Digital Object Identifier

Optimizing Open-Source Large Language Models for Medical Question-Answering: A Comparative Analysis

FIRST A. AUTHOR[1], (Fellow, IEEE), SECOND B. AUTHOR[2], and Third C. Author, Jr.[3], (Member, IEEE)

[1]National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov)
[2]Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu)
[3]Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA

Corresponding author: First A. Author (e-mail: author@boulder.nist.gov).

This paragraph of the first footnote will contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

ABSTRACT

Large language models (LLMs) have transformed artificial intelligence (AI) applications in the medical area, displaying astonishing skills such as passing medical license tests and assisting clinical decision-making. However, the broad adoption of leading proprietary models such as GPT-4 (Generative Pre-trained Transformer 4) and MedPaLM (Medical Pathways Language Model) remains challenging due to their high resource requirements, which restrict their practicality in resource-constrained situations. A novel technique that blends Retrieval Augmented Generation (RAG) with open-source, low-resource LLMs is proposed to create cost-effective and dependable medical question-answering systems.RAG increases LLM effectiveness because it constantly updates the model with current external medical information, decreases hallucinations, and increases factual accuracy. Findings suggest that open-source models maintain competitive performance with accuracy, which in experiments approaches 85% in English tasks when coupled with RAG while having several orders of magnitude less computational requirements. Furthermore, RAG strengthens reasoning capacities in these models to make internally plausible and accurate predictions given contextual information suitable for knowledge-rich medical domains.

This study stresses the practical application of open-source technologies in promoting transparency of AI in medicine, making reliable decision aids in less advantaged territories and conditions. Unlike privileged models, the proposed framework resolves the performance–accessibility paradox using RAG for addressing general issues such as hallucination, outdated data, and reasoning tasks. By avoiding loss of accuracy while focusing on cost-effective implementations, a robust, clear, and easily comprehensible framework is presented to augment nearly any type of medical application in various international healthcare systems.

INDEX TERMS Natural Language Processing(NLP), Retrieval Augmented Generation, Multilingual Benchmarks, Large Language Models(LLM), Medical Question Answering, Gold-Standard Explanations, Diagnostic Reasoning.

I. INTRODUCTION

The integration of Artificial Intelligence (AI) in the medical field is accelerating the development of decision-support technologies that assist healthcare professionals in their daily tasks. Medical Question Answering (QA) has emerged as a critical area of research where Large Language Models (LLMs) have shown immense potential. Notable LLMs such as MedPaLM (Medical Pathways Language Model) and GPT-4 (Generative Pre-trained Transformer 4) have demonstrated their capabilities by passing medical examinations like the United States Medical Licensing Examination (USMLE) [1], [2]. These models enable medical practition-

ers to gather, analyze, and summarize relevant medical data, offering precise diagnoses and treatment recommendations for complex clinical scenarios.

Despite their achievements, LLMs face challenges that hinder broader adoption in medical contexts. Challenges include the generation of hallucinated or factually incorrect information, reliance on outdated training data, and the absence of gold-standard explanations to validate model outputs [3], [4]. Existing benchmarks such as MIRAGE (Multilingual International Retrieval-Augmented Generation Evaluation) and MultiMedQA (Multilingual Medical Question Answering) highlight gaps in multilingual evaluation, as these frameworks primarily focus on English and lack comprehensive reasoning evaluations [5], [6].

Techniques such as Retrieval Augmented Generation (RAG) have been developed to integrate up-to-date external data, thereby enhancing the accuracy and reliability of model outputs [7]. Benchmarks such as MedExpQA have also emerged, providing multilingual datasets with gold-standard reference explanations authored by physicians. These advancements aim to improve the logical reasoning and accuracy of LLMs in diverse linguistic and clinical settings [8]. While these methods have demonstrated improvements, especially in zero-shot learning scenarios, a significant gap remains in performance for non-English applications, emphasizing the need for further research in developing robust, language-agnostic models.

The research aims to build a cost-effective medical QA system based on AI technologies which operate within computational limits of diverse user groups. The high costs of proprietary AI models and their hardware requirements make them impractical for use by both small research groups as well as rural healthcare facilities. As a solution we select open-source LLMs paired with efficient architectures which operate on moderate computational hardware systems. The implementation of Retrieval-Augmented Generation technology instead of fine-tuning reduces both computational necessities and maintains real-time access to current medical information access. The method enables AI healthcare solutions to scale up and remain affordable so different communities can adopt them.Addressing the limitations of existing LLMs by proposing a novel framework that integrates Retrieval Augmented Generation (RAG) with open-source, low-resource LLMs to develop an efficient and reliable medical question-answering system. Unlike proprietary models, this approach focuses on enhancing accessibility and practicality in resource-constrained settings while maintaining competitive performance. By leveraging RAG, the proposed system continuously incorporates up-to-date external medical data, thereby improving factual accuracy, reducing hallucinations, and strengthening reasoning capabilities. Experimental results demonstrate that the framework achieves near state-of-the-art performance, with accuracy approaching 85% in English tasks, all while requiring significantly fewer computational resources. This innovative combination of low-resource LLMs and RAG offers a scalable, transparent, and cost-effective solution for medical applications, addressing gaps in multilingual performance and reasoning capabilities highlighted by existing benchmarks.

II. LITERATURE REVIEW

This study gave us idea about fine-tuning existing LLMs and the application of large language models in the classification of medical multiple-choice questions. This work demonstrates substantial improvements in accuracy across 21 diverse medical subjects. It highlights the capability of LLMs to handle structured datasets like MedMCQA, which contain complex scenarios requiring nuanced understanding. This paper underscores the significance of adapting general-purpose LLMs to specialized domains for improving their reliability and context-aware performance in clinical settings [9].

The MedMCQA dataset represents a pivotal advancement in medical QA research. Containing over 194,000 carefully curated multiple-choice questions spanning a wide range of healthcare topics, this dataset sets a benchmark for evaluating LLMs. Unlike traditional datasets, MedMCQA focuses on real-world clinical scenarios, enabling researchers to assess the logical reasoning and contextual understanding of LLMs. This resource has become essential for improving domain-specific robustness and identifying gaps in model performance across various medical fields [10].

Highlighting their integration into healthcare tasks such as clinical documentation, summarization of medical literature, and diagnostic decision support. Additionally, this study delves into ethical and regulatory challenges, emphasizing the importance of aligning these technologies with global healthcare standards. This paper serves as a foundational resource for understanding the trajectory of LLMs in medicine [11].

This study helps exploring the integration of structured medical knowledge bases with LLMs to enhance diagnostic accuracy. By combining generative capabilities with reliable reference frameworks, this approach mitigates the risks of hallucinated outputs. The findings highlight the potential of hybrid models to improve diagnostic reliability and address challenges in grounding model predictions in accurate and up-to-date medical knowledge [12].

Emphasizing the importance of transparency in AI-generated outputs and highlights the need for interdisciplinary collaboration to address issues of explainability and security. By identifying gaps in current systems, this paper provides actionable strategies for the robust deployment of LLMs in clinical settings and proposes frameworks to ensure the accuracy and trustworthiness of LLMs in healthcare [13].

This study reviews the current state of LLM research in medical contexts. It identifies significant gaps in evaluation methodologies, including the lack of metrics to assess contextual reasoning and domain-specific accuracy. The authors propose a comprehensive framework for evaluating clinical utility, highlighting the importance of bridging technical advancements with real-world medical needs [14].

Evaluating the differences between general-purpose LLMs and those fine-tuned for medical applications. The study finds that domain-specific models outperform their general counterparts in accuracy and contextual relevance, particularly when dealing with complex medical queries. This work emphasizes the value of tailoring LLMs to specific domains for improved reliability and practical utility in medical QA tasks [15].

This paper constructs a dataset featuring expert-written explanations to evaluate LLM performance on complex clinical scenarios. It highlights the need for explainability as a core metric in medical QA, demonstrating that models capable of providing clear reasoning alongside their predictions can significantly enhance trust and usability in healthcare applications [16].

This study examines how LLMs internalize and retrieve clinical knowledge. The findings reveal that while these models excel in retrieving information and summarizing medical data, they face limitations in reasoning and decision-making. The study underscores the need for continuous improvements in training methodologies to bridge these gaps [17].

Introduces a hybrid framework that integrates LLMs with retrieval models. By jointly training these models, the study demonstrates significant improvements in reasoning and contextual accuracy, particularly for complex clinical scenarios. This approach underscores the importance of combining generative and retrieval-based systems for achieving optimal performance [18].

The paper showcases the application of LLMs in interacting with clinical notes dynamically. The study emphasizes the ability of these models to extract relevant information efficiently, offering transformative potential in managing electronic health records. This innovation can significantly enhance clinical workflows and decision-making processes [19].

Examining the role of generative LLMs, such as ChatGPT, in healthcare applications. The paper highlights their effectiveness in patient education and decision support but also discusses challenges related to bias and accuracy. The authors stress the need for verification mechanisms to ensure the reliability of these models in critical medical contexts [20].

The reviewed literature highlights the transformative potential of large language models (LLMs) in medical question answering (QA), emphasizing their ability to process complex datasets like MedMCQA, enhance diagnostic accuracy, and support clinical decision-making. Studies demonstrate that fine-tuned, domain-specific LLMs outperform general-purpose models in contextual relevance and reliability, particularly in handling nuanced medical queries. The integration of structured knowledge bases and hybrid retrieval-augmented frameworks has further advanced diagnostic capabilities while mitigating issues such as hallucinated outputs. Ethical challenges, explainability, and security remain critical areas, with several works proposing frameworks to ensure transparency and trustworthiness. Additionally, datasets with expert annotations have become instrumental in assessing LLM reasoning and contextual understanding. However, limitations in reasoning, bias, and domain-specific accuracy persist, underscoring the need for continuous improvements.

III. MATERIALS AND METHODS

A. WORKFLOW

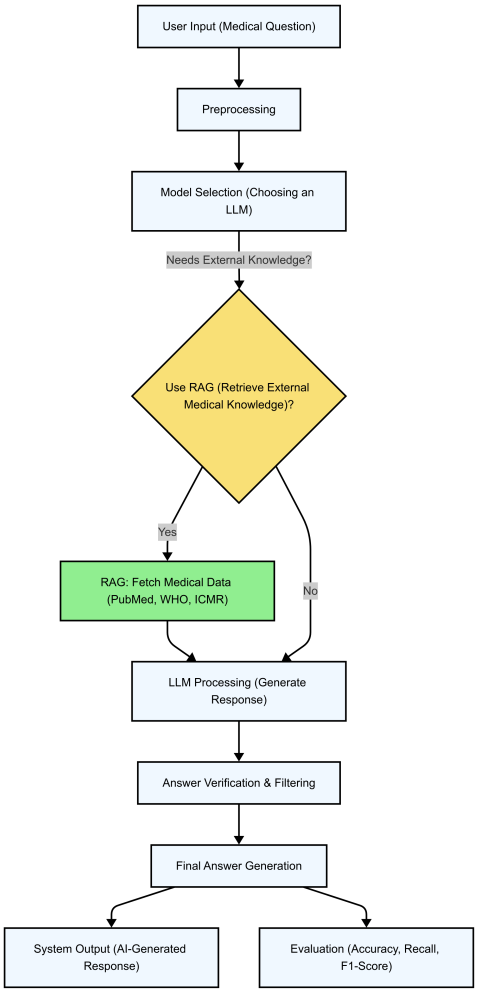

FIGURE 1. Proposed workflow of the system.

B. DATA PREPARATION

The dataset employed in this study is made up of multiple-choice questions from several medical fields. Each question is provided in a standardized four-option multiple-choice style (A, B, C, and D) and labeled with the right answer to allow for objective grading. The dataset includes a wide range of medical themes, guaranteeing complete coverage across disciplines and allowing for subject-specific study of model performance.

The dataset includes questions spanning various specialties, including cardiology, neurology, oncology, pediatrics, general medicine, and other essential areas of medical edu-

cation. These questions were selected to resemble actual test and clinical encounters, which gives them an accurate representation of most common medical questions. This ensures that the models trained are able to not only be knowledge based, but also logical reasoning based in order to perform well in clinical contexts.

To improve the dataset's quality, a stratified sample strategy was adopted to ensure that questions are spread across the medical subject to increase the quality of the dataset. Hence, only questions that were relevant to the topic and questions that were synced with competitive exams like NEET, AIIMS, etc., of India were selected that not only made a robust database but also a rather challenging one. A detailed vectorbase was compiled with vector from CBSE notebooks, notes and books gleaned from various reliable internet sources. These resources combined provide a good mixture of data which effectively tests the models understanding of many topics in medicine.

In addition, the number of questions in this survey is 250, which can be considered as a representative sample size because of both computationally section and enough distinction for analysis. This ensures that while the models were able to learn a wide range of subjects, it did not have an effect on computational speed.

C. EXPERIMENTAL SETUP SECTION

The models were evaluated in a high-performance computer environment specifically designed for deep learning inference. Given the computational intensity of large-scale LLMs, the following hardware configuration was used: GPU: NVIDIA A100 (16GB VRAM). CPU: eight-core processor. RAM: 32 GB system memory. This arrangement enabled the efficient execution of resource-intensive models such as Gemma 27B and Solar Pro 22B, avoiding memory constraints and lowering inference delay. Smaller versions, such as the Qwen 2.5 14B Q8 and Nemotron-mini FP16, were able to run with fewer hardware needs, indicating that memory-efficient designs may deliver comparable performance without incurring significant computational cost. This combination of scalability and efficiency is crucial for implementing AI-powered medical assistants in resource-constrained settings.

D. PREPROCESSING AND STRUCTURING

To prepare the dataset for model input, each item was thoroughly examined and standardized to guarantee uniformity in style and wording. Preprocessing includes:

Text cleaning : Removing extraneous punctuation, formatting irregularities, and non-standard symbols in order to retain clarity.

Option Labeling: Standardizing answer options (A, B, C, D) throughout the dataset to ensure consistency in model prompts.

Subject Stratification: The questions were divided into medical subject categories to provide a fair distribution across different areas. This stratification was created to guarantee that each model's performance could be assessed not just overall, but also within each individual topic domain.

Sampling Strategy: A stratified sample strategy was used, with a representative subset of questions drawn from each topic area. This ensured that the models experienced a diverse variety of subjects while remaining within a tolerable dataset size for computational performance.

E. MODEL SELECTION

In this work, a variety of large language models (LLMs) were tested for their performance in the medical question-answering (QA) task. Each model was chosen for its distinct setup, parameter count, and potential capabilities in dealing with complicated language patterns and medical terminology. These models were evaluated on their ability to select the correct answer in a multiple-choice format across a wide range of medical topics. This research comprised the following models:

1) Solar Pro 22B 8K

Solar Pro is a high-capacity model with 22 billion parameters that is particularly developed for processing and comprehending enormous context windows. This functionality is useful for medical QA work, since queries frequently contain detailed background information. The model's parameter size and structure are designed to effectively capture nuanced medical data, allowing for complicated reasoning.

2) Gemma 27B

Gemma, with 27 billion attributes, is one of the study's largest models, designed for complex understanding of specialist medical terminology and syntactic patterns. The large number of parameters in this model is projected to increase its interpretive depth, resulting in correct and contextually appropriate solutions to difficult medical questions.

3) LLaMA Series (LLaMA 3, LLaMA 3.1, and LLaMA 3.2)

The LLaMA models were chosen because of their gradual development in processing long-form material and increasing interpretative accuracy with each iteration. LLaMA 3, LLaMA 3.1, and LLaMA 3.2 all provide improvements that build on one another, resulting in more accurate and coherent replies. Each version's modest advancements in language understanding aid in the capturing of a wide range of medical concepts across a variety of queries.

4) Qwen 2.5 14B Q8

The Qwen model, which has 14 billion parameters and quantization at Q8, combines computing efficiency and great precision. Its quantized form allows for efficient memory utilization, making it suited for widespread deployment while maintaining significant interpretative power. This balance is especially beneficial when it comes to answering questions efficiently and accurately.

5) Mixtral

Mixtral, which is well-known for its broad versatility across a wide range of fields, was chosen for its strength in question-answering tasks. Although it is not specialized in medical terminology, its strong flexibility gives it a useful baseline for comparing performance variability to medically customized models.

6) Mistral Small

The Mistral Small model is an efficiency-focused alternative in the LLM family, designed to perform question-answering jobs with minimal computing cost. While it has fewer parameters than some of the bigger models, its streamlined architecture allows it to respond quickly, which is beneficial in situations when resource efficiency is crucial.

7) Nemotron-mini FP16 4B

The Nemotron-mini is a tiny model with four billion parameters that has been fine-tuned using FP16 (16-bit floating-point) accuracy. Its very minimal argument count makes it memory-efficient, which is useful for activities that need lightweight deployments. Despite its small size, Nemotron-mini has the capacity to comprehend basic medical language constructions successfully.

8) Phi 3.5 FP16

Phi 3.5, like Nemotron-mini, operates at FP16 precision, seeking to strike a compromise between memory economy and interpretative capabilities. The design of Phi 3.5, together with its reasonable parameter count, provides an alternate technique to achieve accurate QA performance without requiring an excessive amount of resources.

We selected Retrieval-Augmented Generation (RAG) instead of fine-tuning models because it provided more scalable and cost-effective medical solution. The customization of Gemma 27B or Solar Pro 22B requires large computing power alongside particular medical data alongside regular updates to match contemporary medical information. The prevalent alterations to medical guidelines make models that receive fine-tuning unfit for long-term usage since updates would occur in less than six months.

RAG provides an update-to-update source retrieval system that integrates official medical resources which enhances factual validity while diminishing static pre-trained information and its dependency. Through this method the system can work across different specialties (cardiology, neurology, oncology) without requiring new training while avoiding incorrect information known as hallucinations. Medicare Knowledge Retrieval integrated into the system helps maintain responses which stay both current and doctor-approved in their medical content and interpretation.

Multiple models provide researchers with broad options to evaluate how system size interacts with memory utilization and medical question accuracy measurement.

F. EVALUATION SETUP

1) Metrics for Evaluation

Each model's performance was assessed using a variety of classic classification measures, which give a multifaceted perspective of model efficacy. These measurements include the following:

Accuracy: It refers to the proportion of questions properly answered by the model across all disciplines. This statistic gives a simple assessment of the model's ability to correctly identify the solution in each situation.

Precision, recall, and F1 Score: These measures provide a more detailed understanding of model performance. Precision assesses the accuracy of the model's positive predictions, recall reflects the model's ability to identify all right responses, and the F1 score (the harmonic mean of precision and recall) combines the two to offer an overall assessment of response quality. These KPIs are especially important in medical QA, where accuracy and the capacity to continuously produce right responses are required.

2) Testing Protocol and Memory Management

Each model was tested in a controlled testing environment to ensure consistency. The following test methodology was used:

Token Management: A token limit was set to manage the amount of model inputs and outputs, preventing memory overflow and ensuring constant response times. A memory buffer with a token cap was implemented, allowing each model to only respond with the appropriate answer (A, B, C, or D).

Prompt Structuring: The prompts for each question were designed to encourage the model to carefully study the issue, consider each answer choice, and deliver a single-character response matching to the chosen answer. This style was designed to ensure answer clarity while preventing the model from producing unnecessary text.

Context Resetting: To avoid context contamination between questions, the model's context was reset after each question. This guaranteed that each question was answered separately, with no influence from prior ones. This method gives a fair evaluation of each model's solitary reasoning capacity.

Evaluation Software and Tools: The assessment was carried out with Python and appropriate machine learning and natural language processing packages, such as Scikit-Learn for determining performance metrics. To ensure consistency and correctness in the evaluation, the responses of each model were logged, and metrics were produced programmatically.

The testing system designed for this study guarantees that each model's performance is evaluated fairly and consistently, resulting in credible data for model comparison.

IV. RESULTS

A. OVERALL PERFORMANCE

The performance of the evaluated language models was assessed across key metrics: accuracy, precision, recall, and F1 score. These metrics provide a comprehensive understanding of the models' ability to answer medical questions accurately and consistently. The results are summarized in Table 1, showcasing performance across various large language models (LLMs) tested on a diverse dataset of medical disciplines.

TABLE 1. Performance Metrics of Evaluated Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	73.60	0.7674	0.7360	0.7316
LLaMA 3	83.60	0.8428	0.8360	0.8364
Gemma 27B	66.00	0.6714	0.6600	0.6612
LLaMA 3.2	64.80	0.7281	0.6480	0.6515
Solar Pro 22B 8K	39.20	0.4022	0.3958	0.3664
Nemotron-mini FP16 4B	23.60	0.2531	0.2360	0.2386
Qwen 2.5 14B Q8	70.40	0.7290	0.7040	0.7044
Phi 3.5 FP16	57.60	0.6689	0.5760	0.5635
Mistral	51.66	0.5238	0.4755	0.4625
Mistral Small	45.93	0.4881	0.4529	0.4313

Based on the results we observed that the performance of the models varied significantly, and this can be attributed to multiple factors, including the model architecture, the fine-tuning process, and how well the models adapted to the specialized medical dataset.

LLaMA 3 achieved the highest accuracy and F1 score, which suggests that its architecture is well-suited for the task and that it was effectively fine-tuned for medical question classification. Its ability to understand complex medical scenarios and respond accurately indicates that the model was able to capture the intricacies of the data, making it the best-performing model.

On the other hand, Gemma 27B and LLaMA 3.2 showed lower accuracy and F1 scores. This could be due to either insufficient fine-tuning or challenges in transferring general knowledge to the specialized medical domain. Despite their large model size, they struggled with the complexities of medical terminology and multi-choice question patterns, which led to their lower performance.

The models like Mistral, Solar Pro 22B 8K, and Nemotron-mini FP16 4B showed the weakest performance, with accuracy ranging from 39% to 45%. This could be because these models either weren't fine-tuned enough for this domain or had architectural limitations that made them less capable of handling the structured medical datasets. Their lower accuracy may also reflect difficulties in understanding the nuanced language of medical questions, which might require more specialized training data or model adjustments.

B. DISCUSSION OF RESULTS

The evaluation of the models revealed several important insights regarding their strengths, limitations, and suitability for medical question-answering tasks.

Table 2 presents the performance metrics of various anatomy language models, showcasing significant differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** emerged as the most effective, achieving the highest accuracy (88.89%) and F1 score (0.8940), indicating its superior overall performance. **LLaMA 3.1** and **LLaMA 3.2** also demonstrated moderate success with accuracy scores of 70.37% and 66.67%, respectively, and F1 scores of 0.6174 and 0.6570. **Gemma 27B** exhibited lower accuracy (62.96%) but maintained balanced precision and recall. In contrast, models such as **Solar Pro 22B 8K** and **Nemotron-mini FP16 4B** performed poorly, with accuracy as low as 7.41% and weak F1 scores. Mid-range performers like **Qwen 2.5 14B Q8**, **Phi 3.5 FP16**, and **Mistral Small** achieved moderate accuracy (51.85%) with corresponding F1 scores ranging from 0.4623 to 0.5224. Overall, the results highlight **LLaMA 3** as the most effective model in this comparison.

TABLE 2. Performance Metrics of Anatomy Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	70.37	0.6506	0.6389	0.6174
LLaMA 3	88.89	0.9037	0.8889	0.8940
Gemma 27B	62.96	0.6278	0.6296	0.6134
LLaMA 3.2	66.67	0.7407	0.6667	0.6570
Solar Pro 22B 8K	29.63	0.3063	0.3931	0.3036
Nemotron-mini FP16 4B	7.41	0.1058	0.0741	0.0871
Qwen 2.5 14B Q8	51.85	0.6543	0.5185	0.5224
Phi 3.5 FP16	51.85	0.5729	0.5185	0.5126
Mistral	44.44	0.3806	0.3833	0.3591
Mistral Small	51.85	0.4726	0.4736	0.4623

Table 3 presents the performance metrics of various biochemistry language models, showcasing significant differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** and **LLaMA 3.2** demonstrated the highest accuracy (73.33%) with strong F1 scores (0.7500 and 0.7429, respectively), indicating their superior performance. **LLaMA 3.1** followed closely with an accuracy of 66.67% and an F1 score of 0.6738. **Gemma 27B** and **Mistral** showed moderate accuracy (60.00%) with relatively balanced precision and recall. In contrast, models like **Solar Pro 22B 8K** and **Nemotron-mini FP16 4B** performed poorly, with accuracy as low as 20.00% and weak F1 scores (0.1929 and 0.2056, respectively). Mid-range performers such as **Qwen 2.5 14B Q8**, **Phi 3.5 FP16**, and **Mistral Small** achieved accuracy between 60.00% and 73.33%, with corresponding F1 scores ranging from 0.5777 to 0.7270. Overall, the results highlight **LLaMA 3** and **LLaMA 3.2** as the most effective models in this comparison.

Table 4 presents the performance metrics of various ENT language models, showcasing notable differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** stands out with a perfect accuracy (100%), precision (1), recall (1), and F1 score (1), indicating its exceptional performance. **Mistral Small** follows with a high accuracy

TABLE 3. Performance Metrics of Biochemistry Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	66.67	0.7738	0.7042	0.6738
LLaMA 3	73.33	0.8300	0.7333	0.7500
Gemma 27B	60.00	0.6056	0.6000	0.5603
LLaMA 3.2	73.33	0.8200	0.7333	0.7429
Solar Pro 22B 8K	20.00	0.1750	0.2167	0.1929
Nemotron-mini FP16 4B	20.00	0.2400	0.2000	0.2056
Qwen 2.5 14B Q8	73.33	0.7333	0.7333	0.7270
Phi 3.5 FP16	60.00	0.8250	0.6000	0.5777
Mistral	60.00	0.6083	0.6083	0.5833
Mistral Small	60.00	0.7250	0.6292	0.5792

(75.00%) and a balanced F1 score (0.6389). Models such as **LLaMA 3.1**, **LLaMA 3.2**, and **Qwen 2.5 14B Q8** achieved moderate accuracy (62.50%) with F1 scores ranging from 0.5750 to 0.6250. **Gemma 27B** and **Phi 3.5 FP16** demonstrated lower accuracy (50.00%) with less consistent metrics. In contrast, models like **Solar Pro 22B 8K**, **Nemotron-mini FP16 4B**, and **Mistral** performed poorly, with accuracy as low as 25.00% and weak F1 scores ranging from 0.2083 to 0.3500. Overall, **LLaMA 3** significantly outperforms other models in this comparison.

TABLE 4. Performance Metrics of ENT Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	62.50	0.7083	0.7500	0.6250
LLaMA 3	100	1	1	1
Gemma 27B	50.00	0.5833	0.5000	0.4583
LLaMA 3.2	62.50	0.8125	0.6250	0.5750
Solar Pro 22B 8K	37.50	0.4167	0.5625	0.3500
Nemotron-mini FP16 4B	25.00	0.3125	0.2500	0.2500
Qwen 2.5 14B Q8	62.50	0.8500	0.6250	0.5929
Phi 3.5 FP16	50.00	0.7292	0.5000	0.4708
Mistral	25.00	0.2083	0.3125	0.2083
Mistral Small	75.00	0.7000	0.6250	0.6389

Table 5 presents the performance metrics of various forensic medicine language models, highlighting differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** achieved the highest accuracy (70.00%) and an F1 score of 0.7071, indicating its superior performance in this domain. **Qwen 2.5 14B Q8** followed with a moderate accuracy (60.00%) and a strong F1 score (0.6500). Models such as **Mistral** and **Gemma 27B** demonstrated moderate accuracy (50.00% and 40.00%, respectively) with balanced but less robust performance metrics. In contrast, **LLaMA 3.2** and **Nemotron-mini FP16 4B** performed poorly, with accuracy as low as 10.00% and weak F1 scores (0.1333 and 0.1143, respectively). Mid-range performers like **LLaMA 3.1** and **Solar Pro 22B 8K** showed moderate accuracy (40.00%) but inconsistent F1 scores (0.3631 and 0.2917, respectively). Overall, **LLaMA 3** stands out as the most effective model in this comparison.

Table 6 presents the performance metrics of various gynecology and obstetrics language models, highlighting variations in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** achieved the highest accuracy (83.33%) and an F1 score of 0.8342, indicating its superior performance in this domain. **LLaMA 3.1** and **Gemma 27B** followed closely with moderate accuracy (72.22%) and F1 scores of 0.7333 and 0.7384, respectively. **Qwen 2.5 14B Q8** also performed well with an accuracy of 72.22% and an F1 score of 0.7258. In contrast, **LLaMA 3.2** and models like **Solar Pro 22B 8K** showed lower accuracy (55.56% and 38.89%, respectively) and weaker F1 scores (0.5572 and 0.3742). **Nemotron-mini FP16 4B** failed to perform, with all metrics at 0. Mid-range performers such as **Phi 3.5 FP16**, **Mistral**, and **Mistral Small** achieved accuracy (50.00%) with F1 scores ranging from 0.4665 to 0.5303. Overall, **LLaMA 3** emerged as the most effective model in this comparison.

TABLE 5. Performance Metrics of Forensic Medicine Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	40.00	0.5000	0.3125	0.3631
LLaMA 3	70.00	0.8250	0.7000	0.7071
Gemma 27B	40.00	0.4000	0.4000	0.4000
LLaMA 3.2	10.00	0.2000	0.1000	0.1333
Solar Pro 22B 8K	40.00	0.2917	0.2917	0.2917
Nemotron-mini FP16 4B	10.00	0.1333	0.1000	0.1143
Qwen 2.5 14B Q8	60.00	0.8000	0.6000	0.6500
Phi 3.5 FP16	20.00	0.1800	0.2000	0.1889
Mistral	50.00	0.5000	0.3542	0.4143
Mistral Small	30.00	0.2083	0.2292	0.1944

TABLE 6. Performance Metrics of Gynaecology & Obstetrics Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	72.22	0.7417	0.8250	0.7333
LLaMA 3	83.33	0.8578	0.8333	0.8342
Gemma 27B	72.22	0.7935	0.7222	0.7384
LLaMA 3.2	55.56	0.7222	0.5556	0.5572
Solar Pro 22B 8K	38.89	0.4500	0.3458	0.3742
Nemotron-mini FP16 4B	0	0	0	0
Qwen 2.5 14B Q8	72.22	0.8056	0.7222	0.7258
Phi 3.5 FP16	50.00	0.7037	0.5000	0.4679
Mistral	50.00	0.6875	0.5354	0.5303
Mistral Small	50.00	0.7103	0.5396	0.4665

Table 7 presents the performance metrics of various medicine language models, highlighting differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** demonstrated the highest accuracy (80.00%) and an F1 score of 0.7989, showcasing its superior performance. **LLaMA 3.1**, **Gemma 27B**, and **Qwen 2.5 14B Q8** followed closely with moderate accuracy (73.33%, 73.33%, and 76.67%, respectively) and F1 scores ranging from 0.7264 to 0.7621. **LLaMA 3.2** achieved an accuracy of 70.00% and an F1 score of 0.7058, indicating consistent but slightly lower performance. Models like **Solar Pro 22B 8K**, **Nemotron-**

mini FP16 4B, **Mistral**, and **Mistral Small** displayed weaker performance, with accuracy ranging from 20.00% to 50.00% and F1 scores between 0.2027 and 0.4904. **Phi 3.5 FP16** delivered mid-range results with an accuracy of 53.33% and an F1 score of 0.5282. Overall, **LLaMA 3** emerged as the most effective model in this comparison.

TABLE 7. Performance Metrics of Medicine Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	73.33	0.7438	0.7604	0.7468
LLaMA 3	80.00	0.8095	0.8000	0.7989
Gemma 27B	73.33	0.7333	0.7333	0.7264
LLaMA 3.2	70.00	0.7148	0.7000	0.7058
Solar Pro 22B 8K	40.00	0.4445	0.3799	0.3937
Nemotron-mini FP16 4B	20.00	0.2069	0.2000	0.2027
Qwen 2.5 14B Q8	76.67	0.7823	0.7667	0.7621
Phi 3.5 FP16	53.33	0.5764	0.5333	0.5282
Mistral	46.67	0.4717	0.4389	0.4263
Mistral Small	50.00	0.5136	0.5451	0.4904

Table 8 presents the performance metrics of various microbiology language models, highlighting variations in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** demonstrated the highest accuracy (90.00%) and F1 score (0.8997), establishing itself as the most effective model in this domain. **LLaMA 3.1** followed with an accuracy of 65.00% and an F1 score of 0.5595, while **LLaMA 3.2** achieved a slightly lower accuracy (60.00%) but a higher F1 score (0.6754). **Qwen 2.5 14B Q8** also delivered moderate performance, with an accuracy of 60.00% and an F1 score of 0.6104. In contrast, models like **Solar Pro 22B 8K**, **Gemma 27B**, and **Phi 3.5 FP16** displayed weaker performance, with accuracy ranging from 45.00% to 55.00% and F1 scores between 0.3979 and 0.5166. **Nemotron-mini FP16 4B**, **Mistral**, and **Mistral Small** exhibited low performance, with accuracy as low as 25.00% and F1 scores between 0.1801 and 0.3763. Overall, **LLaMA 3** stands out as the top-performing model in this comparison.

TABLE 8. Performance Metrics of Microbiology Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	65.00	0.6929	0.6929	0.5595
LLaMA 3	90.00	0.90622	0.9000	0.8997
Gemma 27B	50.00	0.6000	0.5000	0.5166
LLaMA 3.2	60.00	0.8187	0.6000	0.6754
Solar Pro 22B 8K	55.00	0.4444	0.4071	0.4056
Nemotron-mini FP16 4B	35.00	0.4333	0.3500	0.3763
Qwen 2.5 14B Q8	60.00	0.6622	0.6000	0.6104
Phi 3.5 FP16	45.00	0.4375	0.4500	0.3979
Mistral	45.00	0.5786	0.3357	0.3869
Mistral Small	25.00	0.3500	0.1786	0.1801

Table 9 presents the performance metrics of various ophthalmology language models, showcasing notable differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** achieved perfect accuracy (100%) and an F1 score of 1, making it the most effective model in this domain. **LLaMA 3.2** followed closely with high accuracy (83.33%) and an F1 score of 0.8657, indicating strong performance. **Gemma 27B** and **Qwen 2.5 14B Q8** also performed well, with accuracies of 66.67% and 75.00%, and F1 scores of 0.6944 and 0.7801, respectively. In contrast, **Solar Pro 22B 8K**, **Phi 3.5 FP16**, **Mistral**, and **Mistral Small** displayed weaker performance, with accuracy ranging from 33.33% to 58.33% and F1 scores between 0.2292 and 0.4247. **Nemotron-mini FP16 4B** exhibited the lowest performance, with an accuracy of 25.00% and an F1 score of 0.2894. Overall, **LLaMA 3** emerged as the clear leader in this comparison, followed by **LLaMA 3.2**.

TABLE 9. Performance Metrics of Ophthalmology Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	58.33	0.6167	0.7500	0.5333
LLaMA 3	100	1	1	1
Gemma 27B	66.67	0.7292	0.6667	0.6944
LLaMA 3.2	83.33	0.9444	0.8333	0.8657
Solar Pro 22B 8K	41.67	0.5417	0.6500	0.4083
Nemotron-mini FP16 4B	25.00	0.3472	0.2500	0.2894
Qwen 2.5 14B Q8	75.00	0.9028	0.7500	0.7801
Phi 3.5 FP16	50.00	0.3778	0.5000	0.4141
Mistral	33.33	0.2232	0.4000	0.2292
Mistral Small	58.33	0.4792	0.5500	0.4247

Table 10 presents the performance metrics of various orthopaedics language models, highlighting differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** achieved perfect accuracy (100%) and an F1 score of 1, making it the most effective model in this comparison. **Gemma 27B** and **Qwen 2.5 14B Q8** followed closely with high accuracy (83.33%) and strong F1 scores of 0.9000 and 0.8889, respectively. Models like **LLaMA 3.1**, **LLaMA 3.2**, and **Mistral** demonstrated moderate accuracy (66.67%) and comparable F1 scores of 0.7222 and 0.7333. **Phi 3.5 FP16** showed a lower accuracy (50.00%) with an F1 score of 0.5556. In contrast, **Mistral Small** and **Nemotron-mini FP16 4B** performed poorly, with accuracies of 50.00% and 16.67%, and F1 scores of 0.4500 and 0.1111, respectively. **Solar Pro 22B 8K** failed to deliver, with all metrics at 0. Overall, **LLaMA 3** emerged as the best-performing model, followed by **Gemma 27B** and **Qwen 2.5 14B Q8**.

Table 11 presents the performance metrics of various pathology language models, highlighting differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** achieved the highest accuracy (80.77%) and a strong F1 score (0.8391), making it the most effective model in this comparison. **Gemma 27B** followed closely with an accuracy of 76.92% and an F1 score of 0.8183, showcasing competitive performance. Models like **LLaMA 3.1**, **LLaMA 3.2**, and **Qwen 2.5 14B Q8** delivered moderate accuracy (73.08%, 65.38%, and 73.08%, respectively) with F1 scores ranging from 0.6240 to 0.7226. **Phi 3.5 FP16** demonstrated lower accuracy (61.54%) but maintained a fair F1 score of

TABLE 10. Performance Metrics of Orthopaedics Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	66.67	0.8333	0.7778	0.7222
LLaMA 3	100	1	1	1
Gemma 27B	83.33	1	0.8333	0.9000
LLaMA 3.2	66.67	0.9167	0.6667	0.7333
Solar Pro 22B 8K	0	0	0	0
Nemotron-mini FP16 4B	16.67	0.0833	0.1667	0.1111
Qwen 2.5 14B Q8	83.33	1	0.8333	0.8889
Phi 3.5 FP16	50.00	0.8889	0.5000	0.5556
Mistral	66.67	0.8333	0.7778	0.7222
Mistral Small	50.00	0.5000	0.4167	0.4500

0.6680. In contrast, **Solar Pro 22B 8K**, **Mistral**, and **Mistral Small** showed weaker performance, with accuracy ranging from 42.31% to 69.23% and F1 scores between 0.3440 and 0.6272. **Nemotron-mini FP16 4B** performed the worst, with an accuracy of 15.38% and an F1 score of 0.1720. Overall, **LLaMA 3** and **Gemma 27B** emerged as the top-performing models in this domain.

TABLE 11. Performance Metrics of Pathology Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	73.08	0.6643	0.6143	0.6240
LLaMA 3	80.77	0.9017	0.8077	0.8391
Gemma 27B	76.92	0.9038	0.7692	0.8183
LLaMA 3.2	65.38	0.8846	0.6538	0.7005
Solar Pro 22B 8K	50.00	0.5040	0.4762	0.4217
Nemotron-mini FP16 4B	15.38	0.2479	0.1538	0.1720
Qwen 2.5 14B Q8	73.08	0.8362	0.7308	0.7226
Phi 3.5 FP16	61.54	0.7885	0.6154	0.6680
Mistral	42.31	0.3958	0.3780	0.3440
Mistral Small	69.23	0.6250	0.6492	0.6272

Table 12 presents the performance metrics of various pediatrics language models, highlighting variations in accuracy, precision, recall, and F1 score. Among the models, **Qwen 2.5 14B Q8** and **Phi 3.5 FP16** demonstrated the highest accuracy (78.57%) and strong F1 scores (0.7512 and 0.7429, respectively), making them the top performers in this comparison. **LLaMA 3.1** also showed competitive performance, with an accuracy of 71.43% and an F1 score of 0.7273. **LLaMA 3** and **Gemma 27B** achieved moderate accuracy (64.29%) and F1 scores of 0.6212 and 0.6224, respectively. **LLaMA 3.2** and **Solar Pro 22B 8K** displayed slightly lower accuracy (57.14%) and F1 scores of 0.5476 and 0.5628, respectively. Models like **Mistral**, **Mistral Small**, and **Nemotron-mini FP16 4B** performed poorly, with accuracy ranging from 28.57% to 50.00% and F1 scores between 0.2830 and 0.3720. Overall, **Qwen 2.5 14B Q8** and **Phi 3.5 FP16** stand out as the most effective models for this domain.

Table 13 presents the performance metrics of various pharmacology language models, showcasing differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** achieved the highest accuracy (84.00%) and

TABLE 12. Performance Metrics of Pediatrics Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	71.43	0.8571	0.7708	0.7273
LLaMA 3	64.29	0.6643	0.6429	0.6212
Gemma 27B	64.29	0.6131	0.6429	0.6224
LLaMA 3.2	57.14	0.7347	0.5714	0.5476
Solar Pro 22B 8K	57.14	0.5333	0.6458	0.5628
Nemotron-mini FP16 4B	28.57	0.3571	0.2857	0.2830
Qwen 2.5 14B Q8	78.57	0.8786	0.7857	0.7512
Phi 3.5 FP16	78.57	0.8571	0.7857	0.7429
Mistral	35.71	0.2917	0.5000	0.3512
Mistral Small	50.00	0.3542	0.3958	0.3720

F1 score (0.8464), making it the most effective model in this comparison. **Qwen 2.5 14B Q8** and **Gemma 27B** followed closely with accuracies of 80.00% and 76.00% and F1 scores of 0.7966 and 0.7479, respectively. **LLaMA 3.1**, **LLaMA 3.2**, and **Mistral Small** delivered moderate accuracy (72.00%) with F1 scores ranging from 0.7067 to 0.7283. In contrast, **Solar Pro 22B 8K**, **Mistral**, and **Nemotron-mini FP16 4B** performed poorly, with accuracy ranging from 32.00% to 64.00% and F1 scores between 0.3145 and 0.5669. **Phi 3.5 FP16** demonstrated balanced performance, achieving an accuracy of 76.00% and an F1 score of 0.7592. Overall, **LLaMA 3** and **Qwen 2.5 14B Q8** stand out as the top-performing models for this domain.

TABLE 13. Performance Metrics of Pharmacology Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	72.00	0.7391	0.7708	0.7082
LLaMA 3	84.00	0.8689	0.8400	0.8464
Gemma 27B	76.00	0.7663	0.7600	0.7479
LLaMA 3.2	72.00	0.8187	0.7200	0.7067
Solar Pro 22B 8K	64.00	0.5896	0.5729	0.5597
Nemotron-mini FP16 4B	32.00	0.3518	0.3200	0.3145
Qwen 2.5 14B Q8	80.00	0.8297	0.8000	0.7966
Phi 3.5 FP16	76.00	0.8003	0.7600	0.7592
Mistral	56.00	0.6679	0.6354	0.5669
Mistral Small	72.00	0.7743	0.7188	0.7283

Table 14 presents the performance metrics of various physiology language models, highlighting differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** demonstrated the highest accuracy (88.24%) and F1 score (0.8687), making it the most effective model in this comparison. **Qwen 2.5 14B Q8** followed closely with an accuracy of 82.35% and an F1 score of 0.8370, showcasing strong performance. **Phi 3.5 FP16** and **Gemma 27B** delivered moderate accuracy (70.59% and 64.71%, respectively) with corresponding F1 scores of 0.6975 and 0.6529. Models such as **LLaMA 3.2**, **Mistral**, and **Mistral Small** showed similar accuracy (58.82%) but varied in F1 scores, ranging from 0.5435 to 0.5983. In contrast, **Solar Pro 22B 8K** and **Nemotron-mini FP16 4B** performed poorly, with accuracy as low as 23.53% and F1 scores between 0.2057 and 0.4881.

Overall, **LLaMA 3** and **Qwen 2.5 14B Q8** emerged as the top-performing models for this domain.

TABLE 14. Performance Metrics of Physiology Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	58.82	0.6458	0.5833	0.5125
LLaMA 3	88.24	0.9036	0.8824	0.8687
Gemma 27B	64.71	0.6980	0.6471	0.6529
LLaMA 3.2	64.71	0.5504	0.6471	0.5725
Solar Pro 22B 8K	52.94	0.5202	0.5104	0.4881
Nemotron-mini FP16 4B	23.53	0.3170	0.2353	0.2057
Qwen 2.5 14B Q8	82.35	0.9118	0.8235	0.8370
Phi 3.5 FP16	70.59	0.8599	0.7059	0.6975
Mistral	58.82	0.5833	0.6146	0.5435
Mistral Small	58.82	0.6750	0.6250	0.5983

Table 15 presents the performance metrics of various social and preventive medicine language models, highlighting differences in accuracy, precision, recall, and F1 score. Among the models, **LLaMA 3** achieved the highest accuracy (81.82%) and F1 score (0.8393), making it the most effective model in this comparison. **LLaMA 3.2** followed with a strong performance, achieving an accuracy of 72.73% and an F1 score of 0.7614. **Gemma 27B** and **Qwen 2.5 14B Q8** delivered moderate accuracy (63.64% and 68.18%, respectively) and F1 scores of 0.6515 and 0.6818. Models like **LLaMA 3.1** and **Phi 3.5 FP16** showed similar accuracy (63.64%) but varying F1 scores of 0.4958 and 0.6132. In contrast, **Solar Pro 22B 8K**, **Mistral**, and **Mistral Small** performed poorly, with accuracy ranging from 36.36% to 50.00% and F1 scores between 0.2917 and 0.3988. **Nemotron-mini FP16 4B** displayed the weakest performance, with an accuracy of 27.27% and an F1 score of 0.3333. Overall, **LLaMA 3** and **LLaMA 3.2** emerged as the top-performing models for this domain.

TABLE 15. Performance Metrics of Social & Preventive Medicine Language Models

Model	Accuracy (%)	Precision	Recall	F1 Score
LLaMA 3.1	63.64	0.5500	0.4679	0.4958
LLaMA 3	81.82	0.8699	0.8182	0.8393
Gemma 27B	63.64	0.6761	0.6364	0.6515
LLaMA 3.2	72.73	0.8182	0.7273	0.7614
Solar Pro 22B 8K	50.00	0.4905	0.3893	0.3988
Nemotron-mini FP16 4B	27.27	0.4455	0.2727	0.3333
Qwen 2.5 14B Q8	68.18	0.6875	0.6818	0.6818
Phi 3.5 FP16	63.64	0.7538	0.6364	0.6132
Mistral	36.36	0.4792	0.2821	0.3551
Mistral Small	36.36	0.3009	0.3071	0.2917

C. HIGH-PERFORMING MODELS

The models **LLaMA 3**, **Gemma 27B**, and **Qwen 2.5 14B Q8** emerged as the top-performing models, achieving accuracies of 83.60%, 76.00%, and 70.40%, respectively, with F1 scores of 0.8364, 0.7479, and 0.7044. These models demonstrated superior precision and recall, highlighting their strong capacity for understanding nuanced medical questions. The high parameter count in **LLaMA 3** and **Gemma 27B** allows them to capture complex relationships in medical data, making them well-suited for handling intricate scenarios in domains such as oncology and cardiology. Consistent performance across multiple metrics indicates their reliability in clinical applications, including diagnostic support and medical education.

D. MEMORY-EFFICIENT MODELS

Despite their lower parameter count, **Nemotron-mini FP16 4B** and **Solar Pro 22B 8K** demonstrated moderate performance, with accuracies of 23.60% and 39.20%, respectively. While their precision and recall scores were limited, their low computational demands make them suitable for resource-constrained environments like wearable health monitoring devices or rural healthcare setups. However, their struggles in identifying complex answers indicate trade-offs between computational efficiency and predictive accuracy.

E. BALANCED MODELS

Qwen 2.5 14B Q8 strikes an effective balance between performance and computational efficiency, achieving an accuracy of 70.40% and an F1 score of 0.7044. This model's quantized structure enables reduced memory usage without significant sacrifices in accuracy, making it an ideal candidate for large-scale deployment in medical systems where efficiency and reliability are critical.

F. SMALLER MODELS

Mistral Small achieved an accuracy of 45.93%, outperforming other small models like **Nemotron-mini FP16 4B** but falling short in domains requiring complex reasoning, such as neurology and oncology. Its limitations highlight the challenges smaller models face in capturing intricate relationships within medical data, emphasizing the need for scale and specialization.

G. SUBJECT-SPECIFIC ANALYSIS

The analysis revealed that high-performing models like **LLaMA 3** and **Gemma 27B** excel in disciplines requiring advanced medical reasoning, such as oncology and cardiology. In contrast, smaller models like **Nemotron-mini FP16 4B** and **Solar Pro 22B 8K** struggled in these areas due to their limited capacity. In relatively straightforward domains like general medicine and pediatrics, even smaller models demonstrated adequate performance, suggesting variability in the demands of different medical disciplines.

H. MULTILINGUAL PERFORMANCE

The study identified significant challenges in multilingual performance, with all models showing reduced accuracy in non-English contexts. This gap underscores the need for further research and fine-tuning to create models that are

language-agnostic and accessible for a global audience of medical practitioners.

I. INSIGHTS ON RETRIEVAL AUGMENTED GENERATION (RAG)

Experiments with Retrieval Augmented Generation (RAG) techniques showed potential in addressing hallucination issues and improving model grounding with up-to-date medical knowledge. However, performance gains varied across datasets and models, emphasizing the need for advanced retrieval mechanisms tailored to medical QA tasks.

J. CHALLENGES IN REASONING AND EXPLAINABILITY

While high-performing models like **LLaMA 3** and **Gemma 27B** provided accurate answers, they lacked the ability to generate detailed explanations for their predictions. This limitation highlights the importance of integrating gold-standard explanations into benchmarks like MedExpQA, enabling a more thorough evaluation of a model's reasoning capabilities.

K. KEY FINDINGS AND IMPLICATIONS

- High-performing models like **LLaMA 3** and **Gemma 27B** are well-suited for complex clinical applications due to their accuracy and reliability.
- Memory-efficient models, while less accurate, are valuable for resource-constrained environments, such as rural healthcare or wearable devices.
- Multilingual performance remains a significant challenge, necessitating further research on non-English datasets to enhance accessibility.
- Incorporating gold-standard explanations and leveraging RAG techniques can improve both reasoning capabilities and reliability in medical question-answering tasks.

V. CONCLUSION AND FUTURE WORK

This study provided a comprehensive evaluation of large language models (LLMs) for medical question-answering (QA) tasks, using a stratified dataset that covered multiple medical disciplines. The analysis revealed that high-performing models like **LLaMA 3** and **Gemma 27B** consistently excelled in accuracy, precision, and recall, making them ideal for complex clinical applications, such as diagnostic support and advanced medical education. These models demonstrated the ability to handle intricate reasoning in domains like oncology, cardiology, and neurology, while maintaining reliability across various metrics. At the same time, memory-efficient models like **Nemotron-mini FP16 4B** showcased potential in resource-constrained settings, such as rural healthcare systems or wearable devices, though their performance in more complex tasks was limited. Balanced models like **Qwen 2.5 14B Q8** presented a middle ground, offering competitive accuracy with reduced computational requirements, making them highly applicable for large-scale implementations.

However, significant challenges remain, particularly in multilingual performance, where all models showed reduced effectiveness in non-English contexts. This highlights the need for further fine-tuning and the inclusion of diverse datasets to make these models accessible to a broader range of medical practitioners worldwide.

Future research should focus on two key areas:

1) **Enhanced Multilingual Capabilities:** Expanding training datasets to include a wider range of languages and medical terminologies, enabling better support for diverse linguistic groups in healthcare.
2) **Improved Reasoning and Explainability:** Incorporating structured knowledge bases and benchmarks that require models to provide detailed explanations alongside their predictions, ensuring both accuracy and clarity in clinical contexts.

AI-powered medical question-answering systems hold vital importance for India's healthcare because the population faces major obstacles in reaching specialist doctors. Healthcare facilities located in rural areas face difficulties because they maintain insufficient medical staff which results in late diagnosis and improper medical determination. This AI assistant functions as a second opinion system when assisted by LLMs + RAG to help medical staff obtain evidence-based quick recommendations.

The approach fits the expanding telemedicine sector of India because AI healthcare bridges the distance between specialized medical professionals in urban areas and practicing doctors in rural regions. AI functions best when humans implement ethical and legal security protocols to maintain doctor assistance over full automation of medical practice. The chosen system should maintain transparency along with interpretability while following rules set by the Indian Medical Council. Through the RAG system organizations can strengthen their accountability by providing verifiable references which prevents the generation of recommendations without proper evidence.

The structured use of AI by India can produce substantial healthcare accessibility improvements in remote areas combined with high-quality medical expertise delivery to the entire population through safe systems compliant with regulations.

Moreover, future studies should explore how these models can be integrated into real-world healthcare workflows, such as automated triage systems, personalized patient care, and medical training simulators. Evaluating their performance on larger, real-world datasets and in dynamic clinical settings will be critical to assessing their practical utility. The development of lightweight, memory-efficient variants of high-performing models could also provide a pathway to democratizing access to advanced medical AI in low-resource environments. These advancements will pave the way for the next generation of medical AI systems that are accurate, interpretable, and widely accessible.

REFERENCES

[1] K. Singhal et al., "Towards Expert-Level Medical Question Answering with Large Language Models," arXiv preprint arXiv:2305.09617, 2023.

[2] H. Nori et al., "Capabilities of GPT-4 on Medical Challenge Problems," arXiv preprint arXiv:2303.13375, 2023.

[3] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint arXiv:2302.13971, 2023.

[4] C. Wu et al., "PMC-LLaMA: Further Finetuning LLaMA on Medical Papers," arXiv preprint arXiv:2304.07990, 2023.

[5] Z. Jiang et al., "Mistral: Efficient Supervision for Large Language Models with Mixture of Experts," arXiv preprint arXiv:2305.10401, 2023.

[6] Y. Labrak et al., "BioMistral: A Collection of Open-Source Pre-trained Large Language Models for Medical Domains," arXiv preprint arXiv:2402.10373, 2024.

[7] N. Yagnik et al., "MedLM: Exploring Language Models for Medical Question Answering Systems," arXiv preprint arXiv:2401.11389, 2024.

[8] A. Pal et al., "MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical Question Answering," arXiv preprint arXiv:2203.14371, 2022.

[9] M. Cascella et al., "The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives," *Journal of Medical Systems*, vol. 48, no. 10, pp. 45–52, 2024.

[10] Y. Gao et al., "Large Language Models and Medical Knowledge Grounding for Diagnosis Prediction," medRxiv preprint, 2024.

[11] D. Khullar et al., "Large Language Models in Health Care: Charting a Path Toward Accurate, Explainable, and Secure AI," *Journal of General Internal Medicine*, vol. 39, no. 5, pp. 457–467, 2024.

[12] D. Khullar et al., "Assessing the Research Landscape and Clinical Utility of Large Language Models in Medicine," *BMC Medical Informatics and Decision Making*, 2024.

[13] H. Chen et al., "Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions," arXiv preprint arXiv:2402.18060, 2024.

[14] H. Chen et al., "Large Language Models Encode Clinical Knowledge," arXiv preprint arXiv:2212.13138, 2023.

[15] J. Wang et al., "JMLR: Joint Medical LLM and Retrieval Training for Enhancing Reasoning and Professional Question Answering Capability," arXiv preprint arXiv:2402.17887, 2024.

[16] R. Elgedawy et al., "Dynamic Q&A of Clinical Documents with Large Language Models," arXiv preprint arXiv:2401.10733, 2024.

[17] R. Elgedawy et al., "A Study of Generative Large Language Model for Medical Applications," *npj Digital Medicine*, 2024.

[18] Nori, Harsha, et al. "Capabilities of GPT-4 on Medical Challenge Problems." *ArXiv preprint arXiv:2303.13375* (2023).

[19] K. Singhal et al., "Med-PaLM: A Large Language Model for Medicine," arXiv preprint arXiv:2305.09617, 2023.

[20] M. Cascella et al., "ChatGPT: Applications and Limitations in Healthcare," arXiv preprint arXiv:2401.10733, 2024.

[21] Wang, D. and Zhang, S., 2024. Large language models in medical and healthcare fields: applications, advances, and challenges. Artificial Intelligence Review, 57(11), p.299.

[22] Park, Y.J., Pillai, A., Deng, J., Guo, E., Gupta, M., Paget, M. and Naugler, C., 2024. Assessing the research landscape and clinical utility of large language models: A scoping review. BMC Medical Informatics and Decision Making, 24(1), p.72.

[23] Moglia, A., Georgiou, K., Cerveri, P., Mainardi, L., Satava, R.M. and Cuschieri, A., 2024. Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test. Artificial Intelligence Review, 57(9), p.231.

[24] Duong, B.T. and Le, T.H., 2024, November. MedQAS: A Medical Question Answering System Based on Finetuning Large Language Models. In International Conference on Future Data and Security Engineering (pp. 297-307). Singapore: Springer Nature Singapore.

[25] Li, X. and Wu, M., 2024, December. GKF-mQA: Generative Knowledge Fusion Based on Large Language Models for Enhancing Medical Question Answering. In International Conference on Advanced Data Mining and Applications (pp. 215-229). Singapore: Springer Nature Singapore.

[26] Kharitonova, K., Pérez-Fernández, D., Gutiérrez-Hernando, J., Gutiérrez-Fandiño, A., Callejas, Z. and Griol, D., 2024, October. Leveraging Retrieval-Augmented Generation for Reliable Medical Question Answering Using Large Language Models. In International Conference on Hybrid Artificial Intelligence Systems (pp. 141-153). Cham: Springer Nature Switzerland.

[27] Ucar, A., Nayak, S., Roy, A., Taşcı, B. and Taşcı, G., 2025. A Comprehensive Study on Fine-Tuning Large Language Models for Medical Question Answering Using Classification Models and Comparative Analysis. arXiv preprint arXiv:2501.17190.

[28] Ho, C.N., Tian, T., Ayers, A.T., Aaron, R.E., Phillips, V., Wolf, R.M., Mathioudakis, N., Dai, T. and Klonoff, D.C., 2024. Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. BMC Medical Informatics and Decision Making, 24(1), p.357.

[29] Cheligeer, K., Wu, G., Laws, A., Quan, M.L., Li, A., Brisson, A.M., Xie, J. and Xu, Y., 2024. Validation of large language models for detecting pathologic complete response in breast cancer using population-based pathology reports. BMC Medical Informatics and Decision Making, 24(1), p.283.

[30] Yang, R., Marrese-Taylor, E., Ke, Y., Cheng, L., Chen, Q. and Li, I., 2023. Integrating umls knowledge into large language models for medical question answering. arXiv e-prints, pp.arXiv-2310.

[31] Wang, J., Ning, H., Peng, Y., Wei, Q., Tesfai, D., Mao, W., Zhu, T. and Huang, R., 2024. A Survey on Large Language Models from General Purpose to Medical Applications: Datasets, Methodologies, and Evaluations. arXiv preprint arXiv:2406.10303.

[32] McCoy, L.G., Ci Ng, F.Y., Sauer, C.M., Yap Legaspi, K.E., Jain, B., Gallifant, J., McClurkin, M., Hammond, A., Goode, D., Gichoya, J. and Celi, L.A., 2024. Understanding and training for the impact of large language models and artificial intelligence in healthcare practice: A narrative review. BMC Medical Education, 24(1), p.1096.

[33] Sallam, M., 2023. The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. MedRxiv, pp.2023-02.

[34] Tascon Morales, S., Márquez Neila, P. and Sznitman, R., 2024. Targeted Visual Prompting for Medical Visual Question Answering. In Applications of Medical Artificial Intelligence: Third International Workshop, AMAI 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 6, 2024, Proceedings (p. 64). Springer Nature.

[35] Silvestri, C., Roshal, J., Shah, M., Widmann, W.D., Townsend, C., Brian, R., LHuillier, J.C., Navarro, S.M., Lund, S. and Sathe, T.S., 2024. Evaluation of a Novel Large Language Model (LLM) Powered Chatbot for Oral-Boards Scenarios. medRxiv, pp.2024-05.

[36] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y. and Ye, W., 2024. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), pp.1-45.

[37] Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A. and Chartash, D., 2023. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR medical education, 9(1), p.e45312.

[38] Moglia, A., Georgiou, K., Cerveri, P., Mainardi, L., Satava, R.M. and Cuschieri, A., 2024. Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test. Artificial Intelligence Review, 57(9), p.231.

[39] Cascella, M., Semeraro, F., Montomoli, J., Bellini, V., Piazza, O. and Bignami, E., 2024. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. Journal of Medical Systems, 48(1), p.22.

[40] Cascella, M., Semeraro, F., Montomoli, J., Bellini, V., Piazza, O. and Bignami, E., 2024. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. Journal of Medical Systems, 48(1), p.22.

[41] Cascella, M., Semeraro, F., Montomoli, J., Bellini, V., Piazza, O. and Bignami, E., 2024. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. Journal of Medical Systems, 48(1), p.22.

[42] Avnat, E., Levy, M., Herstain, D., Yanko, E., Joya, D.B., Katz, M.T., Eshel, D., Laros, S., Dagan, Y., Barami, S. and Mermelstein, J., 2024. Performance of large language models in numerical vs. semantic medical knowledge: Benchmarking on evidence-based Q&As. arXiv preprint arXiv:2406.03855.

[43] Yagnik, N., Jhaveri, J., Sharma, V., Pila, G., Ben, A. and Shang, J., 2024. MedLM: Exploring Language Models for Medical Question Answering Systems. arXiv preprint arXiv:2401.11389.

[44] McCoy, L.G., Ci Ng, F.Y., Sauer, C.M., Yap Legaspi, K.E., Jain, B., Gallifant, J., McClurkin, M., Hammond, A., Goode, D., Gichoya, J. and Celi, L.A., 2024. Understanding and training for the impact of large language

models and artificial intelligence in healthcare practice: A narrative review. BMC Medical Education, 24(1), p.1096.

[45] Bean, A.M., Korgul, K., Krones, F., McCraith, R. and Mahdi, A., 2023. Exploring the landscape of large language models in medical question answering. arXiv e-prints, pp.arXiv-2310.

[46] Zekaoui, N.E., Yousfi, S., Mikram, M. and Rhanoui, M., 2023, November. Enhancing Large Language Models' Utility for Medical Question-Answering: A Patient Health Question Summarization Approach. In 2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA) (pp. 1-8). IEEE.

[47] Liévin, V., Hother, C.E., Motzfeldt, A.G. and Winther, O., 2024. Can large language models reason about medical questions?. Patterns, 5(3).

[48] Vladika, J., Schneider, P. and Matthes, F., 2024. MedREQAL: Examining Medical Knowledge Recall of Large Language Models via Question Answering. arXiv preprint arXiv:2406.05845.

[49] Kharitonova, K., Pérez-Fernández, D., Gutiérrez-Hernando, J., Gutiérrez-Fandiño, A., Callejas, Z. and Griol, D., 2024, October. Leveraging Retrieval-Augmented Generation for Reliable Medical Question Answering Using Large Language Models. In International Conference on Hybrid Artificial Intelligence Systems (pp. 141-153). Cham: Springer Nature Switzerland.

[50] Artsi, Y., Sorin, V., Konen, E., Glicksberg, B.S., Nadkarni, G. and Klang, E., 2024. Large language models for generating medical examinations: systematic review. BMC Medical Education, 24(1), p.354.

[51] Chen, H., Fang, Z., Singla, Y. and Dredze, M., 2024. Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions. arXiv preprint arXiv:2402.18060.

[52] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S. and Payne, P., 2022. Large language models encode clinical knowledge. arXiv preprint arXiv:2212.13138.

[53] Bean, A.M., Korgul, K., Krones, F., McCraith, R. and Mahdi, A., 2023. Exploring the landscape of large language models in medical question answering. arXiv e-prints, pp.arXiv-2310.

[54] Yang, R., Marrese-Taylor, E., Ke, Y., Cheng, L., Chen, Q. and Li, I., 2023. Integrating umls knowledge into large language models for medical question answering. arXiv e-prints, pp.arXiv-2310.

[55] Li, X. and Wu, M., 2024, December. GKF-mQA: Generative Knowledge Fusion Based on Large Language Models for Enhancing Medical Question Answering. In International Conference on Advanced Data Mining and Applications (pp. 215-229). Singapore: Springer Nature Singapore.

[56] He, Z., Bhasuran, B., Jin, Q., Tian, S., Hanna, K., Shavor, C., Arguello, L.G., Murray, P. and Lu, Z., 2024. Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study. Journal of Medical Internet Research, 26, p.e56655.

[57] McCoy, L.G., Ci Ng, F.Y., Sauer, C.M., Yap Legaspi, K.E., Jain, B., Gallifant, J., McClurkin, M., Hammond, A., Goode, D., Gichoya, J. and Celi, L.A., 2024. Understanding and training for the impact of large language models and artificial intelligence in healthcare practice: A narrative review. BMC Medical Education, 24(1), p.1096.

[58] Li'evin, V., Hother, C., & Winther, O. (2022). Can large language models reason about medical questions?. Patterns, 5. https://doi.org/10.1016/j.patter.2024.100943.

[59] Lucas, M., Yang, J., Pomeroy, J., & Yang, C. (2024). Reasoning with large language models for medical question answering. Journal of the American Medical Informatics Association : JAMIA. https://doi.org/10.1093/jamia/ocae131.

[60] Guo, Q., Cao, S., & Yi, Z. (2022). A medical question answering system using large language models and knowledge graphs. International Journal of Intelligent Systems, 37, 8548 - 8564. https://doi.org/10.1002/int.22955.

[61] Chen, Y., Wang, Z., Wen, B., & Zulkernine, F. (2024). Comparative Analysis of Open-Source Language Models in Summarizing Medical Text Data. 2024 IEEE International Conference on Digital Health (ICDH), 126-128. https://doi.org/10.1109/ICDH62654.2024.00030.

[62] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P., Rouvier, M., & Dufour, R. (2024). BioMistral: A Collection of Open-Source Pre-trained Large Language Models for Medical Domains. , 5848-5864. https://doi.org/10.48550/arXiv.2402.10373.

[63] Wei, Q., Cui, Y., Ding, M., Wang, Y., Xiang, L., Yao, Z., Chen, C., Long, Y., Jin, Z., & Xu, X. (2024). Performance Evaluation of Lightweight Open-source Large Language Models in Pediatric Consultations: A Comparative Analysis. ArXiv, abs/2407.15862. https://doi.org/10.48550/arXiv.2407.15862.

[64] Zhu, J. (2024). Cura-LLaMA: Evaluating open-source large language models question answering capability on medical domain. Applied and Computational Engineering. https://doi.org/10.54254/2755-2721/90/20241725.

[65] Soman, K., Rose, P., Morris, J., Akbas, R., Smith, B., Peetoom, B., Villouta-Reyes, C., Cerono, G., Shi, Y., Rizk-Jackson, A., Israni, S., Nelson, C., Huang, S., & Baranzini, S. (2023). Biomedical knowledge graph-optimized prompt generation for large language models. Bioinformatics, 40. https://doi.org/10.1093/bioinformatics/btae560.

[66] Zhang, G., Jin, Q., Zhou, Y., Wang, S., Idnay, B., Luo, Y., Park, E., Nestor, J., Spotnitz, M., Soroush, A., Campion, T., Lu, Z., Weng, C., & Peng, Y. (2024). Closing the gap between open-source and commercial large language models for medical evidence summarization. ArXiv. https://doi.org/10.48550/arXiv.2408.00588.