Efficient Antihallucinogenic AI for Tropical Medicine: A Probabilistic Framework for Differential Diagnosis

S. M. Sakeef Sani, Md. Shaown Miah, Taufiq Hasan

mHealth Lab, Department of Biomedical Engineering
Bangladesh University of Enginering and Technology (BUET), Dhaka - 1205, Bangladesh.
Email: taufiq@bme.buet.ac.bd

Abstract

Clinical decision-making, particularly in the context of differential diagnosis in low-resource healthcare settings, poses significant challenges due to the complexity and variety of symptoms presented by patients and the lack of skilled doctors. This study introduces `mLabLLM`, a fine-tuned adaptation of the LLaMA 3.2 3B model designed to enhance clinical decision-making in differential diagnosis. Leveraging domain-specific datasets, including a curated tropical diseases dataset including Dengue, malaria, and chikungunya - prevalent health challenges in South Asian countries and employing optimization techniques like Low-Rank Adaptation (LoRA) and pruning to reduce computational overhead. The model achieves greater efficiency without compromising performance. A probabilistic framework integrates symptom-disease frequencies with Bayesian reasoning, enabling dynamic ranking of diagnoses during patient interactions. Experimental results show that `mLabLLM` significantly outperforms baseline models, achieving an 82.8 % Top-3 accuracy in differential diagnosis, compared to 75.1% for `Phi-3-128k` and 72.4% for `LLaMA 3.2 3B`, positioning it as a scalable and practical solution for real-world clinical applications.

1 Introduction

Accurate and efficient differential diagnosis is a fundamental aspect of medical practice, particularly in resource-limited settings where the availability of skilled professionals and healthcare infrastructure is often restricted. In these environments, clinicians must rely heavily on their experience and available resources to make timely and accurate diagnoses. In addition, when patient burdens are high, primary care physicians may be able to spend only a few minutes discussing the symptoms of the patient (Irving et al. 2017). Artificial intelligence (AI) tools can be of significant benefit in diagnostic assistance in such scenarios if patient symptoms can be processed before the doctor's appointment.

Recent advancements in deep learning models have shown considerable promise in enhancing diagnostic accuracy by aiding clinicians in synthesizing complex patient data. Large language models (LLMs) such as GPT-4 (OpenAI 2023), LLaMA (Touvron et al. 2023), PaLM (Chowdhery and et al. 2022), Gopher (Rae and et al. 2021), and OPT (Zhang and et al. 2022) have emerged as powerful tools, capable of processing vast amounts of medical literature and patient data, offering valuable insights for clinical decision-making. However, these general-purpose models often lack domain-specific training, limiting their effectiveness in navigating the complexities of medical diagnoses in diverse clinical scenarios. Domain-specific models, such as `Med-PaLM` (Carr et al. 2022) and `BioMedLM` (Lee et al. 2021), address this gap by specializing in medical tasks, e.g., question answering and literature mining. However, their computationally expensive nature renders them less practical for resource-constrained environments, where accessibility and scalability are critical. This limitation is particularly evident in low and middle-income countries (LMICs) like Bangladesh, which experienced over 321,000 reported dengue cases during its largest outbreak in 2023, overwhelming its healthcare system and highlighting gaps in surveillance and rural healthcare infrastructure (World Health Organization (WHO) 2024; PLOS Neglected Tropical Diseases 2023). Additionally, approximately 2.4 million annual dengue cases remain underreported due to limitations in diagnostic resources and cultural barriers (Tropical Medicine and Health 2022). These challenges emphasize the need for efficient, domain-specific AI solutions tailored to LMIC settings.

This work presents `mLabLLM`, a fine-tuned version of the `LLaMA 3.2 3B` model, designed to address the challenges of differential diagnosis in resource-constrained healthcare systems. By leveraging domain-specific datasets and optimization techniques, `mLabLLM` is tailored to operate effectively within the unique constraints of real-world medical environments. This study makes the following key contributions to the field of clinical AI and differential diagnosis:

- Domain Adaptation: We use custom dataset focused on tropical diseases alongside publicly available biomedical resources such as PubMed abstracts (PubMed 2022), ensuring broad coverage of health conditions that are particularly relevant in tropical countries and resource-limited regions.
- Efficient Fine-Tuning: To optimize the model's computational efficiency, `mLabLLM` leverages Low-Rank Adaptation (LoRA) (Hu et al. 2021), a technique that reduces the computational cost while maintaining high performance in domain-specific tasks. In contrast, traditional

models often require extensive computational resources.

- Dynamic Probabilistic Diagnosis Framework: We propose a Bayesian-based ranking mechanism that functions on the model's output, enabling it to prioritize potential diagnoses in real-time based on patient-reported symptoms and prior knowledge extracted from annotated medical datasets (van Doorn et al. 2020). This framework facilitates dynamic decision-making and supports clinicians in generating a list of probable diagnoses that are continuously updated as new information becomes available.

2 Dataset Creation and Preprocessing

The effectiveness of a medical LLM heavily depends on the quality and relevance of the dataset used for fine-tuning. For the purpose of this study, we curated a custom tropical diseases dataset, augmented with publicly available biomedical datasets, to address the specific needs of differential diagnosis in resource-constrained settings.

2.1 Tropical Diseases Dataset

This dataset was designed to improve the diagnostic capabilities of the model by providing domain-specific data, thus enhancing its ability to understand and process medical information in these contexts.

Sources: The custom tropical diseases dataset was derived from a combination of authoritative medical resources to ensure its comprehensiveness and reliability. Key sources include relevant books (Kasper et al. 2020),(Cahill 2011), (Rothe 2020),(Meunier et al. 2013) and guidelines from the World Health Organization (WHO), which offer trusted insights into the pathophysiology, diagnosis, and treatment of tropical diseases. Additionally, peer-reviewed research articles and publications sourced from open-access platforms, including PubMed Central, were integrated, allowing for the inclusion of the latest clinical findings and evidence-based practices. This combination of established textbooks and up-to-date research ensures that the dataset covers a broad spectrum of diseases and clinical scenarios relevant to tropical medicine, particularly in resource-constrained settings.

Content and Structure: The dataset encompasses both structured and unstructured data, covering essential aspects of medical knowledge required for accurate diagnosis and treatment in tropical medicine. The structured data includes detailed information on common symptoms associated with tropical diseases, such as fever, rash, and joint pain. It also includes diagnostic data, outlining the relevant diagnostic tools and procedures used to identify these diseases, such as imaging results and laboratory tests. The treatment data features commonly recommended regimens, including antimicrobial drugs and supportive therapies, tailored to manage tropical diseases. Additionally, the dataset contains annotated real-world case studies, providing comprehensive examples of conditions, symptoms, diagnostic approaches, treatments, and outcomes, offering valuable insights into practical, clinical decision-making.

For instance, a typical entry in the dataset might look like:

> "Disease: Malaria → Symptoms: Fever, chills, headache, fatigue → Diagnostics: Blood smear, rapid diagnostic test → Treatment: Artemisinin-based combination therapy (ACT)."

2.2 Additional Datasets

In addition to the tropical diseases dataset, several other relevant datasets were incorporated to further enhance our model training. The *PubMed Abstracts* dataset (PubMed 2022), which consists of biomedical literature abstracts from PubMed, provides valuable clinical insights and background information on a wide range of diseases and medical conditions. To improve the model's ability to handle complex medical reasoning tasks, the *USMLE Dataset* was included, which contains a collection of question-answer pairs from the United States Medical Licensing Examination (USMLE). Furthermore, the *MedQA Dataset* (Jin et al. 2020), a domain-specific question-answer set, was utilized to strengthen the model's performance in answering medical questions.

2.3 Preprocessing Steps

Several preprocessing steps were undertaken to ensure the compatibility and effectiveness of the datasets for model fine-tuning. Firstly, *Text Cleaning*, involved removing irrelevant noise and ensuring consistency across the dataset by eliminating special characters, redundant spaces, and standardizing medical abbreviations and terminology. This was followed by *Tokenization*, where the datasets were processed using the LLaMA tokenizer, which divides the text into smaller subword tokens. This ensures that the data aligns with the model's architecture and facilitates efficient processing during training. To enhance the model's understanding of key medical concepts, *Named Entity Recognition (NER)* was performed using `SciSpacy` (Neumann et al. 2019). This tool annotated important entities such as symptoms, diseases, and treatments within the text. For example, in the sentence, "The patient reported fever and joint pain," the terms "fever" and "joint pain" were identified as symptoms. Finally, to address the 128k-token limit of the LLaMA model, *Text Chunking* was applied to divide longer paragraphs into manageable chunks, ensuring that the model can process larger contexts without truncation while preserving semantic coherence.

2.4 Dataset Statistics

The final dataset for fine-tuning and evaluation comprised key components including 50,000 custom-annotated entries from the Tropical Diseases Dataset. Additionally, the dataset included 1,000 labeled, 61,200 unlabeled, and 211,300 artificially generated entries from PubMedQA, along with approximately 60,000 QA pairs from USMLE and MedQA. A brief overview of the data is shown in Table 1.

Dataset	Entries
Tropical Diseases Dataset (custom annotated)	50,000
PubMedQA (PQA-Labeled)	1,000
PubMedQA (PQA-Unlabeled)	61,200
PubMedQA (PQA-Artificial)	211,300
USMLE and MedQA (QA Pairs)	60,000

Table 1: Dataset statistics for fine-tuning and evaluation

3 Model Fine-Tuning

The `mLab_LLM` model was fine-tuned on medical datasets to enhance its ability to perform clinical reasoning and differential diagnosis tasks. The fine-tuning process incorporated *Low-Rank Adaptation (LoRA)* to reduce computational complexity and memory requirements while maintaining high accuracy. Pruning and quantization further optimized the model for deployment in resource-constrained environments. Figure 1 provides a general overview.

3.1 Low-Rank Adaptation (LoRA)

Fine-tuning all weights of a large model such as `LLaMA 3.2 3B` is computationally expensive. LoRA provides a parameter-efficient alternative by introducing low-rank updates to specific model layers while keeping the pre-trained weights frozen.

Formulation: LoRA modifies the weight matrix W in attention layers. Instead of updating W directly, LoRA introduces a low-rank decomposition:

$$W' = W + \Delta W, \quad \Delta W = A \cdot B^T \tag{1}$$

where $W \in R^{d\times k}$ is original pre-trained weight matrix. $A \in R^{d\times r}$ and $B \in R^{k\times r}$ are Low-rank matrices, with rank $r \ll \min(d, k)$. The forward pass in attention computation is given by:

$$h' = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V + A \cdot B^T \tag{2}$$

where Q, K, V are Query, Key, and Value matrices respectively. d_k is Dimensionality of Q. This decomposition significantly reduces the number of trainable parameters as follows:

$$\text{Trainable Parameters} = r \cdot (d + k) \tag{3}$$

As an example, with $d = 4096$, $k = 4096$, and $r = 8$, the number of trainable parameters is reduced by a factor of approximately 256.

Implementation: The Low-Rank Adaptation (LoRA) technique was implemented by applying LoRA layers to the Query, Key, and Value (QKV) projection matrices within the attention layers. The hyperparameters were carefully selected to balance performance and efficiency. Specifically, the rank r was set to 8, which provided an optimal trade-off between accuracy and computational efficiency. The learning rate for the LoRA parameters was set to 5×10^{-5}, while the batch size was set to 128, achieved through gradient accumulation.

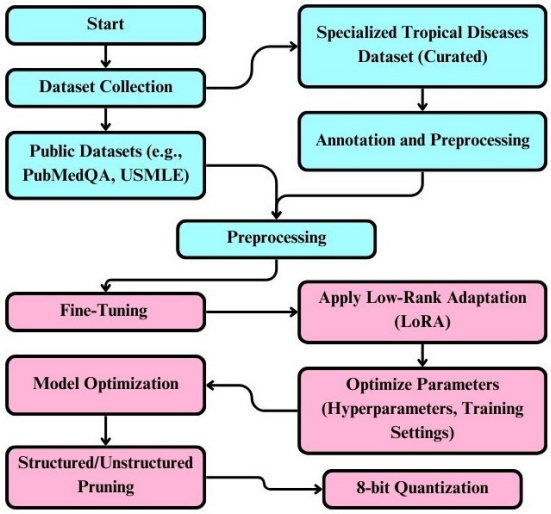

Figure 1: Workflow for data preprocessing and model fine-Tuning.

3.2 Pruning for Optimization

Two pruning strategies were applied in this work. *Structured pruning* involved removing entire attention heads and neurons in feedforward layers based on their contribution to the loss function. The importance of each weight was determined using the following score:

$$\text{Score}(w_{ij}) = \left|\frac{\partial \mathcal{L}}{\partial w_{ij}} w_{ij}\right| \tag{4}$$

Weights with scores below a predefined threshold τ were pruned (He and Xiao 2023). On the other hand, *unstructured pruning* focused on zeroing out individual weights whose magnitudes fell below a specified threshold. This process was carried out iteratively, removing 20%-30% of the parameters, followed by retraining the model to recover its performance. In the *Post-Pruning Retraining* step, the model was fine-tuned using cross-entropy loss to adjust the remaining parameters and maintain accuracy:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{5}$$

where y_i is the true label, and $\hat{y}_i$ is the predicted probability for the i-th token. Finally, a quantization step reduces the precision of model weights to minimize memory usage. An 8-bit quantization scheme was applied:

$$W_{\text{quant}} = \text{round}(W \cdot 2^b)/2^b, \quad b = 8 \tag{6}$$

This approach preserved accuracy while reducing memory consumption by approximately 75%.

4 Differential Diagnosis Framework

A robust differential diagnosis system must iteratively collect and analyze patient information to guide clinical reasoning. `mLab_LLM` employs a dynamic questioning framework,

integrating patient responses with biomedical literature to prioritize potential diagnoses (See Figure 2). The process involves following steps: (1) Iterative Questioning System which engages users in a structured conversation to gather clinical details, with questions dynamically generated based on prior responses and extracted keywords. The process begins with broad, open-ended questions to establish context. For example:

> the system might start with the question: *"What brings you in today?"*
> As the conversation progresses, the model generates more specific follow-up questions based on the symptoms the user reports. For instance, if a user responds with:
> *"I have a fever and feel tired."*
> the model may generate a targeted follow-up question such as:
> *"Have you recently traveled to a tropical region?"*.

This iterative questioning approach allows the model to progressively narrow down the clinical context and gather relevant information to assist in diagnosis. (2) Conversation Continuity i.e. Each response updates a structured **Patient Note**, which is fed back into the model for subsequent queries. (3) Medical Keyword Extraction To enhance the model's reasoning capabilities, relevant medical terms are extracted from user responses and literature. (4) Medical Named Entity Recognition (NER) is performed using tools like *SciSpacy*, extracting terms such as `<symptom>`, `<disease>`, `<treatment>`, and `<risk_factor>`. e.g.

> Input: *"I have a fever and rash."*
> Output: `"fever": <symptom>, "rash": <symptom>`

(5) PubMed Integration: The extracted terms are used to query biomedical literature via PyMed:

> Example Query: *"fever rash tropical disease differential diagnosis"*

The returned abstracts are further processed to extract additional keywords using `SciBERT` (Beltagy, Lo, and Cohan 2019). (6) Symptom-Disease Accumulation: The model maintains a frequency dictionary that combines terms from patient responses and PubMed searches. (7) Construction of a Frequency Dictionary which is a weighted accumulation formula integrates terms:

$$F(t+1) = F(t) + w_1 \cdot F_{\text{response}} + w_2 \cdot F_{\text{pubmed}} \quad (7)$$

where $F(t)$ is the current frequency dictionary, F_{response} are terms extracted from user input ($w_1 = 1.0$), and F_{pubmed} are terms extracted from PubMed abstracts ($w_2 = 0.5$). One example workflow is given below:

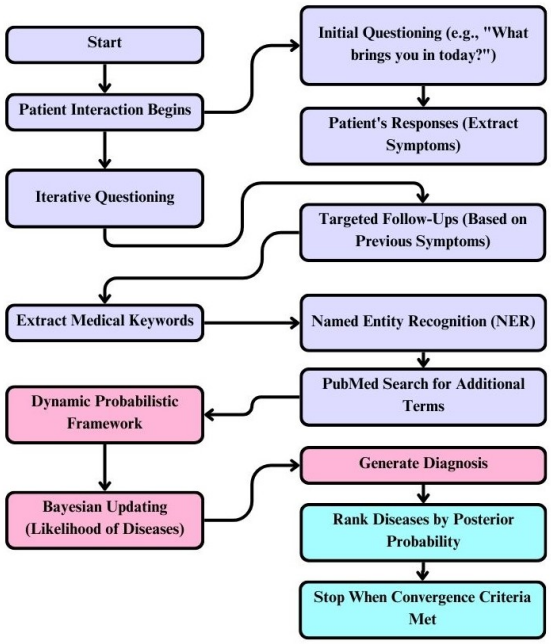

Figure 2: Workflow for Patient Interaction and Differential Diagnosis.

- step 1: User reports fever and fatigue.
  Extracted: `"fever", "fatigue"`.
- step 2: PubMed search adds terms like *"malaria"* and *"dengue"*.
  Updated Frequency:
  `"fever": 1.0, "fatigue": 1.0, "malaria": 0.5, "dengue": 0.5`.

This iterative accumulation informs the probabilistic ranking described in Section 5.

5 Probabilistic Symptom-Disease Ranking

The model employs a probabilistic framework to rank potential diagnoses based on observed symptoms and prior knowledge. This system dynamically integrates patient-reported symptoms, biomedical literature, and clinical knowledge using Bayesian reasoning and frequency-based weighting.

5.1 Bayesian Symptom-Disease Model

The ranking process begins with a prior probability distribution over possible diseases, updated iteratively as symptoms are observed (Friedman, Geiger, and Goldszmidt 1997). Initial disease probabilities $P(\text{disease}_j)$ (aka prior probability) are derived from epidemiological data and prevalence rates:

$$P(\text{disease}_j) = \frac{\text{prevalence}_j}{\sum_{k=1}^{N} \text{prevalence}_k} \quad (8)$$

where prevalence_j is Prevalence of disease j and N is total number of diseases considered. **The likelihood function**

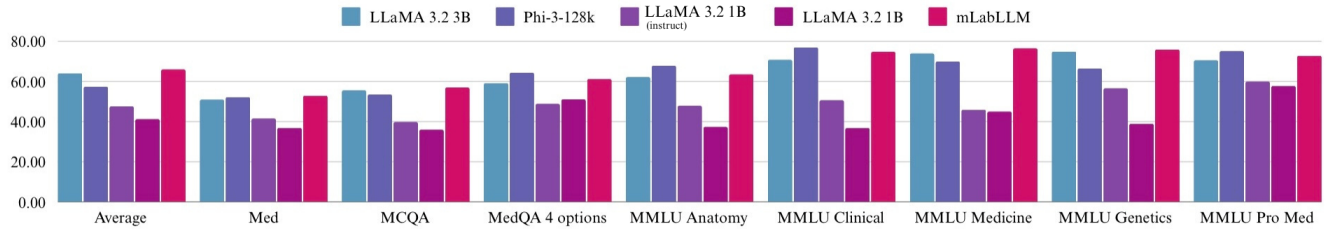

Figure 3: Comparison of mLabLLM with similar models. mLabLLM outperforms all SOTA LLMs across most metrics.

$P(\text{symptoms}|\text{disease}_j)$ quantifies the probability of observing the given symptoms for each disease. This is modeled as:

$$P(\text{symptoms}|\text{disease}_j) = \prod_{i=1}^{M} P(\text{symptom}_i|\text{disease}_j) \quad (9)$$

where M is Number of observed symptoms and $P(\text{symptom}_i|\text{disease}_j)$ is Retrieved from annotated datasets or medical literature. Using Bayes' rule, the posterior probability of disease j is given by:

$$P(\text{disease}_j|\text{symptoms}) \propto P(\text{symptoms}|\text{disease}_j) \cdot P(\text{disease}_j)$$

5.2 Symptom Frequency Integration

The Bayesian probabilities are combined with the symptom frequency dictionary $F(t)$ to refine diagnosis rankings. The relevance of each disease is measured by **Weighted Symptom Scores** defined as:

$$S(\text{disease}_j) = \sum_{i=1}^{M} P(\text{symptom}_i|\text{disease}_j) \cdot F(\text{symptom}_i)$$

where $F(\text{symptom}_i)$ is Weighted frequency of symptom i in the dictionary.

5.3 Stopping Criterion

The system stops questioning when the top-ranked disease scores stabilize. Convergence is defined as:

$$\max_j |S_t(\text{disease}_j) - S_{t-1}(\text{disease}_j)| < \epsilon \quad (10)$$

where $S_t(\text{disease}_j)$ is the score at iteration t and ϵ is an infinitesimal number (e.g. 0.01). Alternatively, a maximum question limit can terminate the process.

6 Results and Benchmarks

The performance of `mLabLLM` was evaluated on various clinical reasoning and diagnostic tasks using both public datasets (See Figure 3). The model demonstrated superior results compared to baseline models, including `LLaMA 3.2 3B` and `Phi-3-128k`.

6.1 Evaluation Metrics

The evaluation metrics used to assess the model's performance are outlined below, with associated mathematical formulas for each task (See Table 2). For the question-answering (QA) tasks, the primary metrics include:

Accuracy: This metric calculates the percentage of correct answers in multiple-choice questions. It is expressed mathematically as:

$$\text{Accuracy} = \frac{\text{Number of Correct Answers}}{\text{Total Number of Questions}} \times 100 \quad (11)$$

Exact Match (EM): This measures the proportion of exact matches between the predicted and true answers. It is defined as:

$$\text{EM} = \frac{\text{Number of Exact Matches}}{\text{Total Number of Questions}} \times 100 \quad (12)$$

For **Summarization tasks**, the metrics used are:

ROUGE-L: This metric measures the overlap of predicted and reference summaries, focusing on the longest common subsequence. The ROUGE-L score is computed as:

$$\text{ROUGE-L} = \frac{\sum_{i=1}^{n} \text{LCS}_i}{\sum_{i=1}^{n} \text{Reference Length}_i} \quad (13)$$

where LCS_i denotes the longest common subsequence for a given summary, and $\text{Reference Length}_i$ is the length of the reference summary (Lin 2004).

BLEU: This evaluates the quality of generated text based on n-gram matches. The BLEU score is given by:

$$\text{BLEU} = \min\left(1, \frac{\text{Candidate n-grams}}{\text{Reference n-grams}}\right) \times \exp\left(\sum_{n=1}^{N} \log p_n\right) \quad (14)$$

where p_n is the precision for n-grams of length n, and N denotes the maximum n-gram length (Papineni et al. 2001).

For **NER and Classification tasks**, the metric used is:

F1 Score: This is the harmonic mean of precision and recall, and it is defined as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (15)$$

where Precision is the number of true positive predictions divided by the number of true positive plus false positive predictions, and Recall is the number of true positive predictions divided by the number of true positive plus false negative predictions.

For **Differential Diagnosis tasks**, the metrics are:

Top-3 Accuracy: This measures whether the correct diagnosis is among the top three predicted diagnoses (Lee et al. 2016). It is calculated as:

$$\text{Top-3 Accuracy} = \frac{\text{Number of Correct Diagnoses in Top-3}}{\text{Total Number of Diagnoses}} \times 100 \quad (16)$$

Models	QA and Reasoning		Differential Diagnosis		Summarization	
	USMLE Accuracy (%)	MedQA Accuracy (%)	Top-3 Accuracy (%)	Symptom Matching F1	ROUGE-L (%)	BLEU (%)
LLaMA 3.2 3B	55.0	51.5	72.4	0.68	42.1	38.5
Phi-3-128k	57.8	54.3	75.1	0.72	44.0	40.2
mLabLLM	**65.2**	**62.1**	**82.8**	**0.79**	**49.3**	**45.6**

Table 2: Performance Comparison Across Tasks: QA and Reasoning, Differential Diagnosis, and Summarization

Symptom Matching F1: This evaluates the model's ability to correctly associate symptoms with diseases. The Symptom Matching F1 is calculated as:

$$F1_{\text{Symptom Matching}} = 2 \times \frac{\text{Precision}_{\text{Symptom}} \times \text{Recall}_{\text{Symptom}}}{\text{Precision}_{\text{Symptom}} + \text{Recall}_{\text{Symptom}}} \quad (17)$$

These evaluation metrics provide a comprehensive assessment of the model's performance across various tasks, ensuring that it can handle different aspects of medical reasoning, summarization, and classification effectively.

6.2 Benchmark Datasets

The following datasets were used for evaluation. The USMLE dataset consists of simulated board exam question-answer pairs, which test the model's ability to perform medical reasoning and decision-making. The MedQA dataset focuses on medical domain question-answer tasks, enhancing the model's performance in answering domain-specific questions. Additionally, a custom Tropical Diseases dataset was created, designed for differential diagnosis and summarization tasks, providing context-specific challenges related to tropical diseases and their clinical management.

6.3 Case Study: Tropical Disease Diagnosis

Scenario: The model is tasked with diagnosing a tropical disease based on the patient's symptoms.

> The input provided by the clinician :
> *"Patient presents with fever, rash, and recent travel to a tropical region."*
> Based on this input, the model generated the following predictions:
> *Dengue: 1.38 (score)*, *Malaria: 1.36 (score)*, and *Chikungunya: 1.21 (score)*.

The model accurately prioritized *Dengue* as the most likely diagnosis, based on high posterior probabilities and symptom frequencies derived from both the input and the model's learned epidemiological data.

6.4 Ablation Study

An ablation study was conducted to evaluate the impact of key components on model performance. These results highlight the importance of both LoRA and the tropical diseases dataset in improving the model's overall performance.

Component Removed	Performance Impact
LoRA	6.7% reduction in Top-3 accuracy
Tropical Diseases Dataset	8.4% decrease in QA accuracy

Table 3: Impact of Key Components on Model Performance

7 Discussion

In this study, we developed a model for tropical disease diagnosis by integrating symptom-based inputs and epidemiological data. The results from the ablation study revealed the critical importance of both the Low-Rank Adaptation (LoRA) approach and the tropical diseases dataset. Removing these components significantly reduced the model's performance, underscoring the effectiveness of these innovations. Specifically, our model, `mLabLLM`, outperformed baseline models such as `LLaMA 3.2 3B` and `Phi-3-128k` across multiple metrics, including achieving an impressive 82.8% Top-3 accuracy in differential diagnosis, compared to 75.1% for `Phi-3-128k` and 72.4% for `LLaMA 3.2 3B`. Additionally, `mLabLLM` demonstrated notable improvements in symptom matching (F1 score of 0.79) and reasoning accuracy, with a significant 62.1% accuracy on MedQA and 65.2% on USMLE questions. The model's performance in text summarization tasks was also strong, with a ROUGE-L score of 49.3% and a BLEU score of 45.6%. These metrics indicate the robustness and scalability of `mLabLLM` in practical clinical environments, particularly for tropical disease diagnosis, where timely, accurate identification is crucial. Future work will aim to refine these results through further dataset diversification, including more rare disease data, and continued optimization of the model's architecture.

8 Conclusions and Future Directions

In conclusion, this work highlights the effectiveness of a probabilistic model for diagnosing tropical diseases based on symptom input and epidemiological data. The integration of LoRA and targeted datasets improved diagnostic accuracy. Future directions involve expanding the dataset to include more diverse diseases, enhancing model generalization, and exploring advanced optimization techniques such as neural architecture search. Additionally, incorporating real-time data and refining the iterative questioning system could further improve the model's clinical applicability and efficiency.

References

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. arXiv:1903.10676.

Cahill, K. M. 2011. *Tropical Medicine: A Clinical Text*. Fordham Univ Press. ISBN 978-0-8232-4060-9.

Carr, A.; Chen, J.; Goodman, P.; et al. 2022. Med-PaLM: A Conversational Model for Medical Question Answering. *arXiv preprint arXiv:2203.08206*.

Chowdhery, A. K.; and et al. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311*.

Friedman, N.; Geiger, D.; and Goldszmidt, M. 1997. Bayesian Network Classifiers. *Machine Learning*, 29(2): 131–163.

He, Y.; and Xiao, L. 2023. Structured Pruning for Deep Convolutional Neural Networks: A survey. ArXiv:2303.00566.

Hu, E.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Irving, G.; Neves, A. L.; Dambha-Miller, H.; Oishi, A.; Tagashira, H.; Verho, A.; and Holden, J. 2017. International variations in primary care physician consultation time: a systematic review of 67 countries. *BMJ open*, 7(10): e017902.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. ArXiv:2009.13081.

Kasper, D. L.; et al. 2020. *Harrison's Principles of Internal Medicine*. McGraw-Hill Education, 21st edition.

Lee, J.; Lee, D.; Lee, Y.-C.; Hwang, W.-S.; and Kim, S.-W. 2016. Improving the accuracy of top-n recommendation using a preference model. *Information Sciences*, 348: 290–304.

Lee, J.; Lee, J.; Kim, S.; et al. 2021. BioMedLM: A Pretrained Biomedical Language Model for Biomedical Text Mining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1401–1410.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Meunier, Y. A.; Hole, M.; Shumba, T.; and Swanner, B. J. 2013. *Tropical Diseases: A Practical Guide for Medical Practitioners and Students*. OUP USA. ISBN 978-0-19-999790-9. Google-Books-ID: ZWcGAQAAQBAJ.

Neumann, M.; King, D.; Beltagy, I.; and Ammar, W. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. ArXiv:1902.07669.

OpenAI. 2023. GPT-4 Technical Report. Accessed: 2024-11-28.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. Philadelphia, Pennsylvania: Association for Computational Linguistics.

PLOS Neglected Tropical Diseases. 2023. Spatio-temporal patterns of dengue in Bangladesh during 2019 to 2023. *PLOS Neglected Tropical Diseases*.

PubMed. 2022. PubMed Abstracts Database. https://pubmed.ncbi.nlm.nih.gov/. Accessed: 2024-11-28.

Rae, J.; and et al. 2021. Gopher: Training a 280 Billion Parameter Language Model. *arXiv preprint arXiv:2112.11446*.

Rothe, C. 2020. *Clinical Cases in Tropical Medicine*. Elsevier Health Sciences. ISBN 978-0-7020-7880-4. Google-Books-ID: iI4EEAAAQBAJ.

Touvron, H.; et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Tropical Medicine and Health. 2022. Twenty-two years of dengue outbreaks in Bangladesh: epidemiology, clinical spectrum, serotypes, and future disease risks. *Tropical Medicine and Health*.

van Doorn, J.; Ly, A.; Marsman, M.; and Wagenmakers, E.-J. 2020. Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's . *Journal of Applied Statistics*, 47(16): 2984–3006. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02664763.2019.1709053.

World Health Organization (WHO). 2024. *Global Report on Neglected Tropical Diseases 2024*. Geneva, Switzerland: World Health Organization.

Zhang, M.; and et al. 2022. OPT: Open Pretrained Transformer Language Models. *arXiv preprint arXiv:2205.01052*.