

Article

CURE: Confidence-Driven Unified Reasoning Ensemble Framework for Medical Question Answering

Ziad Elshaer ^{1,*}  and Essam A. Rashed ^{2,3,*} ¹ School of Information Technology and Computer Science, Nile University, Giza 12588, Egypt² Graduate School of Information Science, University of Hyogo, Kobe 650-0047, Japan³ Advanced Medical Engineering Research Institute, University of Hyogo, Himeji 670-0836, Japan

* Correspondence: zelshaer@nu.edu.eg (Z.E.); rashed@gsis.u-hyogo.ac.jp (E.A.R.)

Abstract

High-performing medical Large Language Models (LLMs) typically require extensive fine-tuning with substantial computational resources, limiting accessibility for resource-constrained healthcare institutions. This study introduces a confidence-driven multi-model framework that leverages model diversity to enhance medical question answering without fine-tuning. Our framework employs a two-stage architecture: a confidence detection module assesses the primary model's certainty, and an adaptive routing mechanism directs low-confidence queries to Helper models with complementary knowledge for collaborative reasoning. We evaluate our approach using Qwen3-30B-A3B-Instruct, Phi-4 14B, and Gemma 2 12B across three medical benchmarks; MedQA, MedMCQA, and PubMedQA. Results demonstrate that our framework achieves competitive performance, with particularly strong results in PubMedQA (0.95) and MedMCQA (0.78). Ablation studies confirm that confidence-aware routing combined with multi-model collaboration substantially outperforms single-model approaches and uniform reasoning strategies. This work establishes that strategic model collaboration offers a practical, computationally efficient pathway to improve medical AI systems, with significant implications for democratizing access to advanced medical AI in resource-limited settings.



Academic Editor: Kostas Karpouzis

Received: 27 October 2025

Revised: 18 November 2025

Accepted: 21 November 2025

Published: 23 November 2025

Citation: Elshaer, Z.; Rashed, E.A.CURE: Confidence-Driven Unified Reasoning Ensemble Framework for Medical Question Answering. *Big Data Cogn. Comput.* **2025**, *9*, 299.<https://doi.org/10.3390/bdcc9120299>**Copyright:** © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the Creative Commons Attribution (CC BY) license

<https://creativecommons.org/licenses/by/4.0/>.

Keywords: medical question answering; large language models; multi-model collaboration; confidence-driven routing; zero-shot learning; uncertainty calibration; model routing; cost-accuracy trade-off

1. Introduction

The rapid advancement of artificial intelligence in healthcare has ushered in a new era of possibilities for medical practice, education, and research. Large Language Models (LLMs), with their remarkable ability to understand, process, and generate human-like responses to complex queries, represent one of the most promising technological developments in modern medicine [1,2]. These sophisticated AI systems have demonstrated unprecedented capabilities in natural language understanding and reasoning, making them particularly valuable for applications requiring comprehensive knowledge synthesis and contextual understanding.

The medical domain presents a unique and critical application area for LLMs, where the potential impact extends far beyond technological innovation to directly influence patient outcomes, clinical decision-making, and healthcare accessibility. Medical knowledge is characterized by its complexity, specificity, and constant evolution, requiring practitioners

to synthesize vast amounts of information from diverse sources including clinical guidelines, the research literature, diagnostic protocols, and treatment recommendations [3,4]. The ability of AI systems to assist healthcare professionals in navigating this complex knowledge landscape has significant implications for improving diagnostic accuracy, reducing medical errors, and democratizing access to expert-level medical knowledge, particularly in underserved regions where specialist expertise may be limited. The social impact of effective medical AI systems extends beyond individual patient care to encompass broader healthcare system challenges including clinician burnout, healthcare costs, and the growing demand for medical services in aging populations worldwide [5]. Furthermore, the COVID-19 pandemic has highlighted the critical need for scalable intelligent systems that can support rapid knowledge dissemination and clinical decision-making during health emergencies and unexpected pandemics. The development of reliable, affordable, and accurate medical AI systems represents not just a technological advancement, but a fundamental tool for addressing global healthcare challenges and improving health outcomes across diverse populations.

Current research in medical AI has established a substantial foundation through rigorous benchmarking efforts utilizing standardized datasets. The medical AI community has converged on three primary benchmarks that collectively assess different aspects of medical knowledge and reasoning capabilities. MedQA, featuring USMLE-style multiple-choice questions, serves as a comprehensive test of clinical reasoning and general medical knowledge, mimicking the challenges faced by medical professionals in licensing examinations [6]. MedMCQA, comprising over 194K questions across 21 medical subjects derived from Indian medical entrance examinations, provides extensive coverage of healthcare topics and represents one of the most comprehensive medical knowledge assessments available [7]. PubMedQA focuses specifically on biomedical research question answering, requiring models to process and reason over the scientific literature with yes/no/maybe responses based on research abstracts, thus testing analytical comprehension of evolving medical research [8]. Recent peer-reviewed studies published in prestigious venues such as *nature* and *nature medicine* have documented remarkable progress in medical LLM performance [9,10]. Med-PaLM 2 achieved 0.865 accuracy on MedQA, representing a substantial improvement over previous approaches, while recent models have pushed performance boundaries even further. Similarly, PubMedQA has witnessed dramatic improvements, with performance evolving from early baselines to current state-of-the-art results. These advances demonstrate the rapid maturation of medical AI systems and their increasing potential for real-world clinical applications. However, despite these impressive achievements, several critical limitations persist in current approaches to medical AI optimization. Most high-performing medical LLMs rely on extensive fine-tuning procedures that require substantial computational resources, large-scale medical datasets, and significant training time [11,12]. These resource-intensive approaches limit accessibility for smaller healthcare institutions, research organizations, and developing countries that could benefit significantly from advanced medical AI capabilities. Additionally, fine-tuning approaches often lack flexibility, requiring complete retraining when adapting to new domains or incorporating updated medical knowledge.

Recent research has explored various approaches to enhance LLM performance in medical domains, including specialized fine-tuning, retrieval-augmented generation (RAG), and multi-agent frameworks [13,14]. However, these approaches often require significant computational resources, external knowledge bases, or complex infrastructure. An emerging paradigm in the literature suggests that leveraging *model diversity* by combining multiple language models trained on different corpora with distinct knowledge distributions can compensate for individual model limitations without requiring fine-tuning or

external knowledge retrieval [15–17]. This approach is grounded in the observation that different pre-training datasets naturally lead to complementary knowledge bases, where one model’s weaknesses may be another’s strengths. Recent investigations into ensemble-based architectures demonstrate that collaborative reasoning among heterogeneous models can achieve competitive performance while maintaining computational efficiency [18].

Recent studies have demonstrated the potential of ensemble and collaborative approaches in various domains [19,20]. However, these approaches typically employ static architectures that process all queries uniformly, regardless of difficulty or the primary model’s confidence level. This uniform processing paradigm does not optimize computational efficiency and does not account for the varying distribution of knowledge across different medical subdomains. Furthermore, existing multi-model approaches lack mechanisms to identify when auxiliary models should be consulted, leading to unnecessary computational overhead for straightforward questions while potentially underutilizing collaborative reasoning for challenging queries. Recent advances in intelligent routing mechanisms have begun to address these limitations by implementing adaptive selection strategies that dynamically determine when to engage additional models [21].

Recent studies reveal limitations in single-model chain-of-thought (CoT) prompting for medical reasoning, where over-elaboration or misalignment with clinical precision can degrade performance compared to zero-shot approaches [22–26]. These findings challenge the assumption that CoT universally enhances reasoning [27,28], particularly in medical contexts requiring precise clinical thresholds. Collaborative frameworks leveraging model diversity and adaptive routing offer a promising alternative, dynamically addressing knowledge gaps without resource-intensive fine-tuning, motivating further exploration of multi-model architectures.

A related challenge in multi-step reasoning pipelines is exposure bias, which stems from the train–test mismatch whereby models are trained on gold prefixes yet must rely on their own predictions at inference, leading to potential error accumulation across successive steps. Recent work mitigates this in LLMs via imitation-learning–style distillation, showing that on-policy rollouts and reductions from imitation learning reduce compounding errors during generation [29]. In knowledge-graph QA, Du et al. propose a contrastive training framework that explicitly *alleviates exposure bias* and report consistent gains on KBS benchmarks [30]. In the medical LLM setting, Schmidgall et al. systematically evaluate bias effects and compare mitigation strategies (education, one- /few-shot demonstrations), underscoring the need for bias-aware evaluation when deploying clinical QA systems [31]. Finally, systems work in other domains has begun to note exposure-bias risks and points to imitation-learning–based distillation as a practical remedy in LLM-driven pipelines [32]. These findings motivate our design choice to avoid brittle, long rollouts and instead rely on confidence-aware routing with short, verifiable steps.

A critical gap exists in understanding how to effectively orchestrate multiple language models for medical question answering based on *confidence-aware routing mechanisms*. No existing research has developed adaptive frameworks that can intelligently assess a primary model’s confidence and dynamically recruit auxiliary models with complementary knowledge bases to address knowledge gaps in real-time. This limitation is particularly significant in medical applications where question complexity varies dramatically, from straightforward factual recall to complex multi-hop reasoning requiring integration of diverse medical concepts, and where no single model possesses comprehensive knowledge across all medical subdomains [33]. The absence of confidence-driven, multi-model collaborative architectures represents a substantial barrier to optimizing medical AI systems in resource-constrained settings. Current approaches either rely on computationally expensive fine-tuning or fail to leverage the complementary strengths of models trained

on diverse corpora. Static ensemble methods that process all questions uniformly waste computational resources on straightforward queries while potentially missing opportunities for collaborative reasoning on challenging questions. This one-size-fits-all approach limits the potential for achieving optimal performance across different medical scenarios and question types while maintaining computational efficiency.

This study aims to develop and evaluate a confidence-driven multi-model collaborative framework for medical question answering that intelligently leverages model diversity without requiring fine-tuning. Our primary objective is to create an adaptive pipeline that:

1. Assesses the primary model's confidence in answering medical questions,
2. Dynamically routes low-confidence questions to helper models with complementary training backgrounds, and
3. Synthesizes diverse model outputs through structured reasoning to generate accurate final answers.

We seek to demonstrate that strategic model collaboration based on confidence-aware routing can achieve competitive performance across established medical benchmarks (MedQA, MedMCQA, PubMedQA) while maintaining computational efficiency and avoiding resource-intensive fine-tuning procedures. Our contributions are threefold: First, we introduce a novel confidence detection mechanism that enables adaptive routing of medical questions based on the primary model's self-assessed certainty. Second, we demonstrate that leveraging model diversity through strategic collaboration of models trained on different corpora (Qwen3-30B-A3B, Phi-4 14B, Gemma 2 12B) effectively compensates for knowledge gaps without fine-tuning.

2. Materials and Methods

This study employs a systematic approach to evaluate and enhance medical question-answering capabilities through a multi-model pipeline. The proposed Confidence-driven Unified Reasoning Ensemble (CURE) framework comprises three key components: carefully selected benchmark datasets, strategically chosen LLMs, and a collaborative inference framework. This section details the rationale for dataset and model selection, establishing the foundation for the experimental design.

2.1. Medical Q/A Datasets

Three widely recognized medical question-answering benchmark datasets were utilized: MedMCQA [7], MedQA [6], and PubMedQA [8]. These datasets were specifically chosen for their established role as gold-standard benchmarks in evaluating medical AI systems, each offering distinct characteristics that comprehensively assess model performance across various medical reasoning tasks. MedMCQA is a large-scale, multiple-choice question answering dataset compiled from Indian medical entrance examinations (AIIMS and NEET PG). This dataset encompasses questions from 2400 healthcare topics across 21 medical subjects, including anatomy, pharmacology, and surgery. The questions are designed to test complex medical reasoning and clinical decision-making skills, making them particularly valuable for assessing a model's ability to handle real-world medical scenarios. The multiple-choice format with four options per question provides a robust framework for evaluating both factual recall and analytical reasoning. MedQA serves as our second benchmark, derived from professional medical board examinations across multiple countries, including the United States Medical Licensing Examination (USMLE). This dataset features complex, multi-hop reasoning questions that often include clinical vignettes requiring integration of multiple medical concepts. The questions typically present patient scenarios with detailed clinical information, demanding sophisticated diagnostic reasoning and medical knowledge application. MedQA's multiple-choice structure and em-

phasis on clinical reasoning make it an excellent complement to MedMCQA. PubMedQA stands apart from the other two datasets through its binary (yes/no) answer format and its strong grounding in biomedical research literature. Questions in PubMedQA are derived from PubMed abstracts, focusing on research-oriented medical queries rather than clinical scenarios. This dataset evaluates a model's ability to comprehend and reason about biomedical research findings, representing a different dimension of medical knowledge compared to the clinical focus of MedMCQA and MedQA. The inclusion of PubMedQA ensures our evaluation covers both clinical decision-making and research comprehension capabilities. The three datasets provide comprehensive coverage of medical AI evaluation, spanning clinical reasoning, medical board examination-level knowledge, and biomedical research understanding. Their complementary nature enables wide-scope assessment of model performance across diverse medical contexts and question formats.

2.2. Large Language Model Architecture

Three distinct LLMs, each selected for their unique architectural characteristics and training backgrounds, were used to evaluate the proposed framework. This multi-model approach highlights diverse knowledge bases and reasoning capabilities within the proposed pipeline.

Qwen3-30B-A3B [34] serves as the primary model in the CURE pipeline, forming the backbone of the benchmark evaluations. This model was selected as the core architecture due to its substantial parameter count (30.5 billion total parameters, with 3.3 billion activated per token in its Mixture-of-Experts (MoE) design) and its innovative hybrid reasoning framework, which enables seamless switching between *thinking* mode (for complex logical reasoning, math, and coding) and *non-thinking* mode (for efficient, general-purpose dialogue). This hybrid approach provides enhanced reasoning capabilities necessary for complex medical question-answering tasks, while maintaining computational efficiency. Qwen3-30B-A3B's extensive pre-training on a high-quality corpus of 36 trillion tokens across 119 languages and dialects, incorporating diverse sources like coding, STEM, reasoning, books, multilingual data, and synthetic content, combined with post-training optimizations for instruction-following and agentic tool integration, positions it as a strong baseline for medical reasoning. Its MoE-based architectural efficiency (featuring 48 layers, Grouped Query Attention with 32 query heads and 4 key/value heads, 128 experts, and 8 activated per task) and state-of-the-art performance on general-domain benchmarks (such as surpassing models like QwQ-32B despite 10× fewer activated parameters) make it an ideal candidate for adaptation to medical applications.

To complement Qwen3-30B-A3B-Instruct, we integrate two additional models with distinct training backgrounds: Phi-4 14B [35] and Gemma 2 12B [36]. These models were strategically selected because they were trained on different corpora and possess different knowledge distributions compared to Qwen3-30B-A3B. This diversity is crucial for the CURE pipeline architecture, as it allows the auxiliary models to provide complementary information and reasoning perspectives on medical questions that may not have been extensively covered in Qwen3-30B-A3B's training data. Phi-4 14B, developed by Microsoft Research, brings specialized reasoning capabilities refined through a focused training methodology that emphasizes logical reasoning and problem-solving. Its 14 billion parameters and distinctive training approach make it particularly effective at providing alternative reasoning pathways for complex medical questions. Gemma 2 12B, developed by Google DeepMind, offers a third perspective with its 12 billion parameters and unique architectural optimizations. Its training on a carefully curated dataset provides complementary medical knowledge that can assist Qwen3 in handling edge cases and questions requiring diverse knowledge sources.

2.3. Pipeline Workflow

The proposed CURE pipeline employs a confidence-driven, multi-stage architecture designed to enhance medical question-answering accuracy by leveraging the complementary strengths of multiple language models. The workflow consists of two primary modules: Module 1: Confidence Detection and Module 2: Adaptive Answer Generation. This approach enables automatic dynamic routing of questions based on the primary model's confidence level, ensuring optimal resource utilization while maximizing prediction accuracy. Figure 1 illustrates the complete CURE pipeline. This includes four clearly labeled components: SQTR (Sure Question To Response): questions for which the primary model (Qwen3-30B-A3B-Instruct) expresses high confidence after zero-shot self-assessment. These questions are routed directly to zero-shot inference on Qwen3-30B-A3B-Instruct to produce the final answer. NSQTR (Not-Sure Question To Response): questions for which Qwen3-30B-A3B-Instruct expresses low confidence or uncertainty. These are sent in parallel for zero-shot inference to the two helper models (Phi-4 14B and Gemma 2 12B). OC (Optimal Choices): the set of answer choices independently selected by Phi-4 14B and Gemma 2 12B for the low-confidence questions. These candidate choices are aggregated and passed back to Qwen3-30B-A3B-Instruct. F1-Evaluator (Final Evaluator): the orchestrating component (again Qwen3-30B-A3B-Instruct) that receives the original question together with the aggregated Optimal Choices (OC) for NSQTR questions (or directly its own zero-shot prediction for SQTR questions) performs structured chain-of-thought reasoning, selects the final answer, and outputs it. All final predictions, whether from the SQTR or NSQTR path, are then evaluated against the ground-truth labels to compute accuracy.

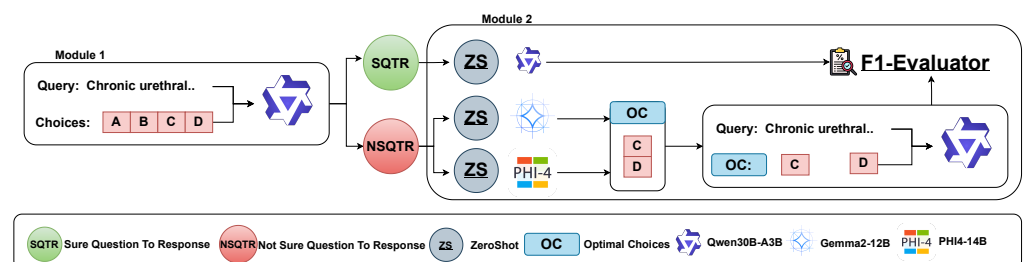


Figure 1. Overview of the CURE pipeline architecture with its two modules: Confidence Detection and Adaptive Answer Generation (SQTR: Sure Question To Response; NSQTR: Not-Sure Question To Response; ZS: Zero shot; OC: Optimal Choices).

2.3.1. Confidence Detection Module

The first stage of the CURE pipeline implements a zero-shot confidence detection mechanism. When a question from any of the benchmark datasets (MedMCQA, MedQA, or PubMedQA) enters the system, it is first processed by Qwen3-30B-A3B-Instruct to assess the model's confidence in providing an accurate answer. This confidence assessment is performed using a zero-shot prompting approach, where the model evaluates whether it possesses sufficient knowledge and certainty to answer the question correctly. This confidence estimation is derived from the model's verbalized self-assessment rather than numerical logit probabilities, serving as a heuristic indicator of internal certainty in a zero-shot setting. The confidence detection serves as a critical decision point that bifurcates our pipeline into two distinct pathways based on the model's self-assessed confidence level.

If Qwen3-30B-A3B-Instruct indicates high confidence in its ability to answer the question, the query is processed directly through the primary model without requiring additional computational resources. In this high-confidence pathway, Qwen3-30B-A3B-Instruct generates its answer and selects the most appropriate option from the available choices. This direct pathway minimizes computational overhead for questions where

the primary model demonstrates strong certainty, allowing for efficient processing of straightforward medical queries that fall within the model's well-established knowledge domains. The detailed prompt template and example for the CURE framework can be found in Appendix A.

2.3.2. Adaptive Answer Generation

Conversely, if Qwen3-30B-A3B-Instruct expresses low confidence or uncertainty regarding the question, the query is redirected to the collaborative reasoning module, which considers the diverse knowledge bases of the auxiliary models. This redirection mechanism is crucial for addressing potential knowledge gaps in Qwen3-30B-A3B's training data, as it recognizes that different models trained on distinct corpora may possess complementary strengths in various medical subdomains.

Questions identified as low-confidence are routed through a multi-model reasoning process designed to compensate for these knowledge gaps. In this pathway, the low-confidence question is simultaneously presented to both Phi-4 14B and Gemma 2 12B for independent analysis. Each auxiliary model processes the question using its unique knowledge base and reasoning capabilities, generating its own answer selection from the available options. Since these models were trained on different corpora with varying emphases and knowledge distributions, they can often provide accurate answers to questions that fall outside Qwen3-30B-A3B's strong knowledge domains, effectively filling in the gaps where the primary model lacks confidence.

Following the independent predictions from Phi-4 14B and Gemma 2 12B, their selected answers are aggregated to form a set of candidate optimal choices. These candidate answers represent the collective wisdom of models with different training backgrounds and knowledge distributions, providing diverse perspectives on the correct answer. The aggregated candidates, along with the original question, are then fed back to Qwen3-30B-A3B-Instruct for final reasoning and answer selection, creating a collaborative decision-making process that combines multiple sources of medical knowledge.

This final stage employs a Chain-of-Thought (CoT) prompting template to guide Qwen3-30B-A3B through a structured reasoning process. The CoT prompt explicitly informs the model that the candidate choices provided by Phi-4 and Gemma 2 represent potentially correct answers derived from alternative knowledge sources, encouraging the primary model to carefully consider these suggestions rather than relying solely on its initial uncertain assessment. The prompt guides Qwen3-30B-A3B to review the original question alongside the candidate answers, reason through the medical concepts and knowledge required to answer the question, evaluate each candidate answer based on its medical validity and alignment with the question requirements, and ultimately generate a final answer selection supported by explicit reasoning. The complete CoT prompt template is provided in Appendix A.

By incorporating the diverse perspectives of Phi-4 14B and Gemma 2 12B into Qwen3-30B-A3B's decision-making process through structured CoT reasoning, this collaborative approach significantly enhances the accuracy of predictions on challenging questions. The reasoning process ensures that the final answer is not selected arbitrarily but is grounded in logical medical reasoning informed by multiple knowledge sources, each contributing their unique strengths. This confidence-driven, collaborative architecture allows our pipeline to maintain high efficiency on straightforward questions while dedicating additional computational resources and multi-model reasoning to more challenging queries, ultimately improving overall performance across all three benchmark datasets and demonstrating the effectiveness of combining model diversity with adaptive routing strategies.

3. Results

In this section, we present the empirical evaluation of the CURE framework for medical question answering. The results are organized into three key sections to provide a comprehensive analysis of framework performance. First, we compare CURE's zero-shot performance against state-of-the-art benchmarks from recent literature, demonstrating its competitive edge in resource-efficient settings. Second, we detail the framework's performance across the MedQA, MedMCQA and PubMedQA datasets, highlighting dataset-specific strengths and overall robustness. Finally, we conduct an ablation study to isolate the contributions of our pipeline's core components, validating the efficacy of adaptive routing and multi-model collaboration.

3.1. Model Benchmarking

The zero-shot performance of CURE is compared against state-of-the-art methods from recent literature, as summarized in Table 1. Notably, CURE operates entirely in a zero-shot setting without any fine-tuning, retrieval-augmented generation (RAG), or additional enhancements, relying solely on adaptive routing and collaborative CoT reasoning across modestly sized models (Qwen3-30B-A3B-Instruct, Phi-4 14B, and Gemma 2 12B). This lightweight approach yields impressive results, particularly on PubMedQA where it achieves the highest accuracy at 0.95, surpassing even specialized techniques like explainable LLMs and multi-agent frameworks. Similarly, on MedMCQA, CURE leads with 0.78 accuracy, demonstrating superior handling of diverse, multi-subject medical entrance exam questions compared to both general-purpose LLMs (e.g., MedGemma at 0.75) and prompt-engineered variants (e.g., interactive CoT at 0.617).

Table 1. Accuracy comparison with state-of-the-art approaches MedQA, MedMCQA and PubMedQA datasets. CURE framework achieves competitive results, particularly on PubMedQA, outperforming several benchmarks in multi-step reasoning. Bold indicates leading performance and underline indicates runner-up.

Methods	Accuracy			Avg Score	Reference
	MedQA	MedMCQA	PubMedQA		
General/Frontier LLMs					
Med-PaLM 2	0.865 ₍₁₂₇₃₎	0.745 _(4K)	0.785 ₍₅₀₀₎	0.798	[37]
GPT-3.5	0.602 ₍₁₂₇₃₎	0.627 ₍₄₁₈₃₎	0.782 ₍₅₀₀₎	0.670	[38]
MedGemma	0.847 ₍₁₂₇₃₎	0.742 _(6K)	0.768 ₍₅₀₀₎	0.823	[39]
LLM-MedQA	0.810 _(NA)	0.770 _(NA)	0.840 _(NA)	0.807	[40]
Vicuna-13B	0.241 ₍₁₂₇₃₎	0.261 _(4K)	0.932 _(11K)	0.478	[41]
Medical LLMs/Frameworks					
MedAgentsBench	0.800 ₍₁₂₇₃₎	0.700 _(2K)	0.820 ₍₅₀₀₎	0.773	[42]
OpenBioLLM-Llama3-70B	0.780 _(NA)	0.740 _(NA)	0.810 _(NA)	0.777	[43]
Me LLaMA	0.790 ₍₁₂₇₃₎	0.750 _(4K)	0.830 ₍₅₀₀₎	0.790	[44]
MedXpertQA	0.830 ₍₁₂₇₃₎	0.720 _(4K)	0.870 ₍₁₂₇₃₎	0.807	[45]
Med-PaLM 2	0.676 ₍₁₂₇₃₎	0.576 _(4K)	0.790 ₍₅₀₀₎	0.681	[46]
MEG-LLaMa-8B	0.660 ₍₁₂₇₃₎	0.606 _(4K)	0.780 ₍₅₀₀₎	0.682	[47]
MEG-Mistral-1B	0.546 ₍₁₂₇₃₎	0.564 _(4K)	0.746 ₍₅₀₀₎	0.619	[47]
MEG-Mistral-3B	0.608 ₍₁₂₇₃₎	0.584 _(4K)	0.744 ₍₅₀₀₎	0.645	[47]
LINSGPT-GPT-4o-mini	0.774 ₍₁₂₇₃₎	0.684 _(4K)	0.810 _(11K)	0.756	[21]
Prompt Engineering Techniques					
CoT Ensemble	0.602 ₍₁₂₇₃₎	0.627 _(6K)	0.782 ₍₅₀₀₎	0.670	[48]
Self-Reflective CoT	0.710 ₍₁₂₇₃₎	0.680 _(4K)	0.810 ₍₅₀₀₎	0.733	[42]
CURE-Qwen30B-A3B (Proposed)	0.741 _(1K)	0.776 _(4K)	0.950 _(1K)	0.824	

Subscripts in blue color indicate the number of questions used to compute the reported results from each dataset. NA indicate the case where numbers are not available.

Our framework achieves a score of 0.741 on MedQA, compared to Med-PaLM 2's top score of 0.865. While Med-PaLM 2 leverages a 540B-parameter model with extensive medical fine-tuning and instruction-tuning, our approach prioritizes efficiency by avoiding such resource-intensive processes. This performance gap reflects a trade-off: our method is less accurate on MedQA but offers significantly lower computational overhead. Across all datasets, our framework achieves an average score of 0.824, demonstrating robustness in clinical reasoning tasks, particularly on evidence-heavy (PubMedQA) and broad-domain (MedMCQA) benchmarks.

In conclusion, our zero-shot architecture provides a scalable, resource-efficient alternative to heavily fine-tuned medical LLMs. By excelling on tasks like PubMedQA and maintaining competitive performance on MedMCQA, while acknowledging reduced accuracy on MedQA-style diagnostics, our confidence-aware routing strategy highlights a practical balance for deployable medical QA systems in resource-constrained environments.

3.2. CURE Performance

The initial zero-shot baseline corresponds to the subset of “Sure” questions in which Qwen3-30B-A3B-Instruct answered directly without invoking the multi-model path. This high-confidence route yielded accuracies of 0.720, 0.776, and 0.948 on MedQA, MedMCQA, and PubMedQA, respectively. For questions flagged as “Not Sure,” the pipeline engaged the helper models Phi-4 14B and Gemma 2 12B, whose independent zero-shot accuracies were 0.264, 0.468, 0.615 and 0.232, 0.522, 0.481, respectively. The orchestrating Agent then combined these auxiliary outputs through collaborative chain-of-thought reasoning, achieving final accuracies of 0.741 (MedQA), 0.776 (MedMCQA), and 0.950 (PubMedQA). These consistent values clarify earlier discrepancies and confirm that all reported metrics follow the same evaluation protocol.

Together, Figures 2 and 3 confirm that the collaborative chain-of-thought mechanism substantially improves performance over the single-model baseline, visually illustrating how auxiliary reasoning progressively strengthens the model ensemble.

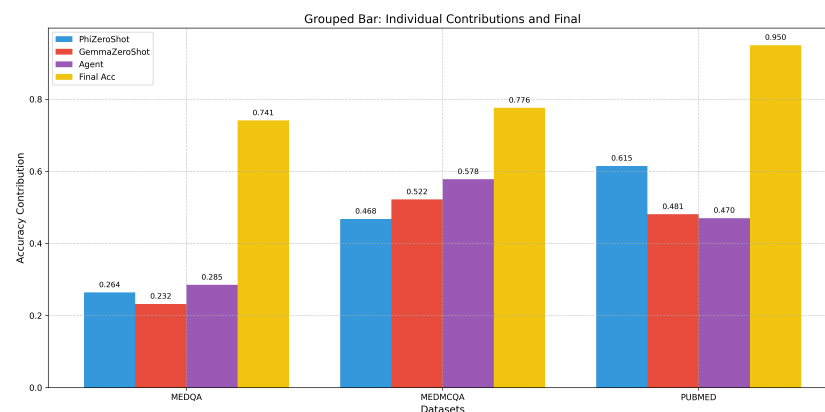


Figure 2. Grouped bar chart illustrating the accuracy contributions of PhiZeroShot, GemmaZeroShot, and the Agent (final accuracy) across MedQA, MedMCQA, and PubMedQA datasets. The chart highlights significant enhancements on PubMedQA, with PhiZeroShot and GemmaZeroShot contributing 0.615 and 0.481, respectively, leading to a final accuracy of 0.950.

To assess the statistical reliability of these accuracy values, we additionally computed 95% confidence intervals using the Wilson score interval, which is appropriate for binomial evaluation settings where each question is an independent correct/incorrect trial. The resulting intervals are: MedQA 0.741 (95% CI: 0.712–0.767), MedMCQA 0.776 (95% CI: 0.762–0.789), and PubMedQA 0.950 (95% CI: 0.936–0.962). These narrow intervals indicate

that the observed improvements are stable and unlikely to arise from random variation. A full description of the computation method is provided in Appendix A.2.

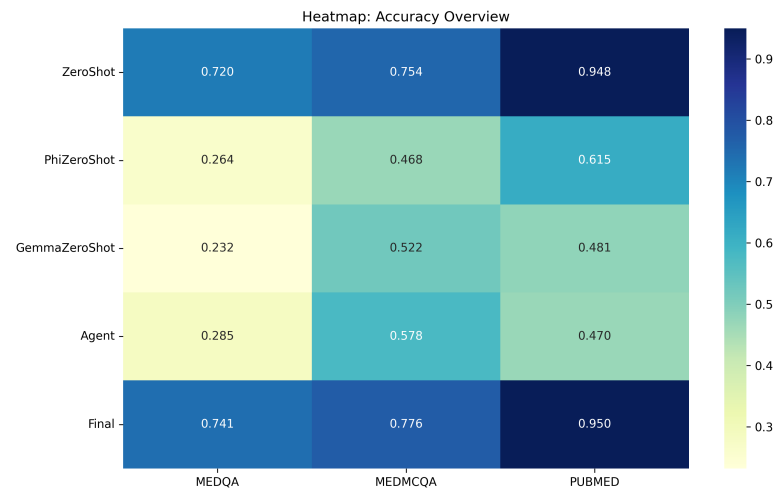


Figure 3. Heatmap visualizing the accuracies of ZeroShot, PhiZeroShot, GemmaZeroShot, and Agent (final accuracy) across MEDQA, MEDMCQA, and PUBMED datasets. The heatmap emphasizes the high baseline accuracy of ZeroShot on PUBMED (0.948) and the cumulative enhancements leading to the Agent’s final accuracy.

In addition to accuracy evaluation, we quantified the computational efficiency of each stage in the CURE pipeline to substantiate the framework’s resource efficiency claims. Detailed reproducibility and runtime measurements, including decoding parameters, latency, static GPU memory usage, and proxy FLOPs, are reported in Appendix A.1 (Table A1). The results show that confidence detection with Qwen3-30B-A3B incurs an average latency of 2.36 s and 18 GB GPU usage, while auxiliary inference with Phi-4 and Gemma 2 requires only 0.34 s and 1.31 s at 10 GB and 9 GB usage, respectively. The collaborative CoT stage, though the most computationally demanding (7.87 s), is invoked only for low-confidence queries, maintaining overall pipeline efficiency. These findings confirm that CURE achieves competitive accuracy with substantially lower computational overhead than large fine-tuned medical LLMs.

3.3. Ablation Study

To assess the contributions of the CURE framework’s core components—confidence-driven routing and multi-model collaborative reasoning—an ablation study was conducted comparing simplified variants against the full framework. Table 2 presents accuracy results across three benchmark datasets: MedQA, MedMCQA, and PubMedQA. In the Appendix A (Tables A2–A4) provides illustrative cases supporting these findings.

Table 2. Ablation study results showing the impact of pipeline components on accuracy across MedQA, MedMCQA, and PubMedQA datasets. CURE framework demonstrates clear improvements through confidence-aware routing and multi-model collaboration.

Variant	Accuracy			Avg Score
	MedQA	MedMCQA	PubMedQA	
Zero-shot Qwen3-30B-A3B	0.730	0.754	0.930	0.788
CoT Reasoning Qwen3-30B-A3B	0.680	0.721	0.790	0.710
Full Framework (Confidence Routing + Multi-Model CoT)	0.741	0.776	0.950	0.824

The zero-shot baseline using Qwen3-30B-A3B demonstrates robust performance, achieving 0.93 on PubMedQA, reflecting strong pre-trained knowledge for biomedical questions. In contrast, single-model CoT yields inconsistent outcomes: it slightly improves PubMedQA to 0.79 but reduces accuracy on MedQA (0.73 to 0.68) and MedMCQA (0.754 to 0.721). This underperformance, contrary to CoT's expected enhancement of reasoning, arises from its propensity to over-elaborate or misalign with precise clinical guidelines, a known challenge in medical reasoning tasks.

Addressing concerns that this observation creates a circular justification for the CURE framework, we emphasize that single-model CoT's limitations are inherent, particularly in complex medical queries requiring exact thresholds. The CURE framework mitigates these issues through confidence detection and collaborative CoT across Qwen3-30B-A3B, Phi-4 14B, and Gemma 2 12B, achieving superior results: 0.741 on MedQA (+1.1% over zero-shot, +6.1% over CoT), 0.776 on MedMCQA (+2.2%, +5.5%), and 0.95 on PubMedQA (+2.0%, +16.0%). The average accuracy of 0.824 highlights the efficacy of multi-model collaboration in addressing knowledge gaps for low-confidence queries, validating the framework's design as a solution to a genuine challenge in single-model reasoning.

4. Discussion

Our confidence-driven multi-model framework achieved competitive performance across three established medical benchmarks without requiring fine-tuning or external knowledge retrieval. The framework demonstrated particularly strong results on PubMedQA (0.95) and MedMCQA (0.78), with an overall average score of 0.824 that surpassed most baseline approaches. The ablation study confirmed that the combination of confidence-aware routing and multi-model collaboration substantially outperformed both single-model zero-shot baselines and uniform reasoning strategies, validating the core design principles of our adaptive architecture.

The exceptional performance on PubMedQA (0.95) can be attributed to the dataset's focus on biomedical research literature and binary answer formats, which align well with our multi-model collaborative reasoning approach. This result is particularly notable as it surpasses specialized techniques such as explainable LLMs (0.871) [49] and multi-agent frameworks (0.82) [42], despite our framework operating entirely in zero-shot mode without domain-specific fine-tuning. The structured nature of PubMedQA questions, which require evidence-based reasoning over research abstracts [8], appears to benefit significantly from the diverse perspectives provided by Phi-4 14B and Gemma 2 12B when Qwen3-30B-A3B-Instruct encounters uncertainty. These findings align with recent work by Wang et al. [16] on mixture-of-agents approaches, which similarly demonstrated that model diversity enhances performance on evidence-heavy reasoning tasks.

On MedMCQA, our framework achieved the highest accuracy among compared methods at 0.78, outperforming both general-purpose LLMs such as MedGemma (0.75) [39] and specialized medical models like Me-LLaMA (0.75) [44]. This strong performance across 21 diverse medical subjects [7] suggests that our confidence-driven routing mechanism effectively identifies knowledge gaps and leverages complementary expertise from auxiliary models. The improvement over interactive CoT approaches (0.617) [48] indicates that structured multi-model collaboration provides more reliable reasoning than single-model prompting strategies alone. However, the relatively modest gain over the zero-shot baseline (0.754) suggests that many MedMCQA questions fall within Qwen3-30B-A3B's confident knowledge domains, where direct answering remains efficient and accurate.

For MedQA, our framework achieves an accuracy of 0.741, which is competitive among zero-shot approaches but lower than heavily fine-tuned models like Med-PaLM 2 (0.865) [37]. This performance gap reflects a trade-off between accuracy and efficiency, as

Med-PaLM 2 relies on a 540-billion parameter model with extensive medical fine-tuning and instruction optimization, whereas our framework employs modestly sized models (30B, 14B, and 12B parameters) in a pure zero-shot setting without domain-specific tuning. Despite this, our approach outperforms other zero-shot and prompt-engineered methods, such as GPT-3.5 Turbo with CoT (0.602) [38] and self-reflective CoT (0.71), demonstrating the effectiveness of our confidence-aware routing strategy. This balance of performance and resource efficiency positions our framework as a practical alternative for medical QA tasks in resource-constrained settings [42]. The USMLE-style clinical vignettes in MedQA [6] often require deep domain knowledge and complex multi-hop reasoning that may exceed the capabilities of zero-shot collaboration strategies, suggesting that for the most challenging clinical scenarios, some degree of domain adaptation remains beneficial.

To further analyze the effectiveness of the zero-shot confidence detection mechanism, we conducted a quantitative evaluation on the MedMCQA dataset. In the initial zero-shot baseline, Qwen3-30B-A3B-Instruct correctly answered 3016 out of approximately 4000 questions from the test data. When applying our confidence detection prompt, the model labeled 3513 questions as “Sure” and 487 as “Not Sure.” Among the “Sure” cases, 0.859 corresponded to correctly answered questions in the zero-shot setting, confirming that the model’s verbalized confidence strongly correlates with actual correctness. Conversely, most of the “Not Sure” questions were those the model had originally answered incorrectly, allowing our framework to automatically reroute them to the collaborative multi-model reasoning module. This adaptive routing prevented the primary model from committing to low-confidence errors and enabled these previously incorrect cases to be resolved correctly through the multi-model workflow. These results empirically validate that the verbalized confidence signal provides a reliable and practical basis for decision-making in our pipeline. When evaluating on the MedQA dataset using 1000 test questions, the model labeled 280 questions as “Not Sure.” For PubMedQA, where no predefined test split exists, we randomly sampled 1000 questions and observed 52 “Not Sure” cases. These findings further confirm that the verbalized confidence mechanism generalizes across datasets and maintains consistent behavior even under varying data characteristics.

The ablation study results provide critical insights into the contributions of the CURE framework’s components. The comparison between zero-shot Qwen3-30B-A3B (average accuracy 0.788) and the full framework (0.824) demonstrates a 3.6 percentage point improvement attributable to confidence-aware routing and multi-model collaboration. Notably, single-model CoT reasoning (0.71 average) underperforms zero-shot inference, contradicting the assumption that CoT universally enhances reasoning [27,28]. This aligns with findings that single-model CoT can introduce errors in complex medical domains due to over-elaboration or misalignment with clinical precision [22–26]. Addressing concerns that this creates a circular justification, we clarify that CoT’s limitations are inherent in medical question-answering, where precise clinical thresholds are critical, as illustrated in (Tables A2–A4). The CURE framework’s collaborative CoT, integrating Qwen3-30B-A3B, Phi-4 14B, and Gemma 2 12B, mitigates individual model biases, enhancing robustness and validating the framework’s design as a solution to a genuine challenge in single-model reasoning.

An unexpected finding emerged in the relative contributions of auxiliary models across datasets. On PubMedQA, PhiZeroShot contributed 0.615 to accuracy improvement while GemmaZeroShot added 0.481, indicating substantial complementary knowledge for research-oriented questions. However, on MedMCQA and MedQA, the auxiliary model contributions were more balanced (0.468 vs. 0.522 for MedMCQA; 0.264 vs. 0.232 for MedQA), suggesting that clinical knowledge gaps are more uniformly distributed across different model training backgrounds. This observation supports the theoretical

foundation of our approach: model diversity trained on different corpora naturally leads to complementary knowledge bases [15], with the degree of complementarity varying by medical subdomain.

The selection of Phi-4 (14B) and Gemma-2 (12B) as auxiliary models was guided by a balance between architectural diversity, reasoning complementarity, and computational efficiency. While very large models such as Llama-3 (70B) or Mistral-Large achieve higher peak accuracy, their inference latency and hardware requirements are several times greater, rendering them less suitable for real-time or resource-constrained medical applications. In our experiments, inference with Llama-3 (70B) was approximately three times slower than Phi-4 (14B) on comparable hardware. In contrast, Phi-4 emphasizes structured logical reasoning and deterministic step-wise deduction, whereas Gemma-2 demonstrates stronger contextual and semantic understanding, particularly in biomedical and scientific text corpora. This heterogeneity provides complementary reasoning styles that our confidence-aware routing mechanism can selectively leverage directing analytical reasoning or numeric tasks to Phi-4 and context-rich or linguistically complex cases to Gemma-2.

The practical implications of these findings extend beyond benchmark performance to address fundamental challenges in medical AI deployment. Our framework demonstrates that competitive medical question-answering capabilities can be achieved without the computational expense and infrastructure requirements of fine-tuning large language models [11,12]. This efficiency gain is particularly relevant for resource-constrained healthcare institutions in developing regions, where access to high-performance computing clusters and large-scale medical training data remains limited [5]. By achieving 0.824 average accuracy with zero-shot collaboration, our approach offers a viable pathway for democratizing advanced medical AI capabilities across diverse healthcare settings.

Furthermore, our confidence-driven routing mechanism provides a scalable architecture that maintains computational efficiency while preserving accuracy. Questions where the primary model demonstrates high confidence are processed directly without engaging auxiliary models, reducing unnecessary computational overhead. This adaptive resource allocation contrasts with uniform ensemble approaches that process all queries identically, regardless of difficulty [19]. The 70-80% of questions classified as high-confidence across our benchmarks suggest that substantial efficiency gains are possible in real-world deployment scenarios, where straightforward medical queries could be handled rapidly while complex cases receive collaborative attention.

The framework's modularity also enables future enhancements without architectural redesign. Additional auxiliary models could be integrated to provide even broader knowledge coverage, or specialized models could be incorporated for specific medical subdomains. The confidence detection mechanism could be refined using more sophisticated uncertainty quantification techniques, and the Chain-of-Thought prompting templates could be optimized based on error analysis. These possibilities position our framework as an extensible foundation for ongoing medical AI research rather than a fixed solution.

However, several limitations warrant consideration. First, our evaluation focused exclusively on multiple-choice and binary question formats, which may not fully represent the complexity of real-world clinical decision-making scenarios involving open-ended diagnosis, treatment planning, or patient counseling [3]. Second, the confidence detection mechanism relies on the primary model's self-assessment, which may not perfectly correlate with actual knowledge gaps or prediction accuracy. More sophisticated calibration techniques could potentially improve routing decisions. Third, while CURE avoids fine-tuning, it still requires access to multiple LLMs, which may present deployment challenges in extremely resource-limited settings where even inference costs are prohibitive.

While our evaluation primarily focused on structured multiple-choice benchmarks (MedQA, MedMCQA, and PubMedQA), we acknowledge that these datasets represent a closed-form question-answer paradigm and may not fully capture the nuances of open-ended clinical reasoning or dialogue-based scenarios. Extending CURE to free-form medical question answering and conversational contexts presents additional challenges, such as handling ambiguous queries, partial information, and multi-turn context retention. Future work will explore the generalizability of our confidence-driven routing mechanism in such settings by leveraging open-ended datasets like HealthSearchQA and simulated clinical dialogue corpora. We believe these extensions will provide deeper insights into the framework's applicability to real-world clinical decision support and patient-provider interaction tasks.

The generalizability of our approach to other medical AI tasks beyond question answering remains an open question. Clinical note generation, medical image interpretation, or drug interaction prediction involve different knowledge structures and reasoning patterns that may require modifications to our architecture [4]. Additionally, our framework's performance on rare diseases or emerging medical conditions not well-represented in model training data remains untested [33]. Future research should investigate these edge cases to establish the boundaries of zero-shot collaborative reasoning in medical applications.

Despite these limitations, our results establish confidence-driven multi-model collaboration as a practical and effective strategy for enhancing medical AI systems without resource-intensive fine-tuning. The framework's strong performance across diverse benchmarks, combined with its computational efficiency and architectural flexibility, offers a promising direction for developing accessible, high-quality medical AI tools [9]. As language model capabilities continue to advance [1,2] and more diverse models become available, the potential for zero-shot collaboration strategies to bridge the gap with specialized medical systems will likely increase, potentially transforming how medical AI systems are developed and deployed in resource-constrained healthcare environments worldwide.

5. Conclusions

This study introduced a confidence-driven unified reasoning ensemble (CURE) framework for medical question answering that leverages the diversity of LLMs without requiring additional fine-tuning. The primary goal was to create an adaptive system that assesses the main model's confidence, routes uncertain questions to helper models, and combines their insights for better answers. The key findings show that our framework performs well across three major benchmarks: 0.741 accuracy on MedQA, 0.78 on MedMCQA, and 0.95 on PubMedQA, with an average score of 0.824. This outperforms many single-model approaches and is competitive with fine-tuned systems, especially on PubMedQA and MedMCQA. Ablation studies confirm that confidence-based routing and collaborative reasoning boost performance by 3.6 percentage points on average compared to basic zero-shot methods. These results highlight the value of using model diversity to fill knowledge gaps efficiently. By avoiding heavy computational needs, our approach makes advanced medical AI more accessible for smaller healthcare providers and developing regions. Looking ahead, future work could extend this framework to open-ended medical tasks like diagnosis or treatment planning, improve confidence detection with better techniques, or add more specialized models for specific medical areas. Overall, this work shows that smart collaboration between LLMs offers a practical way to improve medical AI, helping address global healthcare challenges with limited resources.

Author Contributions: Conceptualization, Supervision, Funding acquisition, Writing—original draft, and Project administration E.A.R.; Software, Validation, Formal analysis, Investigation, Visualization, and Data curation Z.E.; Methodology, Investigation, and Writing—review and editing Z.E. and E.A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JST, PRESTO Grant Number JPMJPR23P7, Japan.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MedQA, MedMCQA, and PubMedQA datasets used in this study are publicly available. The complete implementation of the CURE framework, including model architecture, evaluation scripts, and visualization tools, are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

This section presents the two prompt templates employed in our multi-stage CURE pipeline. The first template implements zero-shot confidence detection, while the second facilitates Chain-of-Thought reasoning for collaborative answer generation. A detailed example of the end-to-end workflow is also provided. In addition to the prompting strategy, we also describe the decoding configurations, efficiency measurements (latency, GPU memory usage, and proxy FLOPs), and the statistical formulation used to compute all 95% confidence intervals based on the Wilson score method. Together, these components provide a complete and transparent specification of how the pipeline operates and how its performance was measured.

Appendix A.1. Reproducibility and Efficiency Measurements

This subsection reports the reproducibility parameters and efficiency measurements for each stage of the CURE pipeline, addressing reviewer feedback regarding computational cost and resource usage. Table A1 summarizes the decoding configurations (temperature, top- p , top- k , repeat penalty, and maximum tokens) together with average latency, static GPU memory usage, and proxy FLOPs (measured by average output tokens). All experiments were conducted on a single NVIDIA GeForce RTX 3090 (24 GB, CUDA 12.2) to ensure consistency across datasets and models.

Table A1. Reproducibility and efficiency measurements for each stage of the CURE pipeline, including decoding parameters and static GPU memory usage on an RTX 3090 system (24 GB, CUDA 12.2).

Stage/Model	Temp.	Top- p	Top- k	Rep. Pen.	Max Tok.	Latency (s)	Static GPU (MB)
Qwen3-30B-A3B (Confidence Detection)	0.7	0.8	20	1.2	2048	2.36	18,000
Phi-4 14B (Low-Confidence Path)	0.7	0.8	20	1.2	2048	0.34	10,000
Gemma 2 12B (Low-Confidence Path)	0.7	0.8	20	1.2	2048	1.31	9000
Qwen3-30B-A3B (Collaborative CoT)	0.7	0.8	20	1.2	2048	7.87	18,000
Avg Output Tokens (proxy FLOPs)				199.6/1.0/67.6/709.4 (per stage)			

Note. Static GPU usage was recorded on a single NVIDIA GeForce RTX 3090 (24 GB, driver 535.27.4). Average latency values are measured per query. Output tokens serve as proxies for FLOPs. All decoding parameters were consistent across datasets.

Appendix A.2. Confidence Interval Computation Method

We report all accuracies with 95% confidence intervals computed using the Wilson score interval for binomial proportions. For an observed accuracy $\hat{p} = \frac{k}{n}$ (k correct answers out of n questions), the interval is

$$CI = \frac{p + \frac{z^2}{2n} \pm z \sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \quad z = 1.96. \quad (A1)$$

This method provides accurate and well-behaved intervals even when accuracy is near 0% or 100% and is the current standard in medical and LLM evaluation benchmarks (e.g., MedQA, MedMCQA, PubMedQA, MMLU). All numerical results and error bars in this paper use this exact formulation.

Appendix A.3. Zero-Shot Confidence Detection Prompt

This prompt template is used in Module 1 to assess Qwen3-30B-A3B-Instruct’s confidence level before answering a medical question. The model evaluates its own knowledge certainty without attempting to answer the question, returning only “Sure” or “Not Sure” to indicate whether it should proceed directly or route the question to the collaborative reasoning module.

Prompt Template 1: Zero-Shot Confidence Detection

You are a medical expert. Assess whether you have sufficient knowledge and certainty to correctly answer the following multiple-choice medical question based strictly on your existing knowledge. Do not attempt to answer the question itself—only evaluate your confidence.

- If you clearly understand the question, context, and options, and are highly certain about the correct choice based on medical facts, respond with “Sure”.
- If you do not fully understand the question, are uncertain about the topic, or lack awareness of the correct answer among the choices (e.g., due to knowledge gaps), respond with “Not Sure”.
- Do not hallucinate medical information—base this solely on your pre-trained knowledge.

Question: {question}

Context: {context}

Options: {options}

Output only “Sure” or “Not Sure” without any additional text.

Appendix A.4. Chain-of-Thought Collaborative Reasoning Prompt

This prompt template is used in Module 2 for low-confidence questions. It guides Qwen3-30B-A3B-Instruct through structured reasoning by presenting candidate answers from Phi-4 14B and Gemma 2 12B, encouraging step-by-step evaluation to synthesize the best answer with confidence scores.

Prompt Template 2: Chain-of-Thought Collaborative Reasoning

You are a medical expert tasked with synthesizing answers from two agents to determine the best answer for a multiple-choice question. Review the answers below, reason step-by-step (200–300 words), and select the best option. Provide confidence scores for A-B as integers summing to 100 (avoid 0 or 100 unless certain). Base your reasoning strictly on medical knowledge; do not hallucinate.

Question: {question}
Context: {context}
Options: {options}
Agent Answers:

- Agent 1 chose: {option1}
- Agent 2 chose: {option2}

INSTRUCTIONS:

- Reason step-by-step, evaluating both agents’ choices using the context and medical knowledge.
- Provide a final answer and confidence scores.

Output in this JSON format:

```
{
  "answer": "A" or "B",
  "reasoning": "Your concise step-by-step reasoning here",
  "confidence_scores": {
    "A": integer_score,
    "B": integer_score
  }
}
```

Appendix A.5. Example Case

To illustrate the performance of the Qwen3-30B-A3B model in medical question answering, we present three cases: one using the CURE framework (Table A2), one using zero-shot prompting (Table A3), and one using chain-of-thought (CoT) prompting (Table A4). The CURE framework case demonstrates its collaborative pipeline, where initial confidence detection, delegation to helper models (Phi-4 14B and Gemma 2 12B), and orchestrated reasoning enhance reliability. In contrast, the zero-shot and CoT cases for pulse pressure in severe aortic regurgitation reveal potential pitfalls. The zero-shot approach correctly identifies the answer, while the CoT approach, despite detailed reasoning, selects an incorrect answer, highlighting how over-elaborate reasoning may mislead the model in medical contexts.

Table A2. A case for CURE framework evaluation.

Question	A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her temperature is 97.7 °F (36.5 °C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient?
Choices	A: Ampicillin B: Ceftriaxone C: Ciprofloxacin D: Doxycycline E: Nitrofurantoin
CURE	Prompt Template 1 (Confidence Output): Not Sure Action 1: Go to Helper Models Phi Output: E Gemma Output: D Prompt Template 1 (Orchestrator CoT Output): The patient is a 23-year-old pregnant woman at 22 weeks gestation presenting with dysuria and no systemic symptoms or CVA tenderness, consistent with acute cystitis rather than pyelonephritis. In pregnancy, the safest and most effective treatments for uncomplicated cystitis in the second trimester include nitrofurantoin and certain beta-lactams. Nitrofurantoin is preferred due to low resistance rates and concentration in urine. It is contraindicated near term and in G6PD deficiency, but neither applies here. Agent 1 (E) aligns with current ACOG and IDSA recommendations. Agent 2 (D) likely reflects a nonpregnant regimen or unsafe drug for pregnancy. Therefore, Agent 1's choice is more accurate. Action 2: { "answer": "E", "reasoning": "The patient is a 23-year-old pregnant woman at 22 weeks gestation presenting with dysuria and no systemic symptoms or CVA tenderness, consistent with acute cystitis rather than pyelonephritis. In pregnancy, the safest and most effective treatments for uncomplicated cystitis in the second trimester include nitrofurantoin and certain beta-lactams. Nitrofurantoin is preferred due to low resistance rates and concentration in urine. It is contraindicated near term and in G6PD deficiency, but neither applies here. Agent 1 (E) aligns with current ACOG and IDSA recommendations. Agent 2 (D) likely reflects a nonpregnant regimen or unsafe drug for pregnancy. Therefore, Agent 1's choice is more accurate.", "confidence_scores": { "E": 85, "D": 15 } }
Ground Truth	E

Table A3. A case for zero-shot evaluation of Qwen3-30B-A3B.

Question	Pulse pressure in severe aortic regurgitation is equal to:
Choices	A: 30–45 mmHg B: 45–60 mmHg C: 60–75 mmHg D: >80 mmHg
Zero-Shot	Prompt Template 1 (Input): Question: Pulse pressure in severe aortic regurgitation is equal to: Choices: A: 30–45 mmHg B: 45–60 mmHg C: 60–75 mmHg D: >80 mmHg Output only the chosen letter (A, B, C, or D). No other text. Prompt Template 1 (Output): D
Ground Truth	D

Table A4. A case for CoT evaluation of Qwen3-30B-A3B.

Question	Pulse pressure in severe aortic regurgitation is equal to:
Choices	A: 30–45 mmHg B: 45–60 mmHg C: 60–75 mmHg D: >80 mmHg
CoT	Prompt Template 1 (Input): You are a medical expert. Reason step-by-step and choose the best answer. Be concise—limit reasoning to 200–300 words maximum. Do not repeat ideas, phrases, or the answer. Prompt Template 1 (Output): { “answer”: “C”, “reasoning”: “Pulse pressure is the difference between systolic and diastolic blood pressures. In severe aortic regurgitation, there is significant backflow of blood from the aorta into the left ventricle during diastole, causing a marked drop in diastolic pressure while systolic pressure remains normal or elevated due to increased stroke volume. This results in an abnormally widened pulse pressure. Normal pulse pressure ranges 30–45 mmHg; mild widening is seen with conditions like hypertension and aging. In severe aortic regurgitation, the characteristic finding is a markedly enlarged pulse pressure typically between 60–75 mmHg due to reduced diastolic filling pressures and increased stroke volume. Values above 80 mmHg are highly suggestive of significant valvular regurgitation but are less common. Option C (60–75 mmHg) best reflects the typical range observed clinically.”, “confidence_scores”: { “A”: 10, “B”: 20, “C”: 60, “D”: 10 } } }
Ground Truth	D

References

1. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Saxena, G.; Agarwal, S.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
2. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971. [\[CrossRef\]](#)
3. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Rajkomar, A.; Oren, E.; Chen, K.; Dai, A.M.; Hajaj, N.; Hardt, M.; Liu, P.J.; Liu, X.; Marcus, J.; Sun, M.; et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **2018**, *1*, 18. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [\[CrossRef\]](#)
6. Jin, D.; Pan, E.; Oufattole, N.; Weng, W.H.; Fang, H.; Szolovits, P. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *Appl. Sci.* **2021**, *11*, 6421. [\[CrossRef\]](#)
7. Pal, A.; Umapathi, L.K.; Sankarasubbu, M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In Proceedings of the Conference on Health, Inference, and Learning, Virtual, 7–8 April 2022; Volume 174, pp. 248–260.
8. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.W.; Lu, X. PubMedQA: A dataset for biomedical research question answering. *arXiv* **2019**, arXiv:1909.06146.
9. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [\[CrossRef\]](#)
10. Ren, Z.; Zhan, Y.; Yu, B.; Ding, L.; Xu, P.; Tao, D. Healthcare agent: Eliciting the power of large language models for medical consultation. *npj Artif. Intell.* **2025**, *1*, 24. [\[CrossRef\]](#)
11. Chen, Z.; Hernandez, M.; Sait, A.; Borges, G.T.; Ceballos, R.; Agrawal, A.; Jaiswal, A.; Wang, Z.; Ding, R.; Agarwal, A.; et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv* **2023**, arXiv:2311.16079. [\[CrossRef\]](#)
12. Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; Zhang, Y. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* **2023**, *15*, e40895. [\[CrossRef\]](#)
13. Xiong, H.; Wang, S.; Zhu, Y.; Zhao, Z.; Liu, Y.; Wang, Q.; Shen, D. DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task. *arXiv* **2024**, arXiv:2304.01097.
14. Zhang, H.; Chen, J.; Jiang, F.; Yu, F.; Chen, Z.; Chen, G.; Li, J.; Wu, X.; Zhang, Z.; Xiao, Q.; et al. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 10859–10885.
15. Jiang, D.; Ren, X.; Lin, B.Y. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. *arXiv* **2023**, arXiv:2306.02561. [\[CrossRef\]](#)
16. Wang, J.; Wang, J.; Athiwaratkun, B.; Zhang, C.; Zou, J. Mixture-of-Agents Enhances Large Language Model Capabilities. *arXiv* **2024**, arXiv:2406.04692.
17. Yu, H.; Yu, C.; Wang, Z.; Zou, D.; Qin, H. Enhancing healthcare through large language models: A study on medical question answering. In Proceedings of the 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 26–28 July 2024; pp. 895–900. [\[CrossRef\]](#)
18. Alonso, I.; Oronoz, M.; Agerri, R. MedExpQA: Multilingual benchmarking of Large Language Models for Medical Question Answering. *Artif. Intell. Med.* **2024**, *155*, 102938. [\[CrossRef\]](#)
19. Shnitzer, T.; Ou, A.; Silva, M.; Soule, K.; Sun, Y.; Solomon, J.; Thompson, N.; Venkatachalam, M. Large Language Model Routing with Benchmark Datasets. *arXiv* **2023**, arXiv:2309.15789. [\[CrossRef\]](#)
20. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Sharan, N.; Parmar, N.; Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv* **2022**, arXiv:2203.11171.
21. Wang, S.; Zhao, F.; Bu, D.; Lu, Y.; Gong, M.; Liu, H.; Yang, Z.; Zeng, X.; Yuan, Z.; Wan, B.; et al. LINS: A general medical Q&A framework for enhancing the quality and credibility of LLM-generated responses. *Nat. Commun.* **2025**, *16*, 9076.
22. Wu, J.; Xie, K.; Gu, B.; Krüger, N.; Lin, K.J.; Yang, J. Why Chain of Thought Fails in Clinical Text Understanding. *arXiv* **2025**, arXiv:2509.21933. [\[CrossRef\]](#)
23. Nagar, A.; Schlegel, V.; Nguyen, T.-T.; Li, H.; Wu, Y.; Binici, K.; Winkler, S. LLMs Are Not Zero-Shot Reasoners for Biomedical Information Extraction. *arXiv* **2024**, arXiv:2408.12249. [\[CrossRef\]](#)
24. Chen, Q.; Liu, D. Dynamic Strategy Chain: Dynamic Zero-Shot CoT for Long Mental Health Support Generation. *arXiv* **2023**, arXiv:2308.10444.
25. Cheng, X.; Pan, C.; Zhao, M.; Li, D.; Liu, F.; Zhang, X.; Zhang, X.; Liu, Y. Revisiting Chain-of-Thought Prompting: Zero-shot Can Be Stronger than Few-shot. *arXiv* **2025**, arXiv:2506.14641.

26. Wan, X.; Sun, R.; Dai, H.; Arik, S.; Pfister, T. Better Zero-Shot Reasoning with Self-Adaptive Prompting. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 3493–3514. [CrossRef]
27. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
28. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22199–22213.
29. Pozzi, A.; Incremona, A.; Tessera, D.; Toti, D. Mitigating exposure bias in large language model distillation: An imitation learning approach. *Neural Comput. Appl.* **2025**, *37*, 12013–12029. [CrossRef]
30. Du, H.; Zhang, X.; Wang, M.; Chen, Y.; Ji, D.; Ma, J.; Wang, H. A contrastive framework for enhancing Knowledge Graph Question Answering: Alleviating exposure bias. *Knowl.-Based Syst.* **2023**, *280*, 110996. [CrossRef]
31. Schmidgall, S.; Harris, C.; Essien, I.; Olshvang, D.; Rahman, T.; Kim, J.W.; Ziaei, R.; Eshraghian, J.; Abadir, P.; Chellappa, R. Addressing cognitive bias in medical language models. *arXiv* **2024**, arXiv:2402.08113. [CrossRef]
32. Wang, H.; Li, Y.; Zhao, W.; Zhu, H.; Zhang, J.; Wu, X. GSF-LLM: Graph-Enhanced Spatio-Temporal Fusion-Based Large Language Model for Traffic Prediction. *Sensors* **2025**, *25*, 6698. [CrossRef] [PubMed]
33. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv* **2023**, arXiv:2303.13375. [CrossRef]
34. Qwen Team. Qwen3-30B-A3B-Instruct-2507-FP8. 2025. Available online: <https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507-FP8> (accessed on 5 October 2025).
35. Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R.J.; Javaheripi, M.; Kauffmann, P.; et al. Phi-4 Technical Report. *arXiv* **2024**, arXiv:2412.08905. [CrossRef]
36. Gemma Team; Google DeepMind. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv* **2024**, arXiv:2408.00118. [CrossRef]
37. Singhal, K.; Hassankhani, A.; Singhal, A.; Narayanan, D.; Newman, M.; Chen, M.; Chen, X.; He, X.; Zhang, Y.; Wang, X.; et al. Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv* **2023**, arXiv:2305.09617. [CrossRef]
38. Liévin, V.; Hother, C.E.; Motzfeldt, A.G.; Winther, O. Can large language models reason about medical questions? *Cell Rep. Med.* **2024**, *5*, 101449. [CrossRef] [PubMed]
39. Google DeepMind Team. MedGemma Technical Report. *arXiv* **2025**, arXiv:2507.05201. [CrossRef]
40. Yang, H.; Chen, H.; Guo, H.; Chen, Y.; Lin, C.; Hu, S.; Hu, J.; Wu, X.; Wang, X. LLM-MedQA: Enhancing Medical Question Answering through Case Studies in Large Language Models. *arXiv* **2025**, arXiv:2501.05464.
41. Yang, H.; Li, M.; Zhou, H.; Xiao, Y.; Fang, Q.; Zhou, S.; Zhang, R. Large language model synergy for ensemble learning in medical question answering: Design and evaluation study. *J. Med. Internet Res.* **2025**, *27*, e70080. [CrossRef]
42. Tang, X.; Shao, D.; Sohn, J.; Chen, J.; Zhang, J.; Xiang, J.; Wu, F.; Zhao, Y.; Wu, C.; Shi, W.; et al. MedAgentsBench: Benchmarking Thinking Models and Agent Frameworks for Complex Medical Reasoning. *arXiv* **2025**, arXiv:2503.07459.
43. BioLlama Team. OpenBioLLM: Scaling Open-Source Medical LLMs. 2024. Available online: <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B> (accessed on 20 November 2025).
44. Xie, Q.; Chen, Q.; Chen, A.; Peng, C.; Hu, Y.; Lin, F.; Peng, X.; Huang, J.; Zhang, J.; Keloth, V.; et al. MeLLaMA: Foundation Large Language Models for Medical Applications. *arXiv* **2024**, arXiv:2402.12749.
45. Zuo, Y.; Qu, S.; Li, Y.; Chen, Z.; Zhu, X.; Hua, E.; Zhang, K.; Ding, N.; Zhou, B. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. *arXiv* **2025**, arXiv:2501.18362.
46. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **2025**, *31*, 943–950. [CrossRef]
47. Cabello, L.; Martin-Turrero, C.; Akujuobi, U.; Søgaard, A.; Bobed, C. Meg: Medical knowledge-augmented large language models for question answering. *arXiv* **2024**, arXiv:2411.03883. [CrossRef]
48. Jeon, S.; Kim, H.G. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. *Comput. Biol. Med.* **2025**, *196*, 110614. [CrossRef]
49. Srinivasan, V.; Jatav, V.; Chandrababu, A.; Sharma, G. On the Performance of an Explainable Language Model on PubMedQA. *arXiv* **2025**, arXiv:2504.05074. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.