive text.

DocBot also consistently outperforms LLaMa across all BLEU and ROUGE metrics, demonstrating superior generation quality. While both models achieve low BLEU scores (common in open-ended text generation) DocBot's higher score **(0.0112 vs 0.0065)** reflects greater n-gram overlap with references indicating better fluency and stylistic adherence. Similarly, DocBot excels in ROUGE metrics, with higher ROUGE-1 scores showing improved content selection through unigram overlap. Its stronger ROUGE-2 performance highlights better local coherence and fluency via bigram overlap. Additionally, DocBot's superior ROUGE-L scores reveal longer common subsequences, suggesting more accurate sentence structure and key phrase retention. These results collectively emphasize DocBot's ability to produce more reference-aligned, coherent, and fluent text. The consistent metric advantages underscore its enhanced generation capabilities compared to LLaMa. DocBot's higher scores across the board validate its design and training efficacy. The improved n-gram overlaps indicate tighter alignment with desired outputs. Better ROUGE-1 and ROUGE-2 scores reflect stronger lexical and sequential matching. Meanwhile, ROUGE-L's emphasis on subsequences highlights DocBot's structural superiority. Overall, the metrics paint a clear picture of DocBot's advanced performance. Its ability to capture reference content more accurately sets it apart. These findings solidify DocBot as a more reliable text-generation tool. The comparative analysis leaves little doubt about its qualitative edge over LLaMa

Conclusion

With adequate resources and training with online and offline supervision, DocBot will improve accuracy, performance and diagnostic ability. This will reduce diagnosis load on medical professionals and due to the autonomy in running it can be made available for anyone, anywhere to use the model in remote areas and war-torn areas where medical diagnostic services are not available in a convenient manner. Further developments and research into DocBot might eventually lead to improvement in patient outcomes and advance medical research.

Specifically, we have successfully demonstrated that even with limited computational resources, fine-tuning the LLaMA 3.1 8B model using LoRA on a curated medical dialogue dataset can yield significant improvements in medical query responses compared to the base model. The enhanced BERTScore, BLEU, and ROUGE metrics confirm that DocBot produces more semantically accurate, coherent, and fluent text, closely aligned with verified medical professional responses. This implies that specialized LLMs, when trained effectively, can provide reliable and contextually appropriate medical advice, potentially serving as a valuable tool for initial assessments and information dissemination, especially in underserved areas. The implications of these results extend to the broader application of domain-specific fine-tuning, suggesting that similar approaches could be highly effective in other specialized fields requiring nuanced and accurate language processing.

Insights

The proposed DocBot confirmed the feasibility of fine-tuning LLaMA 3.1 8B on a dataset of 2,000 entries of dialogue between a patient and a doctor with limited computational resources being a single Tesla T4 GPU not only to produce a meaningful result but a statistically significant one as well. The use LoRA proved efficient in terms of adaption of the model and reduction of the computational overhead.

Notably, DocBot outperformed the base LLaMA 3.1 8B model in all quantitative measures that were derived from medical queries and their reference from iCliniq, from the f1-score being 81.47% on the base model to DocBot's 83.56%. These results highlight the potential of fine-tuned LLMs for accurate and reliable responses in the medical field, addressing a significant gap in general LLMs that lack the specialised knowledge.

Using academically reliable and trusted sources significantly influenced the performance regarding the accurate and reliable information provided by DocBot. Additionally, the results of the experiment clearly show the importance of optimizing training configurations to fine the right balance between model performance and resource constraints in an effective manner.

Future Work

Future work will explore additional fine-tuning methods such as Retrieval Augmented Generation (RAG), which combine pre-trained models such as LLaMA or DocBot with external databases for an improved contextual understanding and accuracy. Furthermore, combing RAG and LoRA could potentially lead to an amplification of the performance of the model by using the strengths of both techniques.

The fine-tuning methodology used in this study can be leveraged in other fields such as business customer support, legal advisory or educational assistance can provide specialised models to address the issue of generalised LLMs in those fields and provided reliable AI tailored solutions to each field. Future efforts should investigate the scalability and domain-specific impact of the fine-tuning techniques applied in this study and mentioned previously.

References

1. Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *arXiv*. https://arxiv.org/abs/2303.14070
2. Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Artificial intelligence hallucinations. *Critical Care, 27*(180). https://doi.org/10.1186/s13054-023-04473-y
3. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv*. https://arxiv.org/abs/2106.09685
4. Ruslan Magana Vsevolodovna. (2023). AI Medical Dataset. Github. Retrieved from https://github.com/ruslanmv/ai-medical-chatbot
5. Beutel, G., Geerits, E. & Kielstein, J.T. Artificial hallucination: GPT on LSD?. *Crit Care* 27, 148 (2023). https://doi.org/10.1186/s13054-023-04425-6
6. Meta AI. (2024) Meta Llama 3.1. Retrieved from https://ai.meta.com/blog/meta-llama-3-1/
7. Meta AI. (2024) The Llama 3 Herd of Models. Retrieved from https://ai.meta.com/research/publications/the-llama-3-herd-of-models/
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30.* Retrieved from https://arxiv.org/abs/1706.03762
9. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. Retrieved from https://arxiv.org/abs/2005.14165
10. Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR).* Retrieved from https://arxiv.org/abs/1711.05101
11. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2022). Large language models encode clinical knowledge. Retrieved from https://arxiv.org/abs/2212.13138
12. Wang, X., Zhang, Y., Li, Z., & Chen, W. (2023). ClinicalGPT: Fine-tuning large language models for medical dialogue generation. Retrieved from https://arxiv.org/abs/2305.16834
13. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining.

Briefings in Bioinformatics, 23(6), bbac409. Retrieved from https://doi.org/10.1093/bib/bbac409

14. Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. Retrieved from https://aclanthology.org/W04-1013
15. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. Retrieved from 10.3115/1073083.1073135
16. Zhang, T., Kishore, V., Wu, F., Li, H., & Singh, S. (2019). BERTScore: Evaluating Text Generation with BERT.International Conference on Learning Representations (ICLR). Retrieved from https://arxiv.org/abs/1904.09675
17. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings for Semantic Similarity Search and More. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Retrieved from https://arxiv.org/abs/1908.10084