Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 270 (2025) 4946–4956

Procedia Computer Science

www.elsevier.com/locate/procedia

29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025)

# Enhancing Clinical Decision Support with LLMs: A Feasibility Study Integrating CoT, RAG, and QLoRA

George Matthew[a], Yulia Hicks[a]

[a] *School of Engineering, Cardiff University, Cardiff CF24 3AA, UK*

**Abstract**

The integration of artificial intelligence (AI) into healthcare holds significant promise for enhancing clinical decision-making and improving patient outcomes. However, general-purpose large language models (LLMs) frequently exhibit limitations such as hallucinations, lack of domain-specific accuracy, and opaque reasoning processes, posing risks in clinical applications. This study addressed these challenges by proposing and exploring an innovative integration of Chain-of-Thought (CoT) prompting, Retrieval-Augmented Generation (RAG), and parameter-efficient fine-tuning using Quantized Low-Rank Adaptation (QLoRA) specifically tailored for medical use.

A distilled 14-billion-parameter variant of the DeepSeek R1 model was fine-tuned using a structured clinical dataset that emphasizes step-by-step reasoning. Additionally, external medical references were incorporated through RAG, employing embedding models for precise context retrieval. The combination of these techniques was systematically evaluated on a challenging set of open-ended medical questions, resulting in accuracy improvements—from a baseline accuracy of 55% to a final performance of 81%.

Further qualitative evaluation involving three practising General Practitioners (GPs) and three fourth-year medical students from Cardiff University underscored the proposed system's clinical utility and transparent reasoning capabilities while also identifying areas for improvement, such as conciseness and explicit adherence to national-specific clinical guidelines.

This research demonstrated that integrating CoT, RAG, and QLoRA provides a practical pathway toward reliable, transparent, and clinically relevant AI support for healthcare professionals. Recommendations for future work include scaling models, incorporating comprehensive patient data, and enhancing customization for clinical application contexts.

*Keywords:* Medical LLMs; Clinical Decision Support Systems; CoT; RAG; QLoRA; Parameter-Efficient Fine-Tuning; Explainable AI

## 1. Introduction

The integration of AI into healthcare has demonstrated substantial potential for enhancing clinical workflows and supporting medical decision-making. LLMs, in particular, excel at natural language understanding and generation; yet, their direct application in clinical settings faces critical challenges, especially around accuracy and trustworthiness. Specifically, current LLMs can generate plausible but incorrect information (hallucination [1]), which can pose serious risks to patient safety [2]. In healthcare contexts, precision is paramount; thus, any AI system must deliver

10.1016/j.procs.2025.09.621

correct information with transparent justification of its reasoning. Equally vital is the placement of AI within a human-in-the-loop framework that keeps ultimate decision authority in the hands of healthcare professionals [3].

A key limitation is the lack of transparency in LLMs, often referred to as their “black-box” nature. When diagnoses and treatment recommendations cannot be easily traced back to clear logical steps, clinicians may be reluctant to rely on AI outputs. Explainable AI approaches address this challenge by prioritizing interpretability and transparency of the underlying reasoning [4]. These elements are critical for building the trust needed for AI-driven systems to move from research prototypes into mainstream clinical practice.

To tackle these concerns, this paper explores an integrated method that combines CoT prompting with RAG and efficient fine-tuning via QLoRA. The goal is to create an AI assistant specifically tailored to medical workflows, offering transparent reasoning supported by domain-specific knowledge and efficient adaptation to healthcare tasks.

The following sections discuss how CoT prompting, RAG, and QLoRA were integrated to enhance factual accuracy, interpretability, and domain fidelity for clinical decision-making. The article then describes dataset usage and the model selection process. Finally, it presents quantitative findings alongside feedback from practicing GPs and medical students, illustrating the feasibility of safe, transparent, and effective AI support in healthcare settings.

## 2. Literature Review

*LLMs in Healthcare: Capabilities and Limitations.* LLMs have demonstrated substantial potential within healthcare, approaching the performance of medical professionals in both clinical assessments and standardized licensing examinations [3]. Recent studies underline this capability; for example, a Generative Pre-trained Transformer based model (GPT) achieved an impressive 79% on the UK Royal College of Radiologists examination, closely approaching the human expert benchmark of 84.8% [3]. Such results highlight the capacity of LLMs to synthesize complex medical information and provide reliable diagnostic suggestions. This ability positions them as valuable tools for clinical decision support, potentially enhancing diagnostic accuracy, efficiency, and patient outcomes in healthcare settings.

Despite these achievements, significant concerns remain. The primary challenge for the clinical use of LLMs is their tendency to produce confidently stated yet incorrect or fabricated outputs, known as hallucinations [5]. Such inaccuracies pose risks to patient safety, necessitating mechanisms to verify and contextualise AI-generated information. Another limitation is the lack of transparency in the decision-making process of LLMs, rendering their reasoning difficult to follow or verify by clinicians, thereby reducing trust and potential adoption in clinical settings.

*Enhancing Reasoning Transparency with CoT Prompting.* Recent advancements, such as CoT prompting, offer promising approaches to mitigate transparency issues in LLM-generated content. CoT encourages LLMs to articulate intermediate reasoning steps, thus enhancing logical coherence and providing clinicians with explicit insights into the AI’s decision-making process [6]. Preliminary research indicates significant improvements in both interpretability and accuracy when applying CoT to clinical reasoning tasks, suggesting a viable strategy for aligning AI-generated reasoning more closely with human clinical thought processes [6].

*Factual Accuracy Through Retrieval-Augmented Generation.* To address the challenge of hallucination and ensure factual accuracy, RAG has emerged as a robust solution. RAG incorporates external verified knowledge sources during generation, grounding AI responses in evidence-based medical literature [5]. This methodology significantly reduces the likelihood of erroneous outputs by providing real-time context, directly improving the reliability and clinical acceptability of generative AI systems [5].

*Domain-Specific Adaptation via Efficient Fine-Tuning.* Efficient fine-tuning approaches, such as LoRA, enable effective adaptation of general-purpose LLMs to specialized medical domains without prohibitive computational resources [7]. By training only a limited number of parameters, LoRA efficiently aligns models to domain-specific vocabulary and practices, ensuring relevance and improved performance in targeted healthcare applications. This method addresses the computational barriers traditionally associated with fine-tuning large-scale models [7].

*Enhancing Diagnostic Safety with AI: Lessons from Martha’s Rule.* Martha’s Rule is a recently advocated policy in the UK introduced following the death of 13-year-old Martha Mills, whose deterioration was not recognised in time to prevent her death from sepsis. This rule aims to prevent such diagnostic errors by empowering patients and their families with the right to request an urgent second medical opinion if they believe their concerns have been overlooked [8]. Although the rule itself does not directly incorporate AI, this context underscores a critical real-world gap that medical AI could address by providing instantaneous, reliable second opinions and mitigating human diagnostic oversights.

*Identified Research Gap.* Despite advancements in AI, specifically LLMs, few current frameworks within a clinical setting provide domain-specific factual accuracy and transparency in a computationally efficient solution. This research directly addresses this gap by combining CoT prompting, RAG for evidence-grounded responses, and parameter-efficient LoRA fine-tuning. By systematically evaluating this integrated approach through both quantitative benchmarks and qualitative expert feedback, this study aims to contribute towards safe, practical, and transparent deployment of AI-assisted decision support systems in healthcare.

## 3. Datasets

*Fine-tuning Dataset.* Enhancing performance through CoT and fine-tuning required a dataset specifically tailored to encapsulate clinical CoT problem-solving. For this purpose, the dataset titled medical-o1-reasoning-SFT [9] was used, enabling the model to learn step-by-step diagnostic and treatment logic aligned with CoT objectives. A medical verifier confirmed its validity, and previous work [9] has demonstrated its effectiveness in enhancing reasoning within the medical domain.

The dataset was constructed by integrating two large-scale Multiple-Choice Question Answering (MCQA) corpora—the United States Medical Licensing Examination (USMLE) [10] and MedQA [11]—which together supplied a diverse range of authentic clinical scenarios essential for refining the model's medical expertise. Each record in the dataset is organized into three interrelated components: a clinical prompt labeled "Question" that sets the scenario, a detailed multi-step CoT labeled "Complex_CoT" capturing the intermediate reasoning process, and a succinct final conclusion labeled "Response." Comprising 25.5 thousand entries [9], this structured format is critical for effectively leveraging the distilled DeepSeek model [12] and fostering robust CoT reasoning.

*RAG Vector Database Dataset.* The second dataset employed is the Textbooks Corpus from Medical RAG (MedRAG/-textbooks [13]), which provides valuable context for retrieval-augmented generation. This corpus consists of content from 18 widely used medical textbooks based on USMLE standards—key references for clinical practice. Given that the fine-tuning dataset was also derived from USMLE material, the two datasets complement each other effectively.

Each record in MedRAG/textbooks represents a short text chunk—restricted to no more than 1000 characters—to enable precise retrieval of pertinent passages. Preprocessing was performed using the RecursiveCharacterTextSplitter from LangChain, yielding 125,847 entries that average approximately 182 tokens each [13]. This extensive coverage increases the likelihood that the embeddings-based language model will locate the relevant contextual information during retrieval.

Structurally, every record is organized into three columns: an identifier (ID), a title indicating the source textbook, and the content of the text snippet. This streamlined format is optimized for real-time lookups, ensuring that only the most contextually appropriate segments are returned. By integrating these concise, thematically coherent chunks into the RAG pipeline, the model is provided with on-demand access to authoritative medical references, a critical factor in grounding its chain-of-thought reasoning with factual accuracy.

## 4. Proposed System Overview

Figure 1 below presents a high-level overview of the proposed system developed. It comprises two main frameworks: Fine-Tuning and Interface. In the Fine-Tuning framework, QLoRA (4-bit quantization + LoRA) is applied to the CoT-style fine-tuning dataset, adapting the LLM for domain-specific tasks and encouraging explicit reasoning. The resulting LoRA adapters are then integrated into the same LLM within the Interface framework. This framework orchestrates RAG by retrieving relevant medical data (via an embedding model and vector store) and appending it to user queries, while CoT prompting leverages these contexts to ground answers and generate step-by-step reasoning.

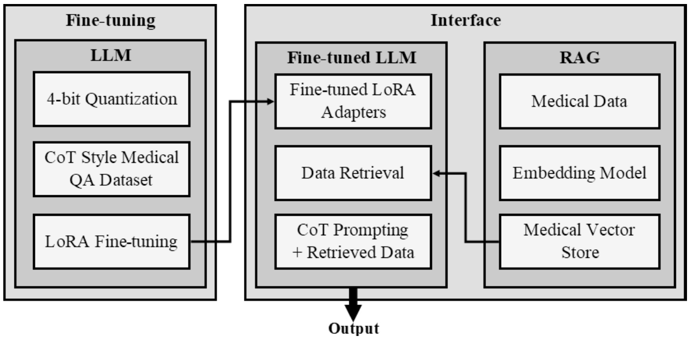

Fig. 1. Proposed System Overview

*QLoRA details.* QLoRA's 4-bit quantisation was used to lower memory use while keeping accuracy close to full-precision LoRA [14]. LoRA adapters were added to all attention projections (query, key, value, output) and the feed-forward gating and up-projection layers, the points that give the greatest adaptation leverage in the transformer architecture [15]. Standard hyper-parameters were chosen—rank 32, scaling $\alpha$ 64, dropout 10%—were adopted for a good balance between capacity and over-fitting. Training and validation losses were logged to monitor training stability and over-fitting.

*RAG details.* Cosine similarity was adopted to compute similarity scores between the user query and stored context vectors, as this metric has demonstrated higher performance than other options [16]; the top k=5 most similar contexts are then retrieved for each query.

## 5. Tested Model Architectures

### 5.1. Available Resources

Efficient deployment of LLMs demands considerable computational capacity [7]. In this study, Google Colab Pro, a cloud-based platform, provided access to high-performance NVIDIA A100 GPUs with 40GB of VRAM, facilitating the fine-tuning process. This 40GB VRAM limit presented a significant constraint, the main factor in using QLoRA. Consequently, selecting an efficient and well-optimized model was essential to balancing performance and computational feasibility.

### 5.2. Text Generation LLM

The work began with using a well-known model from Meta called LLaMA 3.1 8B Instruct [17], which contains 8.03 billion parameters. This model was selected for its efficiency and robust instruction-following performance against its competitors at the time. However, with the release of DeepSeek R1 in January 2025, the study transitioned to a distilled 14.8 billion parameter version of DeepSeek R1 (DeepSeek-R1-Qwen-14B [12]) which enabled the exploration of CoT techniques. Against the graduate-level Google-Proof Question and Answer (GPQA) benchmark [18], a challenging dataset of 448 multiple-choice questions in biology, physics, and chemistry, the distilled DeepSeek 14B scored 59.1% [12], compared to 34.6% for LLaMA 8B Instruct [17]. Remarkably, even LLaMA's flagship 405 billion parameter model scored only 50.7% [19], underscoring the efficiency and potential of the distilled DeepSeek model, which has a parameter size approximately 28 times smaller. To put this in perspective, experts who have or are pursuing PhDs in the corresponding domains achieve 65% accuracy [18].

### 5.3. RAG Embedding LLM

RAG relies on models that transform the text into numerical vectors, facilitating the semantic search for relevant context. This section examines the two embedding models used in the study, E5-Large-v2 [20] and gte-Qwen2-1.5B-instruct [21]. Both models produce embeddings but differ in parameter count, dimensionality, and reported performance on specialized medical benchmarks discussed later in this subsection.

*E5-Large-v2.* E5-Large-v2 comprises 335 million parameters and generates 1024-dimensional embeddings [20]. Its popularity in open-domain retrieval made it a natural starting point, as it balances reasonably strong encoding with modest resource requirements. However, as discussed in Section 6.1.2, practical integration challenges arose when incorporating E5-Large-v2 into the system infrastructure as a result of lower-than-expected retrieval quality for domain-specific content, prompting the investigation of a more capable alternative.

*gte-Qwen2-1.5B-instruct.* gte-Qwen2-1.5B-instruct contains 1.78 billion parameters and yields 1536-dimensional embeddings [21]. Its additional dimensionality and parameter number allow the model to capture finer-grained semantic patterns, though it also increases VRAM and storage demands for the vectors themselves. In practice, this trade-off was worthwhile as it provided more precise retrieval of specialised medical texts. This is detailed in Section 6.1.2.

*Performance on Medical Retrieval Tasks.* Both E5-Large-v2 and gte-Qwen2-1.5B-instruct were evaluated using the Massive Multilingual Text Embedding Benchmark (MMTEB) [22], which was filtered for the medical field. This benchmark measures a model's capacity to identify and retrieve relevant documents in a large medical database.

MMTEB ranks models according to Discounted Cumulative Gain at rank 10 (nDCG@10). This metric compares the top 10 retrieved documents against a perfect "ideal" ranking in which all relevant results appear first, with scores ranging from 0 (worst) to 1 (best). At the time of writing, in comparing models, E5-Large-v2 achieved 105th place with a nDCG@10 of 56.53% [22]. By comparison, gte-Qwen2-1.5B-instruct reached 3rd place with a score of 77.55% [22]. The models that achieved 1st and 2nd provided marginal improvements but were significantly larger with 7 billion parameters and higher dimension embeddings [22]. These larger models were impractical as they would have exceeded VRAM and storage capacity.

## 6. Evaluation

### 6.1. Quantitative Evaluation

In order to evaluate different model configurations, a suitable benchmark was required. Initially, past paper multiple-choice USMLE questions were used, as these reflected the type of content on which the model was trained and stored in the RAG vector store. These questions were highly relevant to the application of an AI medical assistant. An initial test was conducted using the fine-tuned, distilled Deepseek model with CoT and RAG on a USMLE Step 3 sample test paper [10]. This test paper contained four blocks: two labelled Foundations of Independent Practice (FIP) and two labelled Advanced Clinical Medicine (ACM). Five questions were selected from each block, and the answers were manually verified. However, the system achieved a perfect score (100%), raising concerns about the benchmark's ability to provide a robust performance assessment. This was attributed to the multiple-choice format allowing the model to evaluate each option and determine the most likely correct answer—a process evident in its output.

To address this, a search was conducted for challenging and relevant open-ended questions. The FreedomIntelligence/medical-o1-verifiable-problem dataset from HuggingFace [9], comprising USMLE-style open-ended questions paired with detailed answers, was identified as particularly suitable, comprising challenging USMLE-style open-ended questions with answers, was identified as suitable. These questions better reflected the complexity required for a robust evaluation. In order to test a sufficient number of questions across multiple model configurations, an automated evaluation framework was designed, as illustrated in Figure 2 below.

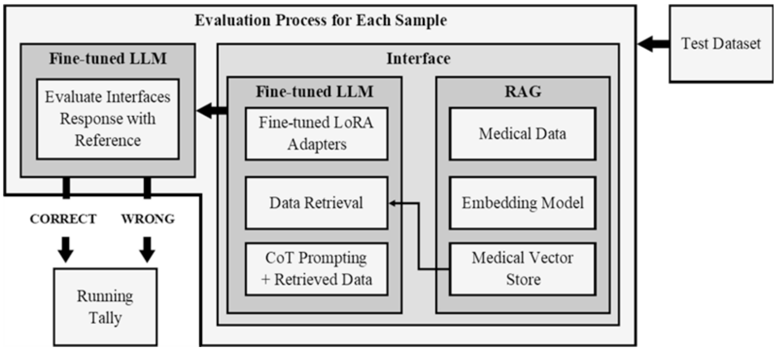

Fig. 2. Automated Evaluation Framework

*Evaluation Pipeline Overview.* Figure 2 presents an automated evaluation pipeline, adapted from the interface framework outlined in Figure 1, designed explicitly for reproducible performance testing. Unlike the original framework intended for real-time user interactions, this pipeline systematically processes a predefined "Test Dataset" (FreedomIntelligence/medical-o1-verifiable-problem) to quantify the model's accuracy.

Within this pipeline, the fine-tuned model first retrieves contextual information using Retrieval-Augmented RAG, then applies CoT reasoning to generate a final answer, consistent with the approach depicted in Figure 1. Subsequently, the framework leverages the same LLM to self-assess by comparing the generated answer against the ground-truth reference, outputting a clear verdict of CORRECT or WRONG.

Each verdict incrementally updates a cumulative tally of results, allowing the calculation of an overall accuracy score upon completion of the testing. This structured approach facilitates transparency, scalability, and reproducibility, essential for rigorous evaluation of model performance in handling complex medical queries.

### 6.1.1. Quantitative Evaluation Results and Discussion

Figure 3 presents a comparative assessment of five distinct model configurations evaluated via the evaluation loop. Each configuration reflected a cumulative integration of methodological enhancements discussed in this article, including CoT prompting and fine-tuning via QLoRA and RAG.

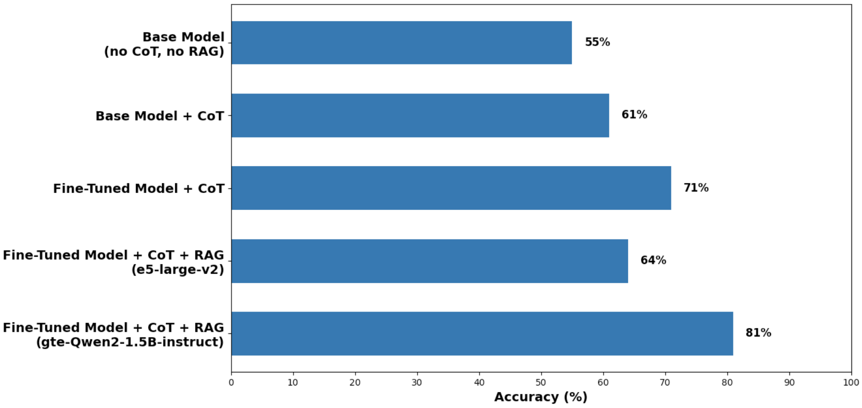

Fig. 3. Model Configurations Performance

### 6.1.2. Quantitative Evaluation Discussion

*Key Observations.* Across all five configurations, accuracy steadily improved from a 55% baseline (distilled DeepSeek-R1-Qwen-14B prompted without CoT) to 61% with CoT prompting, then to 71% after QLoRA fine-tuning. Introducing RAG yielded mixed results: using e5-large-v2 embeddings degraded performance to 64%—a consequence of noisy or tangential retrieval—whereas switching to the larger gte-Qwen2-1.5B-instruct embeddings model boosted accuracy to 81%, the highest observed.

The results attained suggested guiding the model's reasoning processes through CoT and tailoring its parameters via QLoRA fine-tuning yielded measurable gains. In this setup, CoT provided a structured path for logical inference, while the fine-tuned model is better equipped with domain-specific medical knowledge, leading to more robust and coherent answers.

Although RAG holds the potential to enhance accuracy by supplying external, domain-relevant context, it can also act as a "double-edged sword". When the retrieved evidence is high quality and closely aligned with the query, performance significantly improved—as evidenced by the highest accuracy of 81% with gte-Qwen2-1.5B-instruct. Conversely, noisy or tangential retrieval (e.g., e5-large-v2) misdirects the CoT and diminishes overall accuracy. Observations further indicated that a well-structured CoT complements the RAG process by facilitating the systematic integration of retrieved information, thus minimizing the risks posed by irrelevant context.

*Implications for Clinical Applications:.* From a practical standpoint, these results suggested that a medical assistant AI must integrate multiple reinforcing components. Merely prompting the model for CoT or fine-tuning in isolation can enhance clinical reasoning, but the jump to 81% indicated that reliable retrieval of high-fidelity medical knowledge is paramount to mitigating hallucinations and ensuring factually grounded answers. Nevertheless, even an 81% accuracy on open-ended clinical inquiries may still warrant a human-in-the-loop mechanism for real-world safety. This underscores the broader view that clinical decision support tools are more effective when they supplement, rather than supplant, professional judgment.

*Limitations.* Due to time and budget constraints associated with Google Colab usage, 100 questions were used for each configuration. Although this number may appear adequate, the dataset contains 40,600 questions, and the scalable evaluation framework could easily accommodate the entire dataset if resources allowed. The current process took approximately two hours per configuration for the 100 questions; scaling this to the entire dataset would require days of processing — a task that warrants future exploration. Additionally, the automatic evaluator relies on the LLM itself to judge correctness. This introduces the risk of circular reasoning and embedded model bias which may lead to overestimating performance. Although no mistakes were observed, external verification by medical professionals and an independent evaluator LLM would enhance robustness and credibility of the system evaluation. Utilizing the full dataset for evaluation would also likely help mitigate these potential issues. These are considerations for future work.

### *6.2. Medical Expert Qualitative Evaluation*

Having identified the best-performing configuration from the metric-based evaluations, the next phase of this research involves qualitative validation through expert feedback.

*Expert Participants.* This phase of the study involved three fourth-year medical students from Cardiff University and three practising GPs: two recently qualified, Dr Yensi Pardo Gonzalez and Dr María Porto, both practising in Bogota, Colombia, and Dr Elizabeth Long, practising in Dorset, UK, who brings 30 years of clinical experience. Their diverse backgrounds and first-hand clinical insights offered a practical lens for assessing the model's reasoning capabilities and utility in real-world healthcare settings.

*Scenario Generation.* Students and GPs were asked to propose challenging clinical scenarios they deemed most pertinent to everyday practice. Drawing upon their educational and clinical experience, the proposed scenarios were designed to rigorously evaluate the system's diagnostic and therapeutic reasoning in contexts that closely mirror real-world healthcare challenges. It is noteworthy that Dr Long and Dr Porto contributed only one scenario each, a limitation imposed by their busy schedules.

*Feedback Collection Process.* Once the scenarios were developed, the system's responses were shared with the participants. They were asked to score each answer out of 10 and provide qualitative feedback in six categories, defined as follows:

1. **Clinical Accuracy:** Ensuring that the information provided is factually correct and free from hallucinations.

2. **Completeness and Depth:** Evaluating how fully the response addresses the clinical scenario, checking for any missed or omitted details.
3. **Logical Reasoning Flow:** Assessing the CoT component to ensure that the intermediate reasoning steps follow a logical process to arrive at the final answer.
4. **Clinical Utility:** Determining whether the response is practically helpful in real-world settings by providing actionable and relevant information.
5. **Communication Style:** Focusing on the clarity and readability of the final response, ensuring it is easy to understand and well-organized for the end-user.
6. **Potential Pitfalls:** Identifying any risks, omissions, or areas of uncertainty in the response that could lead to misinterpretations or errors in clinical practice, thereby highlighting where additional caution or clarification might be needed.

*Rationale.* The primary goal of this qualitative phase is to supplement automated metrics with expert insights from individuals capable of discerning nuanced issues in patient presentations, diagnostic approaches, and treatment planning. Given the system's intended role as an AI assistant to aid clinicians, input from experienced practitioners and medical students offers a robust measure of real-world feasibility.

By combining these student-supplied scenarios with structured feedback protocols, this qualitative evaluation not only validates the model's performance in more realistic, expert-generated contexts but also illuminates any potential gaps in reasoning or content that automated metrics alone may not detect.

### 6.2.1. Medical Expert Feedback and Discussion

The three medical students provided two scenarios each, Dr Gonzalez provided five scenarios, and Dr Long and Dr Porto provided one scenario each. Overall, as shown in Figure 4, the feedback was highly positive, with combined scores of 95% for Clinical Accuracy, 89% for Completeness and Depth, 94% for Logical Reasoning Flow, 88% for Clinical Utility, 92% for Communication Style, and 80% for Potential Pitfalls. Every participant indicated that an AI system of this kind holds the potential to be used in future clinical practice.

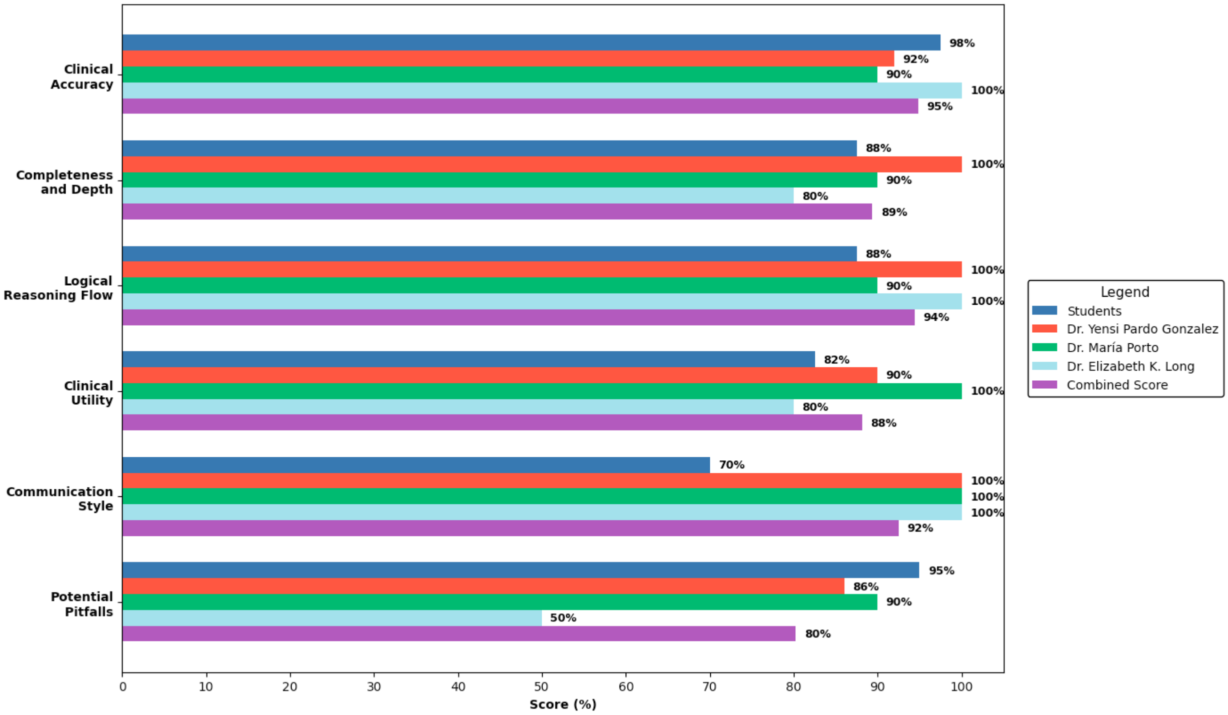

Fig. 4. Medical Expert Results

### 6.2.2. Qualitative Evaluation Discussion

*Key Observations.* Feedback from participants highlighted that the AI consistently delivered accurate and clinically robust diagnostic insights, alongside management strategies that were safe, transparent, and closely aligned with established guidelines and current medical evidence. High ratings—such as Clinical Accuracy scores of 98% from medical students, 92% from Dr Gonzalez, 90% from Dr Porto, and 100% from Dr Long—demonstrated the AI's effectiveness in accurately identifying key clinical issues and navigating complex patient management decisions. Clinical Utility also received strong ratings across participants, with scores reaching as high as 100% from Dr Porto and 90% from Dr Gonzalez, indicating that the AI offered practical, clinically relevant guidance.

Nevertheless, evaluators identified areas for improvement, particularly in explicitly communicating certain mandatory clinical guidelines and recommendations. For example, Dr Long gave a notably lower score (50%) in the Potential Pitfalls category, highlighting that although the AI correctly suggested appropriate protective measures, it did not sufficiently emphasize their mandatory status under established national guidelines, potentially creating ambiguity for clinicians. Communication Style was another area flagged by medical students, who assigned it the lowest rating at 70%, noting that the overly detailed responses might hinder quick understanding in high-pressure clinical situations. In contrast, experienced clinicians highly valued the structured, step-by-step explanations (with 100% ratings from Drs Gonzalez, Porto, and Long), suggesting that while concise guidance might better suit less experienced practitioners, the detailed reasoning enhanced clarity and confidence among seasoned medical professionals.

*Implications for Clinical Applications.* From a practical standpoint, these findings suggest that the system showed considerable promise as a supportive medical assistant. Its robust diagnostic accuracy and generally safe recommendations underscored its potential. Nonetheless, enhancements are needed to clearly emphasise critical clinical national guidelines (e.g. UK and US) and streamline communication, facilitating rapid, straightforward decision-making in clinical environments and reinforcing the importance of having a human in the loop. Emphasising this latter point, future work would also need to consider medico-legal implications such as liability for AI-generated clinical errors, patient safety and deployment risks, including safeguards against misuse, as well as trust calibration in real-world clinical settings, especially given the potential for misinterpretation of advisory versus mandatory clinical guidance.

*Limitations.* Engaging a larger group in future studies could help balance individual variances. Since the tool is designed to assist rather than dictate clinical decisions, the subjective nature of evaluations becomes apparent. For example, Dr. Long awarded a 50% score for Potential Pitfalls, reflecting a strict interpretation, as although the system advised the use of particular protective measures, it did not mandate it. In contrast, Dr. Gonzalez gave a perfect score of 100% for Completeness and Depth, reflecting overall satisfaction with the comprehensive responses despite noting some additional considerations in other categories. A broader range of evaluators and scenarios would likely yield a more balanced and nuanced assessment.

*Additional Comments from Dr Long.* Dr Long suggested exploring the system's capability to read medical records and handle imaging tasks, identifying these as valuable directions for future enhancements. Given her extensive clinical experience, these recommendations highlight important avenues for future development.

## 7. Conclusion

In conclusion, the study indicated that the combination of CoT and RAG supported by QLoRA fine-tuning provided an effective solution for the application of an AI medical assistant. It was shown that fine-tuning provided the model with medical domain-specific expertise, while RAG ensured precise retrieval of relevant contextual information from authoritative medical sources, grounding responses firmly in evidence-based content. CoT contributed structured and clear step-by-step reasoning, notably enhancing RAG by enabling the model to extract only the most relevant information from retrieved contexts. Together, these elements significantly improved the system's accuracy, relevance, and overall clinical utility.

Quantitative evaluations proved this configuration provided significant improvements, with accuracy rising from a baseline of 55% to 81%. These evaluations also highlighted the importance of selecting a high-quality embedding model. The E5-Large-v2 [20] model negatively impacted performance by retrieving irrelevant contexts, whereas using gte-Qwen2-1.5B-instruct [21] improved retrieval quality, leading to an 81% accuracy.

Qualitative assessments by practising GPs and medical students confirmed that the system delivers coherent, evidence-based responses that effectively support clinical decision-making. Participant's opinions varied with respect to the generated response length, with some appreciating the detailed explanations, while others found them overly lengthy, highlighting the need for customizability in the response communication style. Dr Long raised an important point: In the scenario she provided, the generated response recommended a treatment as advisory, when it should have been explicitly marked as mandatory. This underscores the importance of human oversight to ensure rigorous adherence to critical clinical guidelines. Importantly, all participants expressed agreement on the promising potential of this approach and its suitability for future clinical practice.

### *7.1. System Limitations*

The results of this study exceeded expectations, given several notable limitations related to the system itself. Section 4.1 highlighted the significant constraint posed by limited available VRAM of just 40GB. This restricted VRAM influenced the choice of model, resulting in the selection of a distilled version of the flagship DeepSeek R1 model (DeepSeek-R1-Distill-Qwen-14B [12]). Although powerful in its own right, the performance could have improved significantly if the full-scale flagship model had been used. Additionally, the project's short duration necessitated the use of open-source datasets, which limited the quantity and quality of available data. The study would have benefitted from access to proprietary datasets from sources such as the UK National Health Service (NHS), enabling the system to train on a diverse range of healthcare service databases and thereby enhancing its ability to generalise to real-world patient histories.

### *7.2. Future Work*

Future research could significantly enhance the system by exploring several promising avenues. Firstly, employing a larger and more powerful model, such as the flagship DeepSeek R1, would likely improve the accuracy and reliability of responses. Furthermore, integrating national medical records into the dataset would provide comprehensive patient histories, enabling the model to deliver more informed and contextually accurate responses. RAG techniques could be particularly effective with full access to patient medical records. This approach could allow the system to compare a patient's symptoms against historical, real-world medical records from hundreds of similar cases, indicating patterns and facilitating accurate, evidence-based diagnoses. In addition, some pre-trained LLMs, such as the model used in this project, can accommodate multiple languages, allowing the system to be fine-tuned in the required native language.

Addressing feedback provided by Dr Long, the model could be fine-tuned specifically to adhere strictly to UK medical guidelines, thereby enhancing its relevance and compliance within the UK healthcare context. Other considerations include the implications of liability for AI-generated clinical errors, safeguards against misuse and calibrating clinician trust in applied settings. Additionally, to address the mixed feedback from participants regarding communication style, a feature could be incorporated to allow medical practitioners to tailor interactions according to their individual preferences, such as adjusting verbosity and output style, further increasing the system's usability and acceptability in clinical practice.

Expanding beyond purely textual data, future work could investigate the integration of other AI architectures, such as Convolutional Neural Networks (CNNs), to incorporate high-accuracy medical imaging capabilities. Such integration could broaden the system's diagnostic scope. Additionally, other techniques, such as Reinforcement Learning (RL), could be explored to further reduce the likelihood of hallucinations and improve accuracy.

Ultimately, numerous additional technical improvements could be explored beyond those mentioned above. However, successfully implementing such enhancements primarily relies on obtaining and effectively utilizing high-quality, relevant data and addressing the ethical implications of using AI-generated diagnoses. Building on these promising early results, further work investigating the emerging future direction holds significant potential for developing robust AI-assisted tools to support the medical profession.

## 8. Acknowledgements

The Authors would like to acknowledge Dr Elizabeth Long, Dr Yensi Pardo Gonzalez, Dr María Porto, Sophia Evans, Chandu Wijenayake, and Mia Squires for their contribution to the studies qualitative analysis.

## References

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM computing surveys 55 (12) (2023) 1–38.

[2] J.-C. Bélisle-Pipon, Why we need to be careful with llms in medicine, Frontiers in Medicine 11 (2024) 1495582.

[3] E. Sezgin, Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers, Digital health 9 (2023) 20552076231186520.

[4] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, Explainable ai: A brief survey on history, research areas, approaches and challenges, in: Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8, Springer, 2019, pp. 563–574.

[5] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, W. Cheungpasitporn, Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications, Medicina 60 (3) (2024) 445.

[6] Y. Sonoda, R. Kurokawa, A. Hagiwara, Y. Asari, T. Fukushima, J. Kanzawa, W. Gonoi, O. Abe, Structured clinical reasoning prompt enhances llm's diagnostic capabilities in diagnosis please quiz cases, Japanese Journal of Radiology (2024) 1–7.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2) (2022) 3.

[8] M. Mills, Martha's rule: a hospital escalation system to save patients' lives (2023).

[9] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, B. Wang, Huatuogpt-o1, towards medical complex reasoning with llms, arXiv preprint arXiv:2412.18925 (2024).

[10] Federation of State Medical Boards and National Board of Medical Examiners, "United States Medical Licensing Examination." [Online]. Available: https://www.usmle.org/, [Accessed: Mar. 22, 2025].

[11] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, P. Szolovits, What disease does this patient have? a large-scale open domain question answering dataset from medical exams, Applied Sciences 11 (14) (2021) 6421.

[12] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948 (2025).

[13] G. Xiong, Q. Jin, Z. Lu, A. Zhang, Benchmarking retrieval-augmented generation for medicine, in: Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 6233–6251.

[14] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in neural information processing systems 36 (2023) 10088–10115.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[16] A. W. Qurashi, V. Holmes, A. P. Johnson, Document processing: Methods for semantic text similarity analysis, in: 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), IEEE, 2020, pp. 1–6.

[17] Meta AI, "Llama 3.1 8B Instruct," 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, [Accessed: Mar. 22, 2025].

[18] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, S. R. Bowman, Gpqa: A graduate-level google-proof q&a benchmark, in: First Conference on Language Modeling, 2024.

[19] Meta AI, "Llama 3.1 405B Instruct," 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.1-405B-Instruct, [Accessed: Mar. 22, 2025].

[20] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, arXiv preprint arXiv:2212.03533 (2022).

[21] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards general text embeddings with multi-stage contrastive learning, arXiv preprint arXiv:2308.03281 (2023).

[22] K. Enevoldsen, I. Chung, I. Kerboua, M. Kardos, A. Mathur, D. Stap, J. Gala, W. Siblini, D. Krzemiński, G. I. Winata, et al., Mmteb: Massive multilingual text embedding benchmark, arXiv preprint arXiv:2502.13595 (2025).