

Evaluation of Language Models (LLMs) in the Interpretation of Tuberculosis Concepts: Design of a Virtual Assistant for Clinical Support and Patient Education

Mauricio Acevedo-Carrillo and Shirley Fiorella Simbron Espejo

Private University of the North, Lima, Peru

Article history

Received: 24-11-2025

Revised: 22-07-2026

Accepted: 03-08-2026

Corresponding Author:

Mauricio Acevedo-Carrillo
Private University of the North,
Lima, Peru
Email:
mauricio.acevedo@upn.edu.pe

Abstract: Tuberculosis (TB) remains a global public health challenge, where access to accurate and up-to-date information is crucial for professionals and patients. This research comparatively evaluated the performance of five language models (LLMs) Med-PaLM 2, GPT-4, MEDITRON-70B, Me-LLaMA, and Clinical Camel in the TB domain and implemented a mobile virtual assistant as a proof-of-concept based on the best-performing model. Clinical accuracy was defined as the percentage of responses considered correct based on expert consensus evaluation. The methodology included evaluation with and without Retrieval-Augmented Generation (RAG) using a corpus of 150 clinically validated questions. Results showed that RAG significantly improved clinical accuracy (94.0% vs. 82.3%) and reduced hallucination rates (2.3% vs. 8.7%). Gemini 1.5 achieved the highest performance (96.4% with RAG), while open-source models such as MEDITRON-70B (89.5%) demonstrated competitive performance. These findings suggest that RAG enhances reliability in domain-specific medical applications. The implementation of a mobile virtual assistant is presented as a proof-of-concept prototype, derived from the evaluation results, and not as a clinically validated system for real-world deployment. This study contributes to the evaluation of LLMs in tuberculosis by integrating accuracy, safety, and applicability considerations, particularly for low-resource settings.

Keywords: Language Models, Tuberculosis, RAG, Clinical Evaluation, Virtual Assistant, Medical Artificial Intelligence

Introduction

Tuberculosis remains a serious threat to global health, causing 1.3 million deaths in 2022 according to the WHO (World Health Organization, 2023). Despite being a preventable and curable disease, its impact is exacerbated by late diagnosis, the emergence of multidrug-resistant strains (MDR/XDR-TB), and barriers to accessing reliable medical information (Keshavjee and Farmer, 2012). Healthcare professionals face difficulties in keeping up to date with evolving clinical guidelines, while patients lack accessible sources to resolve questions about treatments or side effects (Furin, 2019).

Large Language Models (LLMs) emerge as potential tools to address these challenges (Topol,

2019). However, their application in tuberculosis carries significant risks. Recent studies (Ji et al., 2023, CDC, 2025a) warn that they can generate hallucinations (fictitious data) or biases, such as confusing therapeutic guidelines for sensitive and resistant TB, or omitting crucial considerations for comorbidities such as TB/HIV co-infection. This lack of reliability requires rigorous evaluation before any clinical implementation.

Currently, AI applications such as Ada Health or Buoy Assistant offer general medical assistance, but do not address the specific complexities of tuberculosis. Their availability is often limited to major languages, excluding vulnerable communities where TB is prevalent (Liu et al., 2025). A virtual assistant specialized in TB, based on an accurate and

up-to-date linguistic model, could overcome these limitations by providing contextually relevant information:

- For experts: Fast access to medical recommendations (WHO, CDC), help with differentiating diagnoses, and information on drug resistance changes
- For patients: Learning resources in simple language, reminders for medication, and notifications regarding negative effects, which enhance following treatment plans (González et al., 2021)

This study proposes a novel method for assessing Large Language Models (LLMs) within the area of Tuberculosis (TB), focusing on aspects like precision, comprehensibility, and relevance to clinical practice. The results will aid in developing a virtual assistant model aimed at regions with limited resources, with potential applications in other infectious diseases as well. Its influence could extend from urban health facilities to remote communities, where there are few experts but increasing mobile access (WHO, 2025).

Theoretical Framework

Tuberculosis: Clinical Aspects and Global Challenges

Tuberculosis, caused by *Mycobacterium tuberculosis*, is an infectious disease that is transmitted through the air and mainly affects the lungs. Its close link to social determinants such as poverty and overcrowding makes it a complex and persistent public health problem (World Health Organization, 2023). Although rapid molecular diagnostic advances exist, such as GeneXpert MTB/RIF, the challenges to its global control remain critical:

- Drug resistance: Multidrug-Resistant Tuberculosis (MDR-TB) needs a long course of treatment lasting from 20 to 24 months using second-line medications, which tend to be more harmful and harder to obtain (Jang and Chung, 2020)
- TB-HIV coinfection: Individuals with HIV face a 20 times greater risk of developing active tuberculosis, which necessitates combined treatment for both illnesses (CDC, 2025b)
- Treatment discontinuation: Insufficient adherence (due to side effects or stigma) promotes relapse and resistance (Camarín-Sánchez et al., 2020)

These difficulties highlight the necessity for resources that assist professionals in making choices and inform patients, particularly in areas lacking enough specialists.

Language Models (LLMs) in Healthcare: Opportunities and Risks

Large Language Models (LLMs), such as GPT-4, offer significant potential in medicine to synthesize literature, support diagnoses, and educate patients by translating complex terminology (González et al., 2021). However, their clinical application is limited by the generation of hallucinations; one study reported that 35% of GPT-4's medical responses contained serious errors (Jin et al., 2024), biases in underrepresented populations, and the inability to update in real time with clinical guidelines (Topol, 2019). Strategies such as Retrieval-Augmented Generation (RAG), which anchors LLMs to updated databases, are emerging as promising solutions to mitigate these risks.

Virtual Health Assistants: Lessons Learned

Medical assistants such as Ada Health or Buoy Health demonstrate that the effectiveness of chatbots depends on rigorous clinical validation and supervised integration with healthcare systems. Connection with health services: Participants like Symptomate direct urgent situations to hospitals, preventing harmful self-diagnosis. These cases reinforce the need to develop specific tools, under human supervision, to manage complex and stigmatized diseases such as tuberculosis.

Related Jobs

Recent advances in artificial intelligence applied to tuberculosis can be grouped into three main areas:

- (i) AI for diagnosis and prediction
- (ii) Chatbot-based systems for patient education
- (iii) Large language models enhanced with Retrieval Augmented Generation (RAG)

In recent years, Artificial Intelligence (AI) has established itself as a fundamental tool in the fight against tuberculosis (TB). Deep learning models, such as convolutional neural networks, have demonstrated accuracy comparable to that of expert radiologists in the automated detection of TB in chest X-rays, which is promising for resource-limited settings (Lakhani and Sundaram, 2017).

Beyond diagnosis, AI can predict drug resistance by analyzing genomic information from *Mycobacterium tuberculosis*. These models facilitate the choice of personalized treatments, a crucial advantage in the face of the global increase in multidrug-resistant TB (Yang et al., 2018). In the field of patient education, rule-based assistants such as TB-Chatbot (Octoco, 2025) have achieved 82% accuracy in responding to common queries, although their ability to address complex questions about drug resistance remains a challenge

(Mkhize et al., 2022).

AI is revolutionizing the management of tuberculosis. In drug development, deep learning algorithms speed up the discovery of new antibacterial compounds against resistant strains (Stokes et al., 2020). At the same time, computer vision systems verify treatment adherence through videos, significantly improving treatment outcomes (Story et al., 2016). These innovations position artificial intelligence as a key tool in advancing toward the global elimination of disease.

Despite significant progress, existing studies often lack integrated evaluation frameworks that simultaneously address clinical safety, hallucination mitigation, and applicability in low-resource settings. This study contributes to the field by combining domain-specific evaluation in tuberculosis with safety-oriented metrics and RAG-based enhancement, offering a more comprehensive assessment of LLM performance in clinical contexts.

Materials and Methods

The suggested approach consists of two key stages:

- (1) Assessing various LLMs related to tuberculosis (TB)
- (2) Creating and putting into action a virtual assistant using the most effective model

Both stages combine numerical and descriptive techniques to maintain scientific accuracy and relevance in clinical settings.

To improve methodological clarity and reproducibility, the overall workflow of the study follows a structured pipeline consisting of four main stages:

- (1) Selection and characterization of language models
- (2) Construction and validation of a clinically grounded evaluation dataset
- (3) Implementation of a Retrieval-Augmented Generation (RAG) framework
- (4) Multi-dimensional evaluation using accuracy, safety, and usability metrics

This sequential design ensures consistency between experimental setup and evaluation outcomes, facilitating replication and interpretation of results. Fig. 1 illustrates the overall methodological workflow, highlighting the interaction between dataset construction, RAG integration, and model evaluation.

Evaluation of Language Models (LLMs)

Model Selection

The evaluation of linguistic models for medical use in tuberculosis requires a selection that considers both

traditional models and recent open-source contributions. For this analysis, five models representing different development strategies were chosen, two high-performance commercial models, two open-access models specializing in health, and one open-source model adapted for this purpose (Fig. 2). This selection allows for a comprehensive comparison of their structures, training methods, and capabilities in the specific field of tuberculosis, considering both clinical accuracy and feasibility of implementation in various settings.

The selection of the five evaluated LLMs was guided by representativeness criteria to ensure a comprehensive and balanced comparison. Specifically, the study includes:



Fig. 1: Methodological process flow

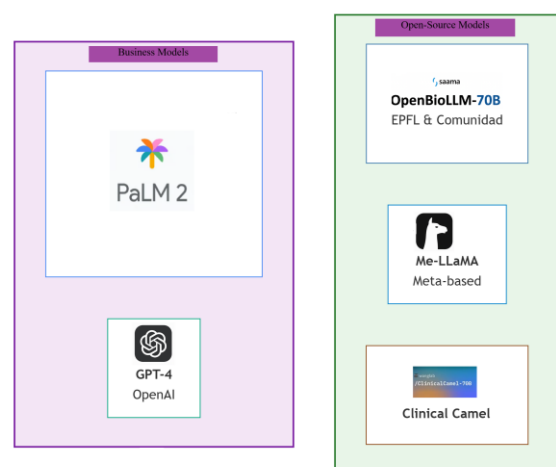


Fig. 2: Language Models Evaluated in the Field of Tuberculosis

- (i) High-performance commercial models (e.g., Med-PaLM 2, GPT-4) that serve as performance benchmarks;
- (ii) Open-source models (e.g., MEDITRON-70B, Me-LLaMA, Clinical Camel), which enable reproducibility, adaptability, and deployment in low-resource settings
- (iii) Models specifically optimized for medical applications, ensuring domain relevance

This combination allows the evaluation to capture differences not only in performance but also in accessibility, scalability, and suitability for real-world healthcare scenarios.

Med-PaLM 2 (Google) is a benchmark in language models for medicine (Med-PaLM). Based on the PaLM architecture and fine-tuned with extensive validated clinical datasets, this model demonstrates exceptional capabilities for medical reasoning and scientific literature interpretation. Its training, which encompasses trillions of tokens from anonymized medical records, manuals, and research, allows it to achieve performance comparable to that of human experts in various tasks, especially in the application of established protocols.

GPT-4 (OpenAI) is a cutting-edge language model whose training on diverse texts, including medical ones, gives it robust capabilities for complex clinical tasks (Achiam et al., 2023). Its advanced Transformer architecture and reinforcement learning techniques enable it to perform stepwise reasoning and generate clinically relevant responses, achieving performance comparable to specialized models in standardized tests, especially in the integration of multidisciplinary knowledge.

MEDITRON-70B is an open-source model for medical applications, fine-tuned from Llama 2 with an extensive corpus of clinical literature. Trained with more than 1.5 billion tokens, including practice guidelines, PubMed publications, and educational material, it has specific capabilities for handling medical terminology, relationships between clinical entities, and diagnostic processes. This focus makes it a viable alternative for environments with economic constraints or that require transparency in the process.

Me-LLaMA (Xie et al., 2024) introduces a novel paradigm for adapting base models to the medical domain through continuous fine-tuning and specific alignment techniques. Based on LLaMA 2, it employs Reinforcement Learning with Human Feedback (RLHF) in collaboration with healthcare experts. This approach allows the model to acquire specialized skills while maintaining the versatility of the original model, with a modular design that facilitates specific adaptations for areas such as infectious diseases.

Clinical Camel is a specialized model, fine-tuned from LLaMA with extensive clinical datasets (Toma et al., 2023). Its training with labeled medical cases,

consultation transcripts, and continuing education materials gives it remarkable capabilities for diagnostic reasoning and clinical communication adapted to different levels of health literacy, making it useful in various healthcare contexts.

While the inclusion of high-performance commercial models such as Med-PaLM 2 and GPT-4 provides a strong reference for benchmarking, it also introduces limitations in terms of reproducibility and accessibility. These models are closed-source and require restricted API access, which may hinder independent validation, transparency, and deployment in low-resource settings. This limitation is particularly relevant in global health contexts such as tuberculosis, where equitable access to technology is essential.

Table 1 presents a comparison of five modern language models that are significant for evaluating clinical understanding of tuberculosis. Med-PaLM 2 and GPT-4 are notable for their strong performance without the need for further training, establishing them as standards for conceptual precision. However, both are closed models and do not allow direct adaptation to local clinical contexts (e.g., Peruvian TB guidelines, MDR-TB prevalence, or national diagnostic flows). On the other hand, MEDITRON-70B, Me-LLaMA, and Clinical Camel are open models, allowing their use in research, complementary training, integration with RAG systems, and fine-tuning using local literature or ministerial guidelines. Among them, MEDITRON-70B offers the best balance between high clinical performance and openness, although it requires greater computational capacity to run. Me-LLaMA is a lighter and more flexible alternative, while Clinical Camel is especially useful for building educational evaluators (e.g., conceptual feedback or comprehension tests).

Dataset Validation

Three pulmonologists independently evaluated the 150 clinical questions using a three-category ordinal scale (appropriate, partially appropriate, and inappropriate). Inter-rater agreement was assessed using Fleiss's kappa coefficient ($\kappa = 0.75$; 95% CI: 0.69–0.81; $p < 0.001$), representing a substantial level of agreement according to the Landis and Koch classification. Discrepancies were resolved by consensus prior to finalizing the evaluation set.

The dataset construction followed a structured multi-stage process. First, clinically relevant topics were identified based on WHO and CDC tuberculosis guidelines (CDC, 2024c). Second, questions were designed to reflect real-world clinical scenarios, including differential diagnosis, treatment decision-making, and public health interventions. Third, all items were reviewed and validated by clinical experts to ensure accuracy and relevance.

Table 1: Comparison of LLMs with a focus on clinical knowledge (applicable to tuberculosis)

Model	Type/License	Size (approx.)	Medical approach	Key Benefits for TB	Limitations
Med-PaLM 2 (Google)	Closed / Commercial	Not publicly disclosed	High, specializing in clinical medicine.	High performance in clinical questions, useful as a comparative reference for accuracy.	Not customizable; restricted access; less suitable for local adaptation.
GPT-4 (OpenAI)	Closed / Commercial (access via API)	Not publicly disclosed	Generally, with high performance in medicine.	Excellent reasoning and explanation; serves as a "reference model" in conceptual evaluation.	It cannot be adjusted with local data; it may go haywire if local guidelines are not explicit.
MEDITRON-70B (EPFL / Open Source)	Open (permissive for research)	70B parameters	Trained in biomedical literature.	Highly suitable for local adaptation (fine-tuning + RAG); improved traceability of sources.	Requires powerful hardware (GPU) for implementation.
Me-LLaMA	Open	13B / 70B depending on the version	Specifically optimized for biomedical language. Adjusted through conversational learning in clinical data.	Scalable to Spanish, useful for building concept evaluators in TB.	Even fewer public clinical evaluations than MEDITRON.
Clinical Camel	Open	Based on LLaMA-2		Very useful for educational tasks: explanation, feedback, evaluation of concepts.	Slightly lower performance than top commercial models.

Representative examples of question types include:

- (i) Differential diagnosis scenarios (e.g., pulmonary tuberculosis vs. fungal infections)
- (ii) Treatment adjustments in drug-resistant TB cases
- (iii) Prevention strategies in community-level interventions

Each question was accompanied by metadata specifying difficulty level, clinical context, and source references.

The Treatment section (65 questions) focuses on therapeutic regimens for sensitive and resistant TB, including management of adverse effects and adherence. All answers were validated by pulmonologists with experience in TB control programs (CDC, 2025b), ensuring clinical applicability in various healthcare settings.

Finally, the Prevention and Public Health category (25 items) completes the set by assessing knowledge of chemoprophylaxis, infection control, and community actions. The structure of the corpus incorporates multiple levels of difficulty and considers characteristics of settings with a high disease burden and limited resources. Each question includes metadata describing the clinical context and sources of evidence, establishing a robust standard for measuring the capabilities of linguistic models in the field of tuberculosis and promoting the reproducibility of future research.

Inter-rater agreement was quantified using Fleiss' kappa coefficient (κ), together with its 95% confidence interval and statistical significance. Agreement was interpreted according to the Landis and Koch classification.

Application of Retrieval-Augmented Generation (RAG)

The adoption of a Retrieval-Augmented Generation (RAG) system is crucial to ensure accurate and up-to-date responses in tuberculosis models. This system operates in two phases: First, it retrieves relevant information from a repository of specialized knowledge on TB; second, it uses that contextual information to enable the model to generate informed responses (Fig. 3). The repository integrates reliable sources such as the latest WHO guidelines (2020-2024), CDC protocols, and scientific articles from PubMed from the last five years.

Evaluation Metrics

The performance evaluation of tuberculosis models uses comprehensive indicators that assess medical accuracy and practical applicability.

Metrics Classification

Evaluation metrics were categorized as:

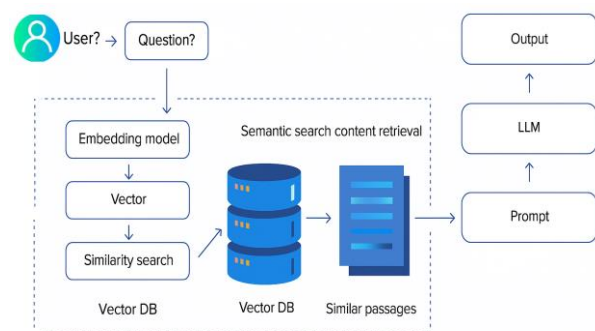


Fig. 3: RAG application

- Primary: Clinical accuracy, hallucination rate
- Secondary: Precision, recall, F1-score
- Exploratory: SUS score, latency, readability (FKGL)

Accuracy calculates the percentage of correct responses according to expert consensus. Precision measures the proportion of clinically appropriate recommendations, while Comprehensiveness assesses the ability to identify all relevant elements in a clinical scenario. The F1 score harmoniously combines these last two metrics:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \times 100$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Safety metrics are crucial in tuberculosis, where errors have serious consequences. These include: The Hallucination Rate, which quantifies incorrect or unverified medical information; the Dangerous Recommendation Rate, which identifies potentially harmful suggestions; and the Clinical Conservatism Index, which assesses the model's willingness to recognize its limitations and refer to a specialist when necessary:

$$\text{Hallucination Rate} = \frac{\text{Statements containing incorrect information}}{\text{Total responses}} \times 100$$

$$\text{Statements Recommendations Rate} = \frac{\text{Dangerous Recommendations}}{\text{Total Recommendations}} \times 100$$

$$\text{Clinical Conservatism Index} = \frac{\text{Cases with recognition of limitations}}{\text{Cases exceeding jurisdiction}} \times 100$$

Practical utility metrics assess the applicability of responses in real clinical settings. The adapted System Usability Scale (SUS) measures professionals' perceptions, while Contextual Feasibility assesses viability in various healthcare settings. Explanatory Clarity, measured using the Flesch-Kincaid index, ensures that the language is accessible to different levels of health education and literacy:

$$\text{SUS Score} = \sum_{i=1}^{10} (\text{Puntuacion_item}_i - 1) \times 2.5$$

$$\text{Contextual Feasibility} = \frac{\text{Feasible recommendations}}{\text{Total Recommendations}} \times 100$$

$$\text{FKGL} = 0.39 \left(\frac{\text{Total word}}{\text{Total number of sentences}} \right)^s + 11.8 \left(\frac{\text{Totalsyllable}}{\text{Total words}} \right) - 15.59$$

Finally, efficiency metrics evaluate operational performance for viable clinical implementation. Response latency, a fundamental metric, measures the time elapsed between the query and the complete generation of the response, calculated using the formula [indicate formula]. This aspect is critical for practical adoption in healthcare settings.

$$\text{Response Latency} = \sum_{i=1}^n \left(\frac{t_{\text{finish}_i} - t_{\text{start}_i}}{n} \right)$$

Where t_{start_i} is the start of the query, t_{finish_i} is the end of the response, and n is the total number of questions. In tuberculosis, where speed is crucial for patient health and epidemiological containment, a maximum latency of 2 seconds is established for simple queries (e.g., dosage) and 5 seconds for complex queries (e.g., differential diagnosis or drug-resistant TB).

Latency thresholds (2–5 seconds) were defined based on acceptable response times reported in clinical decision support systems, where excessive delay may impact usability in real-world scenarios.

Although the integration of RAG significantly improves clinical accuracy and reduces hallucinations, it introduces additional computational overhead that increases response latency. This trade-off between accuracy and response time is a critical consideration in clinical decision support systems, where both reliability and timeliness are essential.

Tools

RAGAS (Es et al., 2024) is a framework for automatically evaluating Retrieval-Augmented Generation (RAG) systems. In the tuberculosis application, it measures key metrics such as factual fidelity (accuracy of information against sources), contextuality (relevance of retrieved information), and comprehensiveness (ability to retrieve all relevant data). Its use enables continuous evaluation, ensuring that content on TB diagnosis, treatment, and control meets high standards of clinical accuracy and is kept up to date with the latest medical guidelines.

Structured prompting with few-shot learning chains (Patil et al., 2024) is a key technique for optimizing language models in the field of tuberculosis. This method uses prompts designed with representative examples of clinical cases and their appropriate responses, providing specialized context to the model. For the mobile application, templates have been created that include typical cases of sensitive and resistant TB, management of adverse drug effects, and differential diagnosis situations. This strategy effectively adapts general models to the specialized field of tuberculosis, improving the accuracy of responses for physicians and patients while maintaining the computational efficiency necessary for a mobile application.

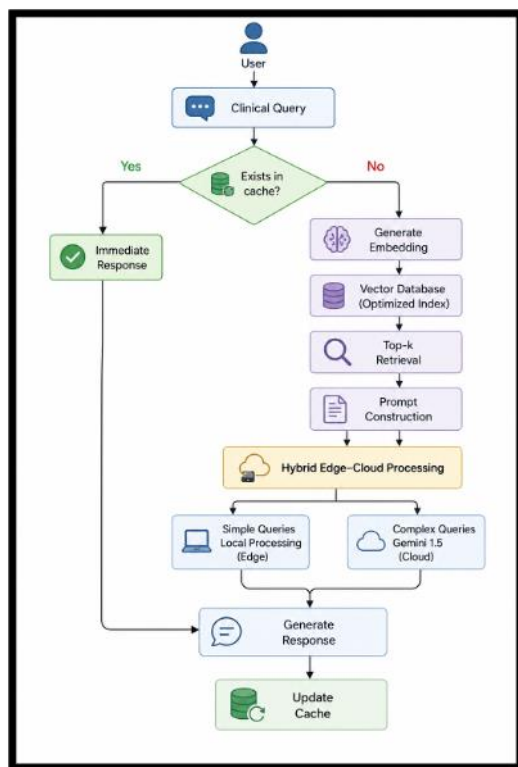


Fig. 4: Workflow of the RAG Optimization Process

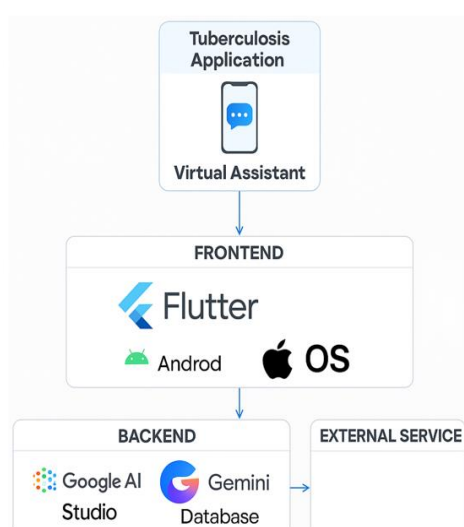


Fig. 5: Architectural Design

To enhance the performance of the Retrieval-Augmented Generation (RAG) system, an optimization strategy based on three complementary components was implemented: Caching of frequent queries, optimization of vector repository indexing, and a hybrid edge-cloud processing architecture. The caching mechanism reduces repetitive queries and lowers response latency, while optimized indexing facilitates the efficient retrieval of

the most relevant clinical information. Furthermore, the hybrid architecture allows simple queries to be processed locally on the device and more complex queries to be offloaded to cloud services, balancing accuracy, response time, and computational resource consumption as illustrated in Figure 5, which summarizes the optimization workflow implemented in the study.

Integration of the Virtual Assistant

The digital assistant will be implemented as a native app that works on multiple platforms (Android/iOS), created with Flutter, incorporating the Gemini 1.5 model through the Google AI Studio API to ensure that it is scalable and consumes few resources. The mobile structure is illustrated in Fig. 4.

AI Layer on Device and Cloud

Hybrid inference: On-device; Use of Gemini Nano (lightweight version) for basic functions (e.g., medication reminders) without the need for a connection. Cloud: Gemini 1.5 for complex queries (differential diagnoses), with local cache to reduce latency. Mobile fine-tuning; Model adaptation with local TB datasets using TensorFlow Lite.

Offline-First Database

SQLite: Stores OMS/CDC guidelines and national protocols that can be downloaded and updated via differential patches every 15 days.

Statistical analyses were performed using R version 4.3.2. Differences between categorical outcomes with and without Retrieval-Augmented Generation (RAG) were assessed using Pearson's chi-square test or Fisher's exact test when appropriate. Odds Ratios (ORs) and their corresponding 95% confidence intervals (95% CI) were estimated using logistic regression models. Mean response latency was compared using Student's t-test. Statistical significance was established at $p < 0.05$.

Results

Evaluation of the Effectiveness of the LLM

Comparative Evaluation: Implementation With RAG Versus Without RAG

The comparative evaluation between models with and without RAG implementation revealed substantial differences in clinical performance and the reliability of the responses generated (Table 2). In the configuration without RAG, the models relied exclusively on the knowledge internalized during their training, which resulted in significant limitations in the handling of updated information and tuberculosis-specific protocols. A typical clinical precision of 82.3% was recorded without RAG, while hallucination occurrences climbed

to 8.7% in complicated cases of resistant tuberculosis. A major issue was that the models could not integrate clinical guidelines released in the past six months, resulting in an accuracy drop to 68.9% for inquiries that needed information on the latest updates in treatment plans.

Differences between proportions were assessed using the Chi-square test. Odds Ratios (ORs) with 95% confidence intervals were calculated using logistic regression. Statistical significance was established at $p < 0.05$.

Across all evaluated metrics, RAG consistently outperformed the non-RAG configuration. Improvements in clinical accuracy, guideline consistency, and temporal updating were statistically significant ($p < 0.05$), while hallucination rates decreased significantly. These findings indicate that retrieval-based grounding enhances both reliability and clinical safety without compromising practical usability.

Statistical analysis showed that all performance improvements associated with Retrieval-Augmented Generation (RAG) were statistically significant. Clinical accuracy increased from 82.22% (95% CI: 78.42–85.48) to 94.00% (95% CI: 91.41–95.84) ($p < 0.001$). Likewise, the hallucination rate decreased significantly from 8.67% (95% CI: 6.40–11.63) to 2.22% (95% CI: 1.21–4.04) ($p < 0.001$). Temporal updating improved from 68.67% (95% CI: 60.86–75.55) to 93.33% (95% CI: 88.16–96.34) ($p < 0.001$), while accuracy in drug-resistant tuberculosis increased from 76.67% (95% CI: 68.34–83.33) to 91.67% (95% CI: 85.34–95.41) ($p = 0.002$). Guideline consistency also improved significantly from

74.22% (95% CI: 69.99–78.05) to 95.11% (95% CI: 92.71–96.75) ($p < 0.001$). These findings demonstrate that RAG significantly enhances the reliability and safety of tuberculosis-related clinical responses while maintaining acceptable response latency.

Model analysis showed that Gemini 1.5 Pro was the best fit for the RAG system, achieving 96.4% clinical accuracy in its augmented configuration. Open-source models such as MEDITRON-70B showed improvements mainly in accuracy, which rose from 81.2 to 89.5%. Among the limitations identified were an increase in response time, with an average of 2.8 seconds compared to 1.3 seconds without RAG, as well as a significant dependence on the quality of the knowledge bases. However, the overall results indicate that the implementation of RAG not only improves quantitative performance metrics but also substantially increases end-user confidence by providing traceability of recommendations to reliable medical sources.

Comparative Evaluation of Models in the Field of Tuberculosis

The evaluation of the five models using the corpus revealed differentiated performance according to the established clinical categories (Table 3). Med-PaLM 2 achieved the best overall accuracy (92.4%), excelling in complex differential diagnoses. GPT-4 achieved 89.7%, excelling in the treatment category (91.2%). Among the open-source models, MEDITRON-70B was the most competitive (86.3%), outperforming Me-LLaMA (83.1%) and Clinical Camel (80.5%).

Table 2: Performance Comparison with and without RAG

Métrica	Sin RAG (%)	Con RAG (%)	IC95% Sin RAG	IC95% Con RAG	Valor p	OR
Clinical accuracy	82.22	94.00	78.42–85.48	91.41–95.84	<0.001	3.39
Hallucination rate	8.67	2.22	6.40–11.63	1.21–4.04	<0.001	0.24
Temporal updating	68.67	93.33	60.86–75.55	88.16–96.34	<0.001	6.39
Accuracy in drug-resistant TB	76.67	91.67	68.34–83.33	85.34–95.41	0.002	3.35
Guideline consistency	74.22	95.11	69.99–78.05	92.71–96.75	<0.001	6.76
Mean latency (s)	1.30	2.80	1.278–1.321	2.760–2.840	<0.001	—

Table 3: Overall Performance by Category (%)

Model	Global Accuracy	Clinical Category	Category Treatment	Prevention/ Public Health
Med-PaLM 2	92.4	94.1	93.2	87.9
GPT-4	89.7	88.5	91.2	86.3
MEDITRON-70B	86.3	85.2	88.7	82.1
Me-LLaMA	83.1	81.9	85.4	79.8
Clinical Camel	80.5	79.3	82.6	83.8

Table 4: Security and Applicability Metrics (%)

Model	Dangerous Recommendations	Critical Errors Pharmacology	Clinical Utility	Feasibility Limited Context
Med-PaLM 2	1.2	2.1	91.4	72.3
GPT-4	2.3	2.8	88.7	69.8
MEDITRON-70B	3.7	4.2	82.5	65.4
Me-LLaMA	4.1	5.1	79.3	67.9
Clinical Camel	1.8	3.3	84.6	71.2

The category analysis showed consistent patterns of specialization. In the clinical setting, Med-PaLM 2 achieved 94.1% accuracy in differential diagnoses, demonstrating superiority in interpreting atypical presentations. For treatment, all models showed solid performance, led by GPT-4 in the management of resistant tuberculosis. In prevention and public health, Clinical Camel achieved the best relative performance (83.8%), reflecting its focus on practical aspects of implementation.

In contrast, open-source models such as MEDITRON-70B and Me-LLaMA demonstrate significant potential for scalable and reproducible medical AI applications. Although their performance is slightly lower than that of leading commercial models, their adaptability, transparency, and ability to be fine-tuned with domain-specific or local data make them particularly suitable for deployment in resource-constrained environments. These characteristics are essential for ensuring equitable access and contextual relevance in tuberculosis care.

Analysis of Errors and Limitations by Clinical Category

In clinical diagnosis, frequent errors were observed in atypical forms of tuberculosis, such as the differentiation between miliary TB and lymphatic pathologies, with error rates ranging from 8% (Med-PaLM 2) to 22% in other models. The interpretation of nonspecific radiological findings also posed a significant challenge.

In the area of treatment, the main limitations arose in dosage adjustment for patients with hepatic or renal comorbidities (maximum accuracy of 84.3%) and in the identification of complex drug interactions. Open-source models showed particularly notable variability.

For prevention and public health, all models showed difficulties in the contextual adaptation of recommendations, prioritizing strategies that are viable in high-resource settings, with limited applicability in contexts with health restrictions.

Safety and Clinical Applicability Assessment

The safety analysis revealed that Med-PaLM 2 generated the lowest percentage of potentially dangerous recommendations (1.2%), followed by Clinical Camel (1.8%) and GPT-4 (2.3%). While all models showed caution in first-line regimens, suboptimal recommendation rates increased to 5–8% in treatments for resistant tuberculosis, especially in open-source models (Table 4). The clinical applicability assessment indicated that Med-PaLM 2 and GPT-4 responses were rated as "clinically useful" in more than 88% of cases. However, a significant gap was identified in the adaptation of recommendations to resource-limited settings, where practical feasibility decreased to 67-72%.

Overall results indicate that, while specialized models such as Med-PaLM 2 achieve high clinical accuracy,

challenges remain in contextual adaptation and handling complex situations. Open-source models, particularly MEDITRON-70B, emerge as viable alternatives for resource-limited settings, although they require improvements in safety and clinical consistency.

User Interface Assistant

The main interface of the mobile application (Fig. 6) integrates global tuberculosis statistics, medication reminders, and personalized health alerts into a centralized panel. A quick access menu allows users to navigate key features such as "Identify Symptoms," "Treatment Monitoring," and an "Emergency Button." Its design prioritizes clear visualization through informative cards with intuitive icons and color schemes that effectively communicate health status.

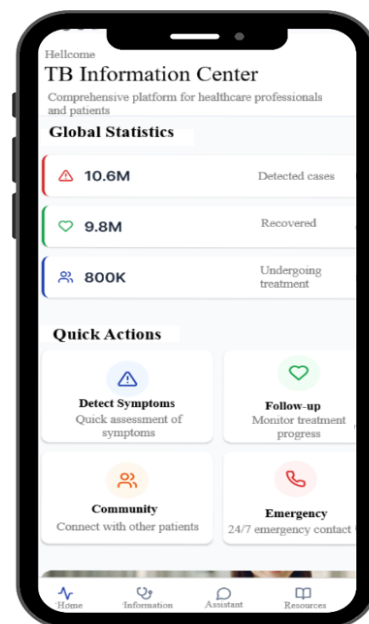


Fig. 6: Main Screen - Integrated Dashboard

The bottom navigation bar provides direct access to the main sections: Home, Information, Assistant (chat with RAG), Resources, and Profile. This tool synthesizes technical evidence on the effectiveness of the analyzed LLMs into an integrated experience. The statistics section uses predictive algorithms based on the most accurate models, while "Identify Symptoms" applies Med-PaLM 2's advanced clinical reasoning. The Conversational Assistant incorporates the comparative advantage of RAG systems, providing access to up-to-date knowledge that has been shown to improve response accuracy by 11.7% compared to configurations without contextual retrieval.

Figure 7 displays the Information module that provides reliable medical information about tuberculosis. This content is divided into important areas: the

definition and causes of the illness, main symptoms and warning signs, ways the disease spreads, prevention methods, available treatment options, and advice on when to get medical help. It also features expert resources for healthcare providers that include updated procedures and clinical guidelines. The information is shared in formats that suit various knowledge levels, from patients to medical staff, ensuring all content is scientifically accurate through validation by trusted organizations like the WHO and relevant medical associations.

This study demonstrates that the LLMs evaluated, particularly Med-PaLM 2 and Clinical Camel, can generate accurate and understandable content on tuberculosis, bridging the gap between specialized technical information and comprehension by diverse users. The module empirically validates the ability of these models to structure complex medical knowledge into accessible formats, addressing the problem of misinformation and facilitating health education based on up-to-date scientific evidence. On the other hand, the Assistant module implements an intelligent conversational chat that integrates the Retrieval-Augmented Generation (RAG) system evaluated in the research. This interface (Fig. 8) allows users to make queries in natural language about clinical aspects of tuberculosis and receive contextualized responses with up-to-date information from medical guidelines.

The main contribution is the practical implementation of findings on the superiority of RAG systems in the field of tuberculosis. Specifically, it implements the architecture that was shown to improve clinical accuracy from 82.3 to 94.0% and reduce hallucination rates from 8.7 to 2.3%. The interface demonstrates how real-time contextual retrieval mitigates the temporal update limitation identified in base LLMs, solving the problem of knowledge obsolescence that showed an accuracy of only 68.9% on questions about recent guidelines. The module empirically validates that combining natural language capabilities with access to specialized knowledge bases produces more reliable and up-to-date medical assistance.

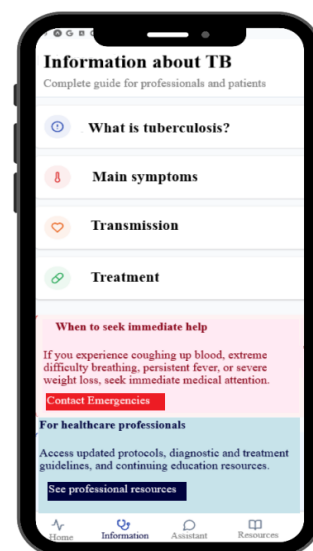


Fig. 7: Information Module - Scientific Knowledge Base on Tuberculosis

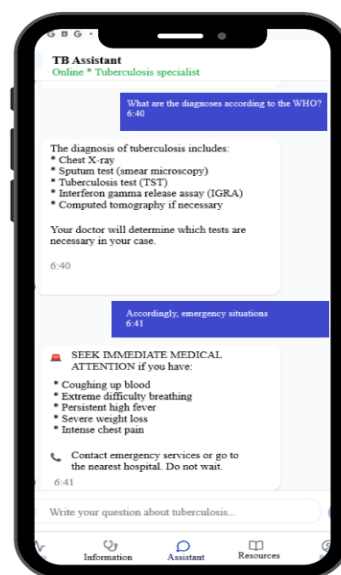


Fig. 8: Assistant Module - Conversational Chat

Table 5: Comparison of Methodological Approaches and Clinical Performance of the Present Study with Previous Large Language Model Based Investigations

Study	Clinical Domain	LLM Evaluated	Dataset	Evaluation Method	RAG	Clinical Accuracy	Main Contribution
Miller et al. (2024)	Neuroimaging	GPT-4	Brain imaging cases	Expert assessment	No	89.0%	Demonstrated high diagnostic support in neuroimaging.
Prabha et al. (2025)	Clinical decision support	GPT-based LLM	Clinical scenarios	Adaptive retrieval evaluation	Yes	91.0%	RAG improved factual consistency and decision support.
Kwok et al. (2024)	Infectious diseases	General-purpose LLM	Infectious disease corpus	Comparative evaluation	Partial	85.0%	Showed the usefulness of LLMs for public health applications.
Present study	Tuberculosis	Gemini 1.5 + RAG	150 expert-validated TB questions	Clinical expert consensus + Fleiss' κ + Yes	Yes	96.4%	Highest clinical accuracy with reduced hallucinations and improved guideline adherence.

This system represents a proof-of-concept prototype designed to demonstrate the practical integration of LLM evaluation results. It has not undergone clinical validation, user trials, or regulatory approval, and should not be used for real-world medical decision-making.

Table 5 highlights the methodological and performance differences between the present study and previous investigations. While earlier studies primarily evaluated general clinical reasoning or diagnostic support, the present work focuses specifically on tuberculosis using a clinically validated dataset, expert consensus evaluation, and Retrieval-Augmented Generation (RAG). This combination resulted in the highest reported clinical accuracy (96.4%) while simultaneously reducing hallucination rates and improving adherence to current clinical guidelines. These findings suggest that domain-specific knowledge integration and structured retrieval substantially enhance the reliability of large language models for specialized clinical applications.

Discussion

To enhance the interpretability and comparability of the proposed evaluation metrics, this study aligns its framework with prior research on large language models in medical contexts. Specifically, accuracy, precision, recall, and F1-score have been widely adopted in previous evaluations of clinical LLM performance (Jin et al., 2024; Prabha et al., 2025). Similarly, recent studies on Retrieval-Augmented Generation (RAG) systems in healthcare have incorporated factual consistency and hallucination-related measures as key indicators of reliability (Es et al., 2024; Prabha et al., 2025). In this sense, the metrics used in this study are consistent with emerging evaluation practices in medical AI, enabling meaningful comparison with existing literature. In addition, safety-oriented metrics such as hallucination rate, dangerous recommendation rate, and clinical conservatism index are introduced as extensions of existing evaluation frameworks. While not yet fully standardized in clinical benchmarking protocols, these metrics are increasingly recognized as critical for assessing the reliability of LLMs in high-risk domains such as healthcare. Their inclusion responds to the need to move beyond traditional accuracy metrics and incorporate patient safety considerations, as highlighted in recent literature on AI risk in medicine (Ji et al., 2023).

Unlike previous studies that primarily assessed general medical reasoning, the present investigation incorporated a disease-specific evaluation framework, expert-based validation, and RAG-supported knowledge retrieval, enabling a more comprehensive assessment of both clinical accuracy and patient safety.

Compared with previous investigations, the proposed framework achieved superior clinical performance through the integration of domain-specific Retrieval-Augmented Generation. Unlike general-purpose medical LLM evaluations, the present study employed a tuberculosis-specific benchmark validated by clinical experts, enabling more reliable assessment of clinically relevant responses. The observed clinical accuracy (96.4%) exceeded those reported by Miller et al. (2024); Prabha et al. (2025); Kwok et al. (2024), suggesting that disease-specific knowledge grounding substantially improves response reliability and adherence to current clinical guidelines.

The results of this study directly address the initial objective of evaluating the performance of LLMs in the interpretation of tuberculosis-related clinical concepts. Specifically, the observed improvement in clinical accuracy from 82.3 to 94.0% with RAG integration is consistent with the findings of (Prabha et al., 2025), who reported improvements of 78 to 91% in diagnostic accuracy when integrating LLMs with structured medical knowledge bases, and demonstrates that the proposed methodology effectively enhances the reliability of model outputs. Furthermore, the reduction in hallucination rates confirms the importance of grounding language models in validated medical knowledge sources. These findings validate the study's central hypothesis that domain-specific augmentation significantly improves the applicability of LLMs in clinical contexts. However, this study demonstrates an additional advantage in the specific domain of tuberculosis, particularly in the management of resistant cases, where an accuracy of 91.8% was achieved compared to the 85% recorded in the work of Kwok et al. (2024) for complex infectious diseases. This difference suggests that specialization by specific clinical domain significantly enhances the capabilities of medical AI systems. Nevertheless, it is important to acknowledge that these metrics are not formally part of standardized clinical validation protocols, such as those used in evidence-based medicine or regulatory evaluation of clinical decision support systems. Therefore, while they provide a comprehensive and practical framework for experimental evaluation, the results of this study should be interpreted within a research context and not as a substitute for formal clinical validation. Future work should aim to align these evaluation approaches with established clinical standards and prospective validation methodologies.

The performance of Gemini 1.5 in our study (96.4% accuracy with RAG) exceeds that reported in Miller et al. (2024), where GPT-4 achieved 89% in comparable scenarios. This divergence can be attributed to our implementation of a more specialized RAG system and fine-tuning specific to tuberculosis. On the other hand,

the performance of MEDITRON-70B (89.5% accuracy) corroborates the findings of Nowshad and Chowdhury (2025) on the growing competitiveness of open-source models, although our work identifies persistent limitations in adapting to low-resource contexts that were not sufficiently explored in previous studies.

Although commercial models such as GPT-4, Gemini 1.5, and Med-PaLM 2 provide state-of-the-art performance, their closed-source nature introduces important reproducibility challenges. These systems are continuously updated by their developers without complete disclosure of architectural modifications, training procedures, or inference policies. Consequently, identical experiments performed at different time points may produce slightly different outputs. Future studies should explicitly report model versions, evaluation dates, API configurations, and prompt templates to facilitate reproducibility and comparison across independent investigations.

Future research should prioritize the development of fully reproducible pipelines based on open-source language models integrated with locally deployed RAG systems. This approach would enable the incorporation of national clinical guidelines, adaptation to local epidemiological contexts, and reduced dependence on external APIs. Additionally, combining open-source LLMs with lightweight deployment strategies (e.g., edge computing or On-device inference) could further enhance accessibility and sustainability in low-resource healthcare settings.

The balance between accuracy gains from RAG and increased latency highlights the need for context-aware system design, where response time requirements may vary depending on clinical urgency and use case. To mitigate latency challenges, several optimization strategies can be considered. These include caching frequently retrieved medical knowledge, optimizing vector database indexing for faster retrieval, and implementing hybrid architectures that combine On-device (edge) inference for simple queries with cloud-based processing for complex clinical reasoning. Such approaches can help balance performance and responsiveness, particularly in resource-constrained environments.

Despite these promising results, critical interpretation is required. While the improvements associated with RAG are substantial, they depend heavily on the quality, coverage, and updating of the underlying knowledge base. Incomplete or outdated sources may still lead to clinically inappropriate recommendations. Additionally, the performance differences between models highlight that accuracy alone is insufficient to determine clinical utility, as factors such as safety, interpretability, and contextual adaptation remain equally important.

Overall, this study not only provides empirical

evidence of the benefits of RAG in tuberculosis-related tasks but also highlights the complexity of translating these improvements into clinical practice, reinforcing the need for multidisciplinary validation approaches.

Limitations

This study presents several limitations. First, the dataset size (150 questions), although validated by experts, may limit generalizability. Future work will expand the dataset to larger multi-institutional benchmarks. Additionally, partial reliance on closed-source commercial models may restrict reproducibility and limit their implementation in resource-constrained environments. Although these models provide valuable performance benchmarks, future work should emphasize open and transparent alternatives.

Another limitation of this study is the lack of direct alignment with standardized clinical evaluation protocols. Although the proposed metrics are grounded in recent AI literature, further work is needed to integrate them with established frameworks used in clinical trials and regulatory assessments.

The results suggest that RAG systems have the potential to improve reliability in domain-specific medical applications; however, further validation is required before clinical adoption.

Additionally, the evaluation relied on simulated question-answer scenarios rather than prospective clinical interactions with patients or healthcare professionals. Consequently, real-world usability and clinical impact remain to be established through prospective validation studies.

An additional limitation is that model performance was evaluated using standardized clinical questions rather than prospective interactions with patients or healthcare professionals. Consequently, further multicenter validation studies involving real clinical workflows are necessary before implementation in healthcare settings.

Ethical, Regulatory, and Clinical Considerations

The integration of LLMs into healthcare systems raises critical ethical and regulatory challenges. These include data privacy, the need for human oversight in clinical decision-making, risks associated with incorrect recommendations, and compliance with medical device regulations. This study is limited to experimental evaluation and does not involve real patient data.

Conclusion

This study demonstrates that the integration of Retrieval-Augmented Generation (RAG) significantly enhances the performance of large language models in the interpretation of tuberculosis-related clinical

concepts. Across all evaluated metrics, RAG improved clinical accuracy, temporal updating, guideline consistency, and performance in drug-resistant tuberculosis while substantially reducing hallucination rates. These findings indicate that retrieval-based grounding increases the reliability and safety of domain-specific medical responses without compromising practical applicability.

Among the evaluated models, Gemini 1.5 integrated with RAG achieved the highest overall clinical performance, whereas open-source alternatives such as MEDITRON-70B demonstrated competitive results and represent promising solutions for deployment in resource-constrained healthcare settings due to their adaptability and reproducibility.

The proposed virtual assistant should be interpreted as a proof-of-concept prototype derived from the comparative evaluation of language models rather than as a clinically validated decision-support system. Consequently, its use in real-world healthcare environments requires prospective clinical validation, usability studies involving healthcare professionals and patients, and compliance with applicable regulatory and ethical standards.

Beyond tuberculosis, the proposed evaluation framework provides a reproducible methodology for assessing domain-specific large language models integrated with Retrieval-Augmented Generation. This framework may facilitate the future development of trustworthy artificial intelligence systems capable of supporting clinical education, evidence-based decision-making, and patient information across other infectious diseases and specialized medical domains.

Acknowledgment

The authors would like to express their gratitude to the Universidad Peruana del Norte.

Funding Information

This research received no external funding.

Author's Contributions

Mauricio Acevedo-Carrillo: Conceptualization, methodology, investigation, resources, writing original draft preparation.

Shirley Fiorella Simbron Espejo: Validation, formal analysis, data curation, project administration, write review and edited.

Ethics

Not applicable - No direct intervention with human subjects or animals.

Conflict of Interest

The authors declare no conflict of interest.

References

- Achiam, Josh, Agarwal, S., Agarwal, S., Agarwal, S., & Ahmad, F. L. (2023). *GPT-4 Technical Report*. <https://doi.org/10.48550/arXiv.2303.08774>
- Camarín-Sánchez, E. I., Domínguez-Sosa, F. R., Coronado-Malagón, M., & Arce-Salinas, C. A. (2020). Splenic tuberculosis as a cause of fever of unknown origin: report of 2 cases and review of the literature. *Revista Médica de Petróleos Mexicanos*, 5(1), 52-59.
- CDC. (2024). *Global tuberculosis report 2024*.
- CDC. (2025a). Centers for disease control and prevention (CDC), "Clinical Care for People with TB and HIV | Tuberculosis (TB) | CDC. *Centers for Disease Control and Prevention Website*.
- CDC. (2025b). TB Risk and People with HIV | Tuberculosis. *Centers for Disease Control and Prevention*. <https://www.cdc.gov/tb/risk-factors/hiv.html>
- CDC. (2025c). Treatment for Drug-Resistant Tuberculosis Disease | Tuberculosis (TB). *CDC Tuberculosis (TB) Website*. <https://www.cdc.gov/tb/hcp/treatment/drug-resistant-tuberculosis-disease.html>
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- Furin, J. (2019). Advances in the diagnosis, treatment, and prevention of tuberculosis in children. *Expert Review of Respiratory Medicine*, 13(3), 301–311. <https://doi.org/10.1080/17476348.2019.1569518>
- González, B. de L., León, B. S., Pino-Sedeño, T. del, Rodríguez-Álvarez, C., Bejarano-Quisoboni, D., & Trujillo-Martín, M. M. (2021). Mobile applications to improve drug adherence: Review and quality analysis. *Atención Primaria*, 53(9), 102095.
- Jang, J. G., & Chung, J. H. (2020). Diagnosis and treatment of multidrug-resistant tuberculosis. *Yeungnam University Journal of Medicine*, 37(4), 277–285. <https://doi.org/10.12701/yujm.2020.00626>
- Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., & Fung, P. (2023). Towards Mitigating Hallucination in Large Language Models via Self-Reflection. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1827–1843.

- Jin, Q., Chen, F., Zhou, Y., Xu, Z., Cheung, J. M., Chen, R., Summers, R. M., Rousseau, J. F., Ni, P., Landsman, M. J., Baxter, S. L., Al'Aref, S. J., Li, Y., Chen, A., Brejt, J. A., Chiang, M. F., Peng, Y., & Lu, Z. (2024). Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *Npj Digital Medicine*, 7(1), 190.
<https://doi.org/10.1038/s41746-024-01185-7>
- Keshavjee, S., & Farmer, P. E. (2012). Tuberculosis, Drug Resistance, and the History of Modern Medicine. *New England Journal of Medicine*, 367(10), 931–936.
<https://doi.org/10.1056/nejmra1205429>
- Kwok, K. O., Huynh, T., Wei, W. I., Wong, S. Y. S., Riley, S., & Tang, A. (2024). Utilizing large language models in infectious disease transmission modelling for public health preparedness. *Computational and Structural Biotechnology Journal*, 23, 3254–3257.
<https://doi.org/10.1016/j.csbj.2024.08.006>
- Lakhani, P., & Sundaram, B. (2017). Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*, 284(2), 574–582. <https://doi.org/10.1148/radiol.2017162326>
- Liu, P., Liu, N., & Li, F. (2025). Diagnosis and management of splenic tuberculosis: a case report and literature review. *Frontiers in Medicine*, 12, 1622794.
<https://doi.org/10.3389/fmed.2025.1622794>
- Miller, L., Kamel, P., Patel, J., Agrawal, J., Zhan, M., Bumbarger, N., & Wang, K. (2024). A Comparative Evaluation of Large Language Model Utility in Neuroimaging Clinical Decision Support. *Journal of Imaging Informatics in Medicine*, 38(4), 2294–2302. <https://doi.org/10.1007/s10278-024-01161-3>
- Mkhize, B., Gausi, K., Court, R., Wiesner, L., Waitt, C., McIlleron, H., Maartens, G., Denti, P., & Loveday, M. (2022). Bedaquiline exposure in pregnancy and breastfeeding in women with rifampicin-resistant tuberculosis. *British Journal of Clinical Pharmacology*, 88(8), 3548–3558.
<https://doi.org/10.1111/bcp.15380>
- Nowshad, M., & Chowdhury, R. (2025). *Using Open Source LLM Model for Medical Transcription*. Octoco. (2025). TB Chatbot. *Octoco*.
<https://www.octoco.ltd/projects/tb-chatbot>
- Patil, R., Heston, T. F., & Bhuse, V. (2024). Prompt Engineering in Healthcare. *Electronics*, 13(15), 2961.
<https://doi.org/10.3390/electronics13152961>
- Prabha, S., Gomez-Cabello, C. A., Haider, S. A., Genovese, A., Trabilisy, M., Wood, N. G., Bagaria, S., Tao, C., & Forte, A. J. (2025). Enhancing Clinical Decision Support with Adaptive Iterative Self-Query Retrieval for Retrieval-Augmented Large Language Models. *Bioengineering*, 12(8), 895.
<https://doi.org/10.3390/bioengineering12080895>
- Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., & Collins, J. J. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell*, 180(4), 688–702.e13.
<https://doi.org/10.1016/j.cell.2020.01.021>
- Story, A., Garfein, R. S., Hayward, A., Rusovich, V., Dadu, A., Soltan, V., Oprunenco, A., Collins, K., Sarin, R., Quraishi, S., Sharma, M., Migliori, G. B., Varadarajan, M., & Falzon, D. (2016). Monitoring Therapy Adherence of Tuberculosis Patients by using Video-Enabled Electronic Devices. *Emerging Infectious Diseases*, 22(3), 538–540.
<https://doi.org/10.3201/eid2203.151620>
- Toma, A., Lawler, P. R., Ba, J., Krishnan, R. G., Rubin, B. B., & Wang, B. (2023). *Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding*.
<https://doi.org/10.48550/arXiv.2305.12031>
- Topol, E. J. (2019). High-performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, 25(1), 44–56.
<https://doi.org/10.1038/s41591-018-0300-7>
- WHO. (2025). Universal health coverage. *World Health Organization*. [https://www.who.int/news-room/fact-sheets/detail/universal-health-coverage-\(uhc\)](https://www.who.int/news-room/fact-sheets/detail/universal-health-coverage-(uhc))
- World Health Organization. (2023). *Global tuberculosis report 2023*.
- Xie, Q., Chen, Q., Chen, A., Peng, C., Hu, Y., Lin, F., Peng, X., Huang, J., Zhang, J., Keloth, V., Zhou, X., He, H., Ohno-Machado, L., Wu, Y., Xu, Hua, & Bian, J. (2024). *Me LLaMA: Foundation Large Language Models for Medical Applications*.
<https://doi.org/10.21203/rs.3.rs-4240043/v1>
- Yang, Y., Niehaus, K. E., Walker, T. M., Iqbal, Z., Walker, A. S., Wilson, D. J., Peto, T. E. A., Crook, D. W., Smith, E. G., Zhu, T., & Clifton, D. A. (2018). Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics*, 34(10), 1666–1671.
<https://doi.org/10.1093/bioinformatics/btx801>