

Article

Collaborative Medical Intelligence: A State-Machine-Based Multi-Agent Architecture for Multimodal Diagnostic Reasoning

Xueyan Zhang ^{1,2}, Xin Nie ^{1,*} , Yuankun Liu ¹  and Zongjun Wang ²¹ Faculty of Data Science, City University of Macau, Macau 999078, China; d23092100093@cityu.edu.mo (X.Z.); d22092100201@cityu.edu.mo (Y.L.)² School of Information Work Office, Shandong Youth University of Political Science, Jinan 250103, China; 080617@sdyu.edu.cn

* Correspondence: d23092100179@cityu.edu.mo

Abstract

Medical artificial intelligence systems often rely on a single model, which may give inconsistent answers and does not reflect the team-based nature of clinical consultation. This study develops a collaborative framework in which several agents analyze a medical question from different professional perspectives, discuss the evidence, vote on the proposed answer, and revise unresolved questions up to three times. Each step and stopping condition is recorded in a time-ordered workflow so that the decision process can be reviewed. The framework was evaluated on three regional versions of a medical examination dataset and three medical image question-answering datasets. It achieved an accuracy of 80.1% on the Mainland examination dataset and produced the highest reported results among the evaluated systems on ADAM. Stronger image models performed better on ACRIMA and Covid CT, showing that collaboration remains limited by the underlying visual model. Additional model-based quality scores are reported only as exploratory results because they were not validated by an independent model or human experts. These findings show that structured collaboration can make medical artificial intelligence workflows more transparent.

Keywords: Multi-agent System (MAS); medical diagnosis; artificial intelligence (AI); state machine; collaborative reasoning

MSC: 92B20

1. Introduction

Medical errors remain a major patient-safety concern and account for more than 250,000 deaths annually in the United States [1]. Diagnostic bias, failures in team communication, and the difficulty of interpreting complex clinical evidence all contribute to these errors [2]. Clinicians must also work with a growing range of data, including electronic health records, medical images, genomic information, and real-time monitoring signals [3]. Decision-support systems that can organize this evidence may help reduce workload and support safer diagnosis.

Large language models (LLMs) can interpret natural language, perform several forms of reasoning, and combine information from different sources [4]. These capabilities have led to their use in clinical decision support, medical question answering, and computer-aided diagnosis [5]. A single model, however, may produce hallucinated facts or inconsistent answers, which limits its suitability for high-stakes diagnostic use [6]. Most current



Academic Editor: Duarte Valério

Received: 12 June 2026

Revised: 2 July 2026

Accepted: 7 July 2026

Published: 16 July 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

systems also omit an important feature of clinical practice: multidisciplinary teams compare interpretations, question assumptions, and discuss difficult cases before reaching a decision [7].

Multi-agent systems (MASs) offer one way to represent this collaborative process. Studies in general-purpose reasoning suggest that coordinating specialized agents can improve robustness and reduce some errors made by individual models [8,9]. Multi-agent frameworks have also been used for task planning and interaction in embodied intelligence and robotics [10,11]. Related work on iterative feedback, tool-based critique, and evolutionary strategies has examined other ways to improve model reliability [12–14]. Their application to structured, iterative medical consultation remains less developed.

This study proposes a multi-agent diagnostic framework based on temporally unrolled DAG executions. The workflow combines task classification, dynamic role assignment, discussion, and gated voting. Its state representation exposes transitions and termination conditions, although structural auditability does not guarantee faithful explanations or clinical validity. The experiments provide evidence for this configuration on specific benchmarks rather than proof of general clinical superiority. This study makes three contributions. It models multidisciplinary consultation through role-specific analysis, discussion, and strict-majority voting; represents each finite diagnostic run as a temporally unrolled DAG with auditable routing and termination; and evaluates the framework on multilingual MedQA and three medical VQA subsets, including same-backbone comparisons and an explicit account of reliability limitations.

2. Related Work

Research on LLMs for medical diagnosis initially focused on fine-tuning pretrained models with medical corpora to encode domain knowledge [5]. Later studies examined whether prompting could elicit clinical reasoning without extensive retraining [15]. Continued pretraining on biomedical literature and clinical records has also produced specialized models that perform competitively on medical licensing examinations [16]. These single-model approaches nevertheless remain prone to factual hallucinations and generally lack the verification mechanisms needed for clinical deployment [6].

Multi-agent systems have been studied as a means of improving reasoning on complex tasks. In embodied intelligence, role specialization and consensus mechanisms support task planning and environmental interaction [9,10]. Iterative feedback and tool-based critique can likewise reduce some prediction errors [12,13]. Medical diagnosis presents a different setting, with stricter requirements for accuracy, traceability, and safety, and general-purpose agent designs do not directly satisfy these requirements [17].

Recent work has begun to specialize agent collaboration for medicine. MedAgents organizes multidisciplinary experts into iterative zero-shot discussions across medical question-answering benchmarks [18]. MDAgents selects solo or group collaboration according to estimated task complexity and combines moderator review with external medical knowledge [19]. MMedAgent focuses on multimodal tool selection, routing inputs to specialized medical models across modalities [20]. In a clinical-diagnosis setting, the Multi-Agent Conversation framework uses doctor agents and a supervisor to conduct rare-disease consultations and reports that supervisor ablation reduces diagnostic performance [21]. These studies establish the value of specialization, adaptive collaboration, tools, and supervision, but they leave complementary questions concerning explicit state-transition semantics, deterministic tie handling, bounded refinement, and cross-regional plus multimodal evaluation within one auditable workflow.

Table 1 positions the proposed method against these representative approaches. Unlike MedAgents, our method formalizes the finite execution as a temporally unrolled DAG

and specifies a strict-majority tie protocol. Unlike MDAgents, it emphasizes explicit state transitions rather than adaptive selection among Solo, Multidisciplinary Team (MDT), and Interdisciplinary Team (ICT) structures. Unlike MMedAgent, it uses role-based deliberation rather than primarily selecting external specialist tools. Compared with MAC, it evaluates both examination-style question answering and medical images, whereas MAC provides a more focused rare-disease consultation study with a dedicated supervisor ablation.

Table 1. Representative recent agent-based approaches relevant to medical reasoning.

Study	Domain	Core Mechanism	Evaluation Emphasis	Difference from This Work
MedAgents [18]	Medical QA	Multidisciplinary, multi-round expert discussion	Nine text QA datasets	No explicit temporally unrolled DAG or strict tie protocol
MDAgents [19]	Medical decision making	Complexity-adaptive Solo/MDT/ICT collaboration	Ten text, image, and video benchmarks	Adapts collaboration structure; our workflow emphasizes explicit, auditable state transitions
MMedAgent [20]	Multimodal medicine	Agent selection of specialized medical tools	Seven tasks across five modalities	Tool routing rather than role-specialized consensus refinement
MAC [21]	Rare-disease diagnosis	Multiple doctor agents plus supervisor	302 rare-disease cases	Focused rare-disease consultation; text-only scope
This work	Medical QA and VQA	Multi-label routing, role discussion, strict-majority gate, bounded rewriting	Multilingual MedQA, ACRIMA, ADAM, Covid CT	Temporally unrolled DAG with explicit tie and termination semantics

Traditional clinical decision-support systems rely mainly on rules or statistical models and may adapt poorly to changing clinical settings [22]. Deep learning has improved pattern recognition in medical images and electronic health records (EHRs) [23], while recent systems increasingly combine multimodal data with external knowledge bases [24]. These systems rarely model the deliberation through which clinicians from different specialties compare evidence and reach a shared decision [2].

3. Methodology

3.1. Overview of the Framework

We define a medical reasoning framework in which each finite run is represented as a temporally unrolled directed acyclic graph (DAG). The framework approximates the consultation process of a clinical team. Its compact finite-state template includes a conditional transition from query rewriting back to discussion and is therefore cyclic. During a diagnostic run, however, each visit to a repeatable state is assigned an iteration index k . This produces an execution graph $\mathcal{G}^{(K)} = (\mathcal{N}^{(K)}, \mathcal{E}^{(K)})$, where $0 \leq k \leq K$. The nodes represent task classification, role-based discussion, consensus voting, query rewriting, summarization, and failure at the iteration limit.

Let (s, k) denote state type s instantiated at refinement index k . For a maximum of K refinements, the unrolled graph contains:

$$\begin{aligned}
 \mathcal{N}^{(K)} = & \{(\text{Classify}, 0), (\text{Allocate}, 0)\} \\
 & \cup \bigcup_{k=0}^K \{(\text{Discussion}, k), (\text{Vote}, k), (\text{Summarize}, k)\} \\
 & \cup \bigcup_{k=0}^{K-1} \{(\text{Rewrite}, k)\} \cup \{(\text{Failure}, K)\}.
 \end{aligned} \tag{1}$$

The admissible transitions are:

$$\begin{aligned}
 \mathcal{E}^{(K)} = & \{(\text{Classify}, 0) \rightarrow (\text{Allocate}, 0), (\text{Allocate}, 0) \rightarrow (\text{Discussion}, 0)\} \\
 & \cup \bigcup_{k=0}^K \{(\text{Discussion}, k) \rightarrow (\text{Vote}, k), (\text{Vote}, k) \rightarrow (\text{Summarize}, k)\} \\
 & \cup \bigcup_{k=0}^{K-1} \{(\text{Vote}, k) \rightarrow (\text{Rewrite}, k), (\text{Rewrite}, k) \rightarrow (\text{Discussion}, k+1)\} \\
 & \cup \{(\text{Vote}, K) \rightarrow (\text{Failure}, K)\}.
 \end{aligned} \tag{2}$$

Only one outgoing transition from a Vote node is selected in a realized run. Assign within-index ranks 0, 1, 2, 3, 4 to Classify, Allocate, Discussion, Vote, and the terminal or Rewrite state, respectively. Every within-index transition increases the rank, and every Rewrite transition increases k . Lexicographic ordering by $(k, \text{rank}(s))$ is therefore a topological ordering of $\mathcal{G}^{(K)}$. Because K is finite, the unrolled graph is finite and acyclic.

The data flow starts from an input query Q , which contains text or an image with accompanying text. A successful run produces a diagnostic answer Y after discussion and voting; a run that reaches the refinement limit without a strict majority returns a failure status.

3.2. Medical Task Classification and Role Allocation

Upon receiving query Q , the system asks the language model to return a JSON object with one ‘disease_type’ string and one ‘modality_type’ string. The supported input modalities are Text and Image. These outputs are generated labels rather than values computed by a deterministic classifier; so, the prompt call is not represented as a mathematical function. The disease label is a routing description rather than a mutually exclusive clinical diagnosis. When domains overlap, the General Practitioner (GP) agent can provision specialists from several relevant disciplines using the full question even though the classifier returns one broad disease label. The GP agent then generates a finite set of role–task pairs $\mathcal{R} = \{(r_i, T_i)\}_{i=1}^n$. Appendix A provides the exact prompts used for classification and role provisioning.

3.3. Multi-Agent Discussion and Gated Voting

For each role–task pair $(r_i, T_i) \in \mathcal{R}$, the Collaborative Discussion and Voting (CDV) module prompts the same underlying model with the query and role-specific responsibility. Because model responses are sampled text, they are denoted by S_i without defining a deterministic response function. In the recorded implementation, each role receives a structured positive and negative assessment derived for that role and judges whether that role-specific summary supports the answer. It does not vote on a single summary shared by all roles. A valid serialized ‘yes’ is encoded as $v_i = 1$ and a valid ‘no’ as $v_i = 0$. Any other output is invalid and must be handled before routing rather than silently interpreted as agreement.

The supplied process workbook records the final serialized gate field ('binary_result') for one representative text case and one representative image case; both values are 'yes'. It does not preserve the individual v_i values or their order. The workbook therefore confirms only the serialized gate field and cannot be used to reconstruct per-agent vote vectors.

For a round k containing n_k valid votes, the vote counts are:

$$C_{\text{agree}}^{(k)} = \sum_{i=1}^{n_k} v_i^{(k)}, \quad C_{\text{disagree}}^{(k)} = \sum_{i=1}^{n_k} (1 - v_i^{(k)}) = n_k - C_{\text{agree}}^{(k)}. \quad (3)$$

A routing rule maps the current refinement index and current-round vote counts to an outgoing transition:

$$\text{next}(k, C_{\text{agree}}^{(k)}, C_{\text{disagree}}^{(k)}) = \begin{cases} \text{Summarize,} & C_{\text{agree}}^{(k)} > C_{\text{disagree}}^{(k)}, \\ \text{Rewrite,} & C_{\text{agree}}^{(k)} \leq C_{\text{disagree}}^{(k)} \text{ and } 0 \leq k < K, \\ \text{Failure,} & C_{\text{agree}}^{(k)} \leq C_{\text{disagree}}^{(k)} \text{ and } k = K. \end{cases} \quad (4)$$

The Rewrite and Failure branches share the absence of a strict majority but have disjoint round conditions. Rewrite is available only for $k \in \{0, \dots, K-1\}$, whereas Failure applies only at $k = K$. A perfect tie is therefore routed to Rewrite while another round remains and to Failure at the iteration limit. Disagreement is never overridden by a tie. This computational gate must not be interpreted as clinical authorization. In a clinical deployment, a tie, persistent disagreement, or any safety-critical dissent would require abstention and escalation to a qualified human clinician rather than autonomous action.

Repeated exposure to shared text can still produce groupthink. A valid analysis would require the vote vector, role identity, candidate answer, and correctness label at every cycle. The historical files do not contain these fields; so, this study neither defines nor reports a post hoc consensus-polarization statistic. This logging limitation is stated explicitly in the Discussion.

3.4. Iterative Query Refinement and Summarization

If the voting threshold is not met at round k , the Query Rewrite module uses the discussion feedback to produce $Q^{(k+1)}$ from $Q^{(k)}$. The rewrite is intended to resolve ambiguity or request missing clinical context. The next discussion runs on a new node (Discussion, $k+1$) rather than returning to (Discussion, k). Temporal expansion is limited to $K = 3$. If no strict majority has been reached when this limit is exhausted, the run terminates with a failure status and does not proceed to moderator summarization. The retained experimental CSV files do not contain convergence or failure flags or per-cycle vote counts; so, failed cases cannot be evaluated separately without rerunning the experiment. We report this limitation without inferring an unsupported subgroup result. The implementation therefore has a bounded recurrent transition in its compact form, while every temporally unrolled execution graph remains a DAG.

For runs routed to Summarize, the Moderator Agent is prompted with the original query and retained discussion text to generate the final answer. This is described as a generation step rather than a deterministic mathematical function. Runs routed to Failure do not receive a moderator-generated answer.

Figure 1 summarizes the end-to-end architecture, including role provisioning, voting, rewriting, termination, and the distinction between the compact transition template and its finite temporal unrolling.

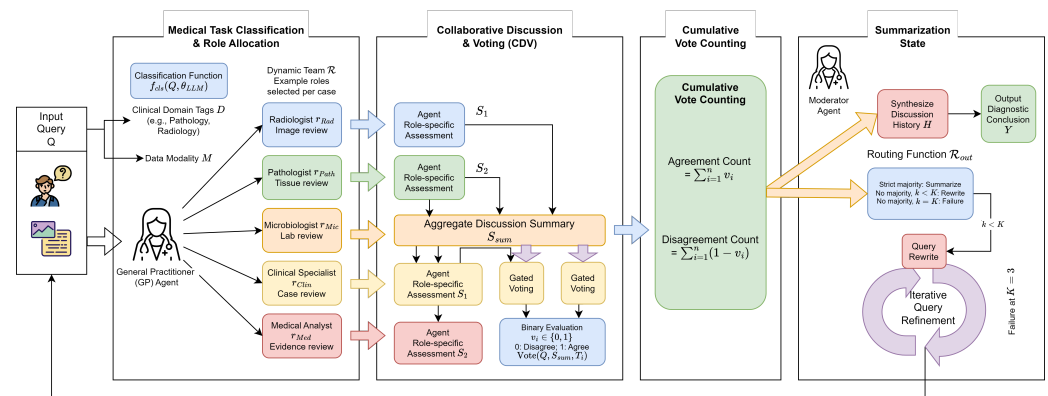


Figure 1. Architecture of the proposed multi-agent diagnostic framework. Black arrows denote data and control flow between states. Colored arrows connect allocated specialist roles to their corresponding role-specific assessment streams. Blue arrows indicate transitions between major workflow modules. Yellow arrows show routing from vote counting to summarization or query rewriting. The circular arrows denote iterative query refinement. A strict majority proceeds to moderation; a tie or disagreement creates a new refinement round, and the run is marked as failed if no strict majority is reached by $K = 3$.

4. Experiments

4.1. Experimental Setup

4.1.1. Model Configurations

We evaluate the proposed framework using a heterogeneous base model strategy. For textual reasoning and multi-agent coordination, we use Qwen2.5-7B-Instruct as the primary reasoning model because it follows structured instructions reliably. Clinical images are processed with MiniCPM-V-4.5. All agents use a temperature of $\tau = 0.5$, and query refinement is limited to $K = 3$. A common temperature keeps sampling behavior constant across roles, but we do not claim that it is optimal. Triage and voting may benefit from lower temperatures, while specialist reasoning may benefit from greater sampling diversity. The archived runs contain only final outputs; so, they cannot support a retrospective temperature analysis. Role-specific calibration therefore remains future work.

4.1.2. Hardware and Infrastructure

All experiments were conducted on a server equipped with an NVIDIA RTX 5090 GPU (Blackwell architecture, 32 GB GDDR7 VRAM, NVIDIA, Santa Clara, CA, USA). Inference used FP8 or BF16 precision for the text and multimodal tasks. The software stack consists of Python 3.10, PyTorch 2.4, and LangGraph 0.1.5 for workflow orchestration.

Table 2 reports the hardware and model configuration used for the experiments.

Table 2. Hardware and Model Specifications.

Component	Specification
Text Backbone	Qwen2.5-7B-Instruct
Vision Backbone	MiniCPM-V-4.5
GPU	NVIDIA RTX 5090 (32 GB)
Architecture	Blackwell
VRAM Standard	GDDR7
Framework	LangGraph/PyTorch
Inference Precision	FP8/BF16

4.2. Datasets

We evaluate the proposed configuration on textual medical question-answering and multimodal medical Visual Question Answering (VQA) datasets.

4.2.1. Textual Diagnostic Dataset: MedQA

We use MedQA [25] to assess clinical reasoning across languages. The dataset contains multiple-choice questions derived from professional medical licensing examinations. The evaluation includes the English MedQA-USMLE subset and the simplified- and traditional-Chinese MedQA-MCMLE subsets based on licensing examinations from Mainland China and Taiwan. These tasks challenge the agents to synthesize patient history and symptoms to select the correct diagnosis from five options.

4.2.2. Multimodal Medical VQA: OmniMedVQA Subsets

The multimodal evaluation uses three subsets of OmniMedVQA [26]. ACRIMA contains retinal fundus images for glaucoma screening and tests whether the system can identify optic-disc changes associated with glaucoma. ADAM, the Automatic Detection of Age-related Macular Degeneration challenge, contains fundus photographs with findings such as drusen, exudates, and hemorrhages. Covid CT contains computed tomography scans for binary COVID-19 classification based on findings that include ground-glass opacities and consolidation.

4.3. Evaluation Metrics

4.3.1. Traditional Metrics

We use accuracy, precision, recall, and F1-score to evaluate performance across datasets. For each class, these measures are calculated from true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs).

4.3.2. Automated Qualitative Scoring (LLM-as-a-Judge)

We use Qwen3-8B as an automated rater for an exploratory assessment of response quality. The resulting judge-rated scores are not treated as objective clinical measurements. The model grades the final output Y on Factual Accuracy (FA), which measures agreement with the reference answer and medical facts; Logical Coherence (LC), which assesses whether Y follows from the discussion history $\mathcal{H}$; and Clarity of Expression (CE), which covers fluency, terminology, and organization. Each dimension is scored from 0 to 100 using zero-shot prompting. Qwen3-8B and the Qwen2.5-7B reasoning backbone belong to the same broad model family; so, shared stylistic preferences may affect the scores. The reported runs were not evaluated by an independent model family or by human experts. We therefore treat the judge scores as exploratory and use task accuracy as the primary endpoint.

Table 3 defines the three judge-rated dimensions and their abbreviations: Factual Accuracy (FA), Logical Coherence (LC), and Clarity of Expression (CE). Here, FA always denotes the automated judge score and is distinct from answer-key-based classification Accuracy.

Table 3. Exploratory judge-rated scoring rubric using Qwen3-8B.

Metric	Description	Score Range
Factual Accuracy (FA; judge-rated)	Rater-assessed alignment with medical evidence and ground truth	0–100
Logical Coherence (LC)	Consistency of the reasoning chain from discussion to summary	0–100
Clarity of Expression (CE)	Readability, linguistic fluency, and professionalism	0–100

Figure 2 summarizes the quantitative and exploratory judge-based evaluation pathways.

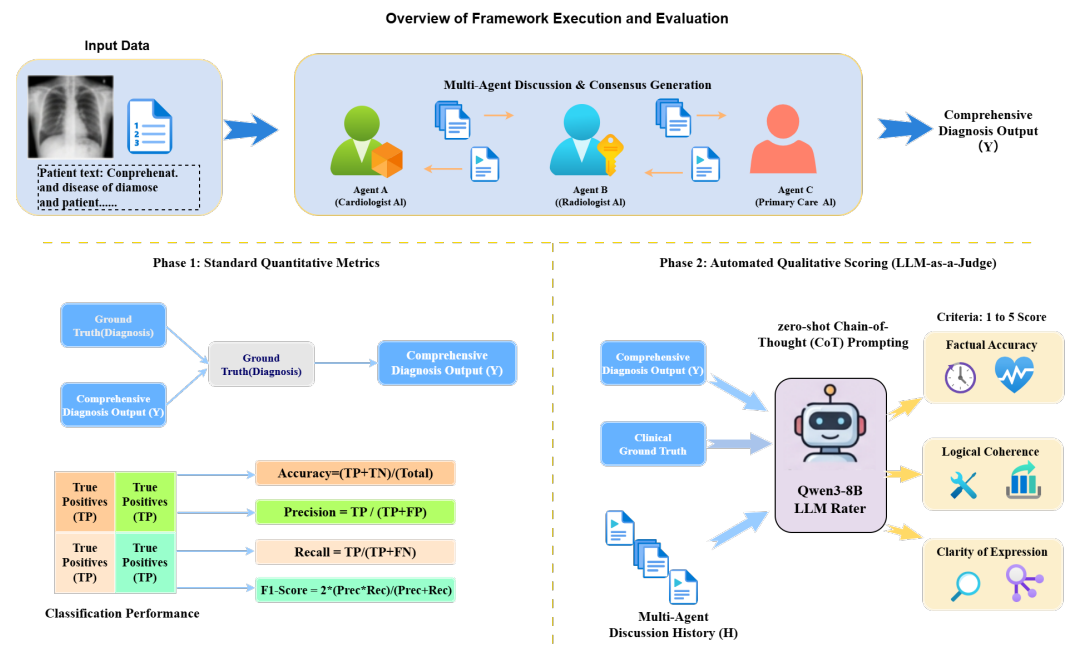


Figure 2. Overview of the proposed diagnostic framework and evaluation methodology. The pipeline consists of three main stages: (**Top**) Multi-agent discussion and consensus generation from multimodal input data; (**Bottom-Left**) Quantitative evaluation using traditional metrics derived from a confusion matrix; (**Bottom-Right**) exploratory LLM-as-a-Judge assessment (Qwen3-8B) of judge-rated factuality, logical coherence, and clarity. These secondary scores are distinct from classification accuracy.

4.3.3. Cross-Regional Evaluation and Metric Analysis

Table 4 reports the listed metrics across the three MedQA subsets. On Mainland MedQA, Multi_Agent records a multiple-choice accuracy of 0.801 and an F1-score of 0.582, compared with 0.558 and 0.443 for Vibe-Thinker. The Qwen2.5-7B row uses the same text backbone without multi-agent orchestration and is included as a direct reference point.

Table 4. Experimental results on MedQA dataset across different regions. Our Multi_Agent framework is compared against various baseline LLMs.

Region	Model	Accuracy	Precision	Recall	F1
Mainland	Multi-Agent (Ours)	0.801	0.591	0.574	0.582
	Llama3-8B	0.173	0.308	0.107	0.157
	Qwen3-8B	0.511	0.481	0.366	0.416
	Qwen2.5-7B	0.383	0.381	0.279	0.297
	Vibe-Thinker	0.558	0.505	0.399	0.443
Taiwan	Multi-Agent (Ours)	0.773	0.534	0.522	0.528
	Llama3-8B	0.198	0.243	0.100	0.141
	Qwen3-8B	0.302	0.390	0.204	0.268
	Qwen2.5-7B	0.265	0.256	0.155	0.168
	Vibe-Thinker	0.525	0.467	0.354	0.402
US	Multi-Agent (Ours)	0.715	0.515	0.514	0.513
	Llama3-8B	0.045	0.394	0.034	0.061
	Qwen3-8B	0.373	0.433	0.268	0.331
	Qwen2.5-7B	0.234	0.312	0.170	0.216
	Vibe-Thinker	0.395	0.359	0.287	0.312

Across the Taiwan and US subsets, the reported values show a similar ranking for several models. For the Llama3-8B US run, 1052 of 1273 records (82.64%) did not contain

a valid extracted option answer. Many affected generations were not literally blank; they were unusable responses, such as repetitions of the question or text that did not provide a parseable option. Because the generation script used unconstrained free text and did not enforce an option token, the reported value of 0.045 is affected by generation and answer-extraction failure.

On Taiwan MedQA, the proposed configuration records an F1-score of 0.528, compared with 0.268 for Qwen3-8B in the reported run. On US MedQA, its precision and recall are 0.515 and 0.514.

Llama3-8B exhibits an extreme imbalance between precision and recall in the affected US run, while Vibe-Thinker also shows metric differences across regions.

Table 5 reports secondary Qwen3-8B judge scores across the Mainland, Taiwan, and US subsets. The Multi-Agent output receives a judge-rated Factual Accuracy (FA) score of 81.26% for Mainland and Logical Coherence (LC) scores above 83% in all three subsets. These values must not be confused with classification accuracy: 0.801 in Table 4 is an answer-key-based multiple-choice accuracy, whereas 81.26% is the mean score assigned by an automated judge on a 0–100 rubric. Because the judge is not independent of the Qwen model family, these results are exploratory and do not by themselves establish suppression of hallucination or clinical superiority.

Table 5. Exploratory Qwen3-8B judge scores across Mainland, Taiwan, and US versions of MedQA. FA denotes judge-rated Factual Accuracy and is distinct from answer-key-based classification accuracy; LC denotes Logical Coherence and CE denotes Clarity of Expression.

Region	Model	FA (%)	LC (%)	CE (%)
Mainland	Llama-3-8B	22.62	45.41	58.93
	Qwen3-8B	49.96	75.50	84.00
	Qwen2.5-7B-Inst.	45.31	73.58	82.44
	Vibe-Thinker	48.79	66.17	75.63
	Multi-Agent (Ours)	81.26	84.53	85.81
Taiwan	Llama-3-8B	26.40	51.04	63.53
	Qwen3-8B	44.11	71.51	81.51
	Qwen2.5-7B-Inst.	43.53	72.10	81.23
	Vibe-Thinker	49.81	65.87	75.74
	Multi-Agent (Ours)	79.91	83.36	84.90
US	Llama-3-8B	38.16	63.23	75.56
	Qwen3-8B	47.00	72.08	82.55
	Qwen2.5-7B-Inst.	41.98	68.90	81.06
	Vibe-Thinker	45.86	68.27	77.02
	Multi-Agent (Ours)	78.46	83.69	85.20

Figure 3 visualizes the quantitative and judge-rated MedQA results across the three regional subsets.

4.4. Glaucoma Detection on ACRIMA Dataset

As illustrated in Table 6, the proposed framework achieves an Accuracy of 0.735 and an F1-score of 0.495, matching MiniCPM-V 2.5 in Accuracy and showing a 0.029 difference between Precision and Recall. Qwen3-VL-8B achieves the highest performance on this dataset (Accuracy 0.861; F1-score 0.697), exceeding the proposed configuration by 12.6 and 20.2 percentage points, respectively.

As illustrated in Table 7, the exploratory judge scores for the Multi-Agent framework are 75.18% Factual Accuracy (FA), 82.42% Logical Coherence (LC), and 84.84% Clarity of

Expression (CE). These model-based scores are secondary to the classification metrics and are subject to the evaluator-family limitation described above.

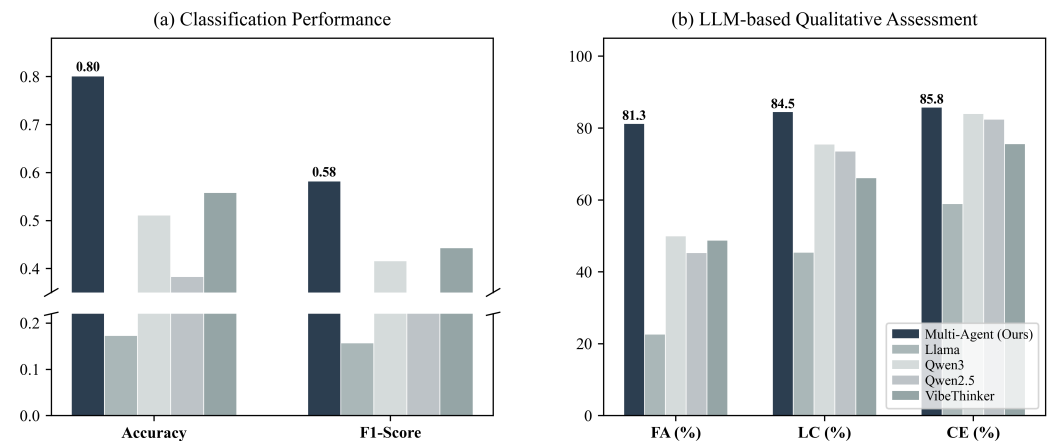


Figure 3. MedQA evaluation across three regional subsets. (a) Answer-key-based classification metrics. (b) Exploratory Qwen3-8B judge scores; these do not constitute independent evidence of hallucination reduction or clinical reliability.

Table 6. Experimental results on the ACRIMA dataset for glaucoma diagnosis. Performance is evaluated using Accuracy, Precision, Recall, and F1-score.

Model	Accuracy	Precision	Recall	F1-Score
Multi-Agent (Ours)	0.735	0.511	0.482	0.495
MiniCPM-V 2.6	0.358	0.404	0.233	0.295
MiniCPM-V 2.5	0.735	0.524	0.486	0.503
Qwen2.5-VL-7B	0.679	0.573	0.525	0.548
Qwen3-VL-8B	0.861	0.706	0.688	0.697

Table 7. Exploratory Qwen3-8B judge scores on ACRIMA: Factual Accuracy (FA), Logical Coherence (LC), and Clarity of Expression (CE).

Model	Inst.	FA (%)	LC (%)	CE (%)
MiniCPM-V 2.6	OpenBMB	52.73	74.52	82.10
MiniCPM-V 4.5	OpenBMB	73.05	83.23	85.69
Qwen2.5-VL-7B	Alibaba	63.42	77.57	82.07
Qwen3-VL-8B	Alibaba	74.84	80.25	84.30
Multi-Agent (Ours)	Authors' framework	75.18	82.42	84.84

Figure 4 provides a compact visual comparison of the ACRIMA classification metrics and judge-rated scores.

4.5. Experimental Results on ADAM Dataset

Table 8 reports an accuracy of 0.850 and an F1-score of 0.855 for the proposed configuration, the highest values among the evaluated systems in this ADAM experiment. Precision (0.867) and recall (0.853) are numerically close. These results are descriptive and benchmark-specific: because the visual backbones and orchestration settings are not controlled factorially, they do not establish that agent coordination caused the observed differences or improved clinical reliability.

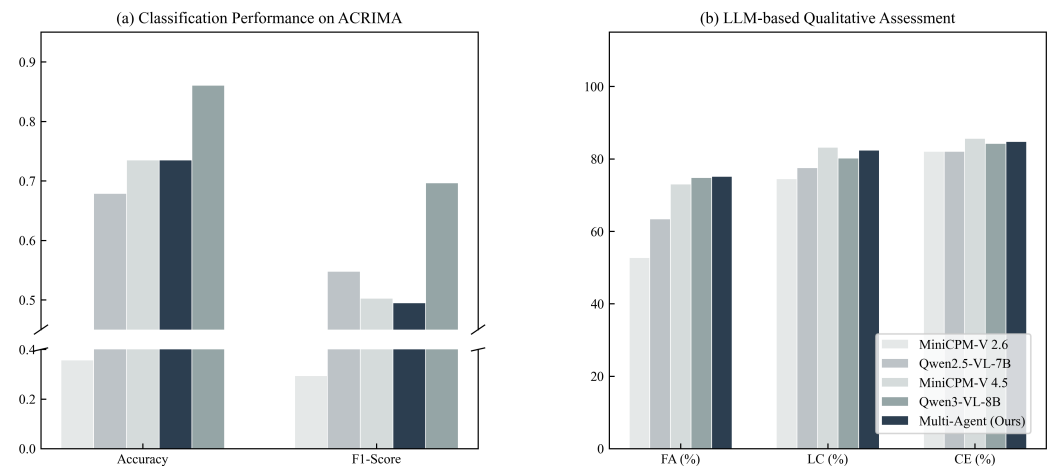


Figure 4. Quantitative and qualitative performance comparison on the ACRIMA dataset. (a) Classification results including Accuracy and F1-score across different models, with a broken y-axis highlighting competitive margins in the high-performance range. (b) exploratory LLM-based evaluation assessing Factual Accuracy (FA), Logical Coherence (LC), and Clarity of Expression (CE). Numerical annotations denote the reported scores; the judge-rated panel is exploratory and does not independently validate framework effectiveness.

Table 8. Experimental results on the ADAM dataset. The performance of our Agent framework is compared against state-of-the-art vision-language models.

Model	Accuracy	Precision	Recall	F1-Score
Multi-Agent (Ours)	0.850	0.867	0.853	0.855
MiniCPM-V 2.6	0.540	0.436	0.427	0.430
MiniCPM-V 4.5	0.609	0.641	0.606	0.608
Qwen2.5-VL-7B	0.770	0.573	0.512	0.540
Qwen3-VL-8B	0.701	0.650	0.560	0.598

Table 9 reports exploratory judge scores for ADAM. The Multi-Agent outputs receive 79.65% FA, 84.71% LC, and 85.63% CE. These values describe the Qwen3-8B rater's assessment and are not human-expert validation or direct classification accuracy.

Table 9. Exploratory Qwen3-8B judge scores on ADAM: Factual Accuracy (FA), Logical Coherence (LC), and Clarity of Expression (CE).

Model	Inst.	FA (%)	LC (%)	CE (%)
MiniCPM-V 2.6	OpenBMB	56.55	77.06	83.16
MiniCPM-V 4.5	OpenBMB	65.57	81.49	85.57
Qwen2.5-VL-7B	Alibaba	70.11	81.89	83.85
Qwen3-VL-8B	Alibaba	60.45	77.12	82.58
Multi-Agent (Ours)	Authors' framework	79.65	84.71	85.63

Figure 5 summarizes the quantitative and judge-rated comparisons on ADAM.

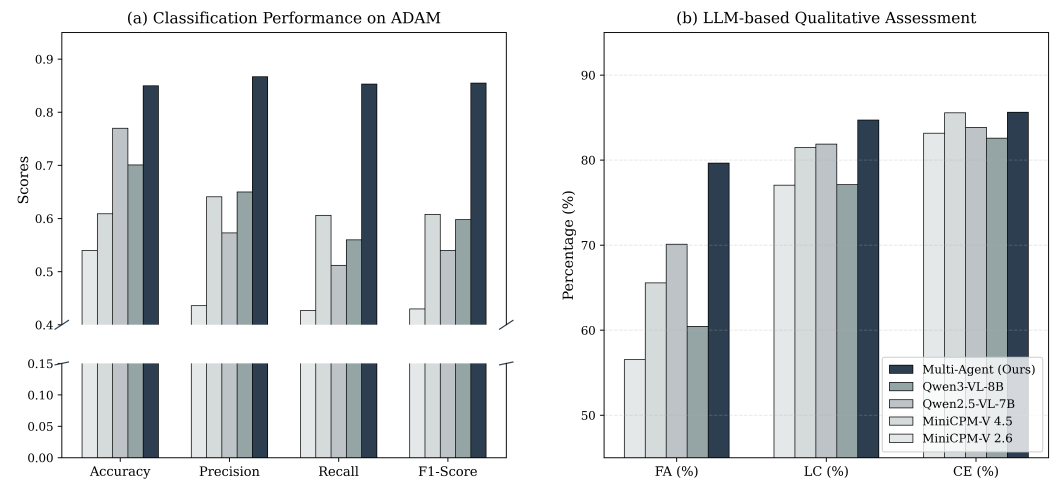


Figure 5. Performance comparison on the ADAM dataset. **(a)** Classification Performance across four quantitative metrics. **(b)** Exploratory Qwen3-8B judge scores for Factual Accuracy (FA), Logical Coherence (LC), and Clarity of Expression (CE).

4.6. Performance Analysis of Multi-Agent Framework on Covid CT Dataset

Table 10 summarizes the performance comparison for pneumonia triage on Covid CT. The proposed framework records an accuracy of 0.758 and an F1-score of 0.805, exceeding MiniCPM-V 2.6 and Qwen2.5-VL-7B on the reported F1-score. MiniCPM-V 4.5 records the highest accuracy and F1-score in this experiment and outperforms the proposed framework on all four reported metrics.

Table 10. Experimental results on the Covid CT dataset for pneumonia triage. Performance metrics include Accuracy, Precision, Recall, and F1-score.

Model	Accuracy	Precision	Recall	F1-Score
Multi-Agent (Ours)	0.758	0.800	0.814	0.805
MiniCPM-V 2.6	0.542	0.506	0.473	0.468
MiniCPM-V 4.5	0.783	0.842	0.834	0.833
Qwen2.5-VL-7B	0.723	0.630	0.626	0.628
Qwen3-VL-8B	0.743	0.804	0.808	0.806

Table 11 reports exploratory judge scores for Covid CT. The Multi-Agent outputs receive 72.21% FA, 82.08% LC, and 84.82% CE. These automated assessments are reported unchanged but should not be interpreted as independent clinical validation.

Table 11. Exploratory Qwen3-8B judge scores on Covid CT: Factual Accuracy (FA), Logical Coherence (LC), and Clarity of Expression (CE).

Model	Inst.	FA (%)	LC (%)	CE (%)
MiniCPM-V 2.6	OpenBMB	66.13	79.82	85.45
MiniCPM-V 4.5	OpenBMB	72.56	81.15	85.22
Qwen2.5-VL-7B	Alibaba	68.41	79.92	83.29
Qwen3-VL-8B	Alibaba	68.44	79.47	83.31
Multi-Agent (Ours)	Authors' framework	72.21	82.08	84.82

Figure 6 visualizes the corresponding quantitative and judge-rated results on Covid CT.

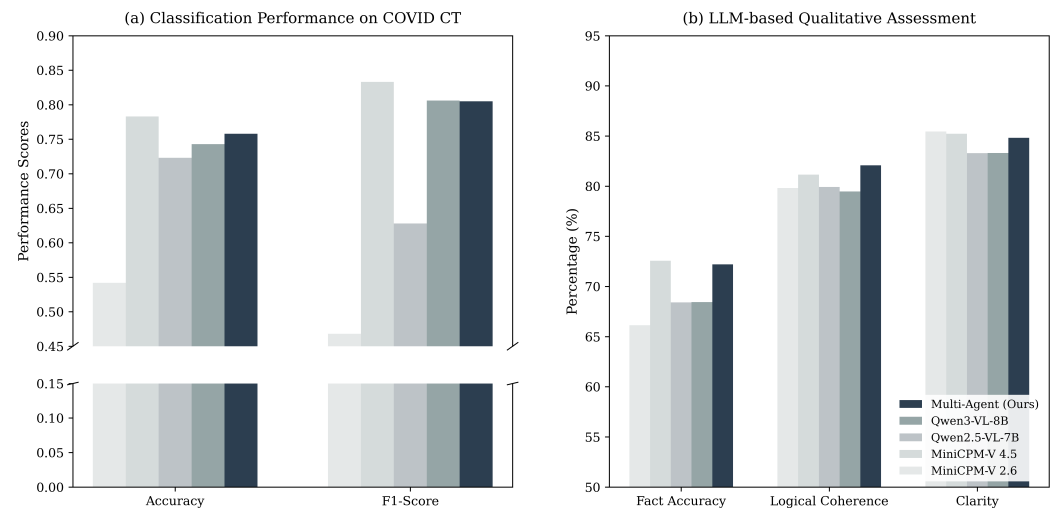


Figure 6. Performance evaluation of the proposed Multi-Agent framework on the Covid CT dataset. (a) Quantitative metric comparison across Accuracy, Precision, Recall, and F1-Score. A broken y-axis is employed to visualize the competitive range (0.5–0.9). (b) Exploratory judge scores for FA, LC, and CE; these are secondary model-based assessments rather than clinical measurements.

5. Discussion

This study examines a multi-agent medical reasoning framework based on temporally unrolled DAG executions. The answer-key metrics and exploratory judge scores warrant further study of structured collaboration, but they do not establish clinical safety or provide independent expert validation.

The framework combines role specialization, gated voting, and bounded refinement, but its relationship to prior agent work requires careful qualification. Zhang et al. [9] decompose embodied cooperation into modular perception, communication, and action roles; our study adopts the modular-role principle but applies it to diagnostic evidence and explicit voting rather than physical action planning. SMART-LLM [10] allocates robot subtasks according to agent capabilities; similarly, our GP agent assigns specialty responsibilities, but our routing objective is diagnostic consensus and uncertainty handling. Self-Refine [12] uses feedback from the same model to iteratively revise an output, whereas our rewrite transition is triggered by disagreement among role-specialized agents. CRITIC [13] grounds correction in external tools; our current verification is predominantly peer-based and therefore lacks CRITIC’s external evidence check. Thus, these studies motivate individual design components, but they do not directly validate medical effectiveness. Table 12 summarizes the similarities and differences.

The 81.26% Mainland value is judge-rated Factual Accuracy (FA), not classification accuracy; the latter is 0.801. The relatively close clarity scores across models may reflect a ceiling in the chosen automated rubric, but the shared-family evaluator prevents stronger conclusions about factual reliability or reasoning quality. The principal methodological contribution is therefore the explicit, auditable orchestration protocol rather than a claim of independently validated clinical superiority.

The results across textual diagnosis and multimodal visual question answering show that orchestration does not remove dependence on the underlying perceptual model. On ACRIMA, the Multi-Agent framework reports the same accuracy as one MiniCPM baseline (0.735) and remains 12.6 percentage points below Qwen3-VL-8B. This is evidence of a visual-backbone bottleneck rather than evidence that collaboration universally improves perception: agents cannot deliberate about image features that the visual encoder fails to extract. MiniCPM was retained as the primary vision component to preserve the originally executed resource configuration and to study orchestration under a compar-

actively lightweight local backbone; substituting Qwen3-VL-8B would change both the backbone and computational budget and would constitute a different experiment. A backbone-controlled study using Qwen3-VL-8B both alone and inside the same agent graph is therefore required before attributing multimodal gains to orchestration.

Table 12. Relationship between the proposed framework and the general agent studies discussed in Section 5.

Study	Relevant Finding or Mechanism	Similarity	Difference and Implication
Zhang et al. [9]	Modular cooperation among embodied agents	Both separate roles and coordinate their outputs	Embodied action planning versus diagnostic evidence aggregation; provides design motivation, not clinical validation
SMART-LLM [10]	Capability-aware multi-agent task allocation	Both assign tasks according to agent specialization	Robot subtasks versus medical responsibilities and voting
Self-Refine [12]	Iterative revision using model-generated feedback	Both use bounded feedback-driven refinement	Single-model self-feedback versus disagreement-triggered multi-agent rewriting
CRITIC [13]	Tool-interactive critique and correction	Both introduce an explicit verification step	External-tool grounding versus primarily peer-based verification; motivates future evidence-tool integration

Limitations and Analyses Not Supported by the Archived Data

The temporally unrolled state graph makes the workflow easier to audit, but the system is not ready for autonomous clinical use. Several requested analyses cannot be recovered from the archived artifacts; so, we report these gaps directly rather than estimate them after the fact.

The same-backbone standalone comparison controls only the text backbone; it is not a factorial ablation of classification, role allocation, voting, rewriting, and moderation. The causal contribution of each module therefore remains unmeasured. In particular, the experiment does not isolate the consensus module, and the Llama3-8B US run is incomplete; differences in the reported metrics therefore cannot be attributed to noise filtering, bias mitigation, or clinical reliability. The reported numerical comparisons describe the evaluated runs and should not be interpreted as clinical validation or evidence that the current decision threshold is optimal. The acyclicity argument establishes a topological property of each finite execution, not diagnostic correctness or safety, and this study did not collect a quantitative measure of path interpretability or explanation fidelity.

The final result files do not preserve per-cycle vote vectors, complete role lists, convergence or failure flags, or reasoning provenance. The supplied process workbook provides two illustrative traces for demonstrating the workflow; these examples are not dataset-wide telemetry. The consensus-polarization rate, dataset-level team-size distribution, rewrite frequency, and performance of failed cases therefore cannot be reconstructed from these artifacts. Under the $K = 3$ rule, a case that still lacks a strict majority at the limit is marked as failed and does not receive a moderator-generated diagnosis. In a clinical setting, such a case should be referred to a qualified clinician.

All roles used the same temperature, $\tau = 0.5$, and no role-specific sensitivity experiment was performed. Because sampling-based generation can produce different responses for the same input, run-to-run variation requires a dedicated repeated-generation analysis.

Such an analysis was outside the scope of the reported evaluation; confidence intervals and inter-run agreement are therefore not reported.

Qwen3-8B judge scores may share family-level stylistic biases with the Qwen2.5-7B backbone. Because no independent-family judge or blinded human-expert ratings were collected, judge–expert correlation and agreement cannot be reported, and these scores cannot support claims of clinical reliability.

The Llama3-8B US result of 0.045 is affected by invalid generations: 1052 of 1273 archived records lack a valid extracted option answer. Typical failures include repeating the question and producing text without a parseable choice; these records should not be described simply as blank model responses. We retain the reported value for transparency but do not use it to support architectural conclusions or treat it as a reliable estimate of model capability. No corrected rerun was available; so, a constrained-decoding rerun is required for a fair baseline comparison.

The recorded voting stage evaluates role-specific supporting and opposing assessments. Correlated model outputs can still produce information cascades, even without a single shared voting summary. Future experiments should record n_k , $v_i^{(k)}$, role identities, answer changes, correctness, convergence status, decoding settings, run identifiers, latency, and judge output at every round. They should also report module-wise ablations, path-fidelity measures, role-specific temperature tests, confidence intervals from repeated independent generations, consensus-polarization and non-convergence strata, and correlations with blinded expert ratings. Multi-agent coordination also increases inference latency and cannot recover visual features that a weaker vision backbone fails to detect.

The formalization through temporally unrolled DAG executions provides a basis for modeling diagnostic reasoning as a sequential process with explicit verification states. Whether this structure suppresses hallucination, polarization, or clinical error requires future backbone-controlled experiments with per-cycle logs and independent expert assessment.

6. Conclusions

We presented a multi-agent medical reasoning framework whose finite runs are represented as temporally unrolled DAGs. The workflow combines generated routing labels, multi-specialty role assignment, specialist discussion, strict-majority voting, bounded rewriting, conditional moderator summarization, and failure at the iteration limit. The experiments report answer-key performance on MedQA, ACRIMA, ADAM, and Covid CT, along with secondary Qwen3-8B judge scores. The same-backbone comparison is limited and cannot replace module-level ablation, and the ACRIMA results show that performance depends on the visual backbone. The framework is a research prototype, not an autonomous clinical system. Further validation requires per-cycle provenance, controlled ablations, repeated runs, role-specific decoding studies, and independent expert review.

Author Contributions: Conceptualization, X.Z. and X.N.; Methodology, X.Z., X.N. and Z.W.; Software, X.Z., Y.L. and Z.W.; Validation, X.Z. and Y.L.; Formal analysis, X.Z. and Y.L.; Writing—original draft, X.Z. and Y.L.; Writing—review & editing, X.Z., X.N. and Z.W.; Visualization, X.Z.; Supervision, X.N.; Project administration, X.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are derived from publicly available open-source datasets, including MedQA and OmniMedVQA. These datasets are freely accessible for research purposes.

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

Appendix A. Reproducibility Disclosures and Role-Provisioning Prompts

The implementation used the following prompt logic for classification and role provisioning (placeholders are shown in braces and whitespace is normalized):

Classification: “You are given a question: {question}. Select the applicable medical question type(s) and modality. Return a JSON object containing ‘disease_type’ and ‘modality_type’. Output the result directly without the thought process.”

Role provisioning: “Given the modality type {modality_type}, disease type(s) {disease_type}, and patient question {question}, identify the relevant specialist doctor(s), radiologic technologist(s), and other specialist(s) from the perspective of a General Practitioner. Assign tasks to each identified role, specifying the actions, tests, or examinations needed for diagnosis and treatment. Output the result directly without the thought process.”

Structured extraction: “Extract the relevant medical roles and tasks. One role may correspond to multiple tasks; combine its tasks into one responsibility statement. Return a JSON object containing ‘roles’, in which every item contains one ‘role’ and its corresponding ‘tasks’. At least one role and task are required.”

Specialist response: “You are the {role}. Your assigned responsibility is {tasks}. Review the original query {question} and provide a concise evidence-based assessment relevant to your responsibility. State the proposed answer and the evidence supporting or opposing it. Do not assume that another agent’s statement is correct.”

Voting: “For your assigned role, review the original query and the supporting and opposing assessments generated for that role. Return ‘binary_result=yes’ only if the role-specific assessment adequately supports the candidate conclusion; otherwise return ‘binary_result=no’. Do not resolve uncertainty in favor of agreement.”

Query rewriting: “Given the current query, role-tagged assessments, vote reasons, and unresolved disagreements, rewrite the query for the next round. Preserve all original clinical facts, identify the disputed point explicitly, and request the missing evidence needed to resolve it. Do not add patient facts. Return only the rewritten query.”

Moderator summarization: “Synthesize the original query and the complete role-tagged discussion history. Report the final answer, the supporting evidence, and any residual disagreement or uncertainty. Do not conceal contradictory evidence and do not invent clinical facts.”

The communication protocol is sequential and schema-constrained: classification and role provisioning produce structured JSON; each specialist receives the same original query plus its own responsibility; each role produces a role-specific pair of supporting and opposing assessments; the voting state serializes a binary gate result for that role; Rewrite receives the current query and unresolved disagreement; and the Moderator produces the final synthesis after routing. These prompt templates describe observable inputs, outputs, and transitions without relying on disclosure of hidden chain-of-thought.

Audit of the Supplied Process Workbook

We additionally examined two workflow examples from the supplied workflow workbook: one MedQA-style text question and one fundus-imaging question. These examples

illustrate the processing stages and are not presented as dataset-wide telemetry. Both contain seven observable stages: Question, Triage Agent, Disease Classification, Doctor Task Planning, Discussion, Voting, and Final Conclusion. The complete fields are reproduced below as dialogue boxes.

The following two records are complete workflow examples supplied for process illustration. They are not dataset-level telemetry or statistical samples.

Example A: Text Medical Question

Question

"A junior orthopaedic surgery resident is completing a carpal tunnel repair with the department chairman as the attending physician. During the case, the resident inadvertently cuts a flexor tendon. The tendon is repaired without complication. The attending tells the resident that the patient will do fine, and there is no need to report this minor complication that will not harm the patient, as he does not want to make the patient worry unnecessarily. He tells the resident to leave this complication out of the operative report. Which of the following is the correct next action for the resident to take?"options": {"A": "Disclose the error to the patient but leave it out of the operative report", "B": "Disclose the error to the patient and put it in the operative report", "C": "Tell the attending that he cannot fail to disclose this mistake", "D": "Report the physician to the ethics committee", "E": "Refuse to dictate the operative report"}

Triage Agent

I need to address the user's question. A resident performing carpal tunnel release inadvertently cut a flexor tendon, and the attending instructed the resident not to document the complication in the operative report. The task is to determine the resident's correct next action.

The question concerns ethical and legal duties in medical practice, particularly whether a medical error must be disclosed to the patient and how it should be documented in the operative report. Relevant sources may include professional ethics guidance, institutional policy, and applicable law.

The available medical knowledge retrieval tool can be used to find relevant guidance. Because the issue is ethical and legal rather than purely diagnostic, the search should focus on error disclosure and accurate operative documentation.

Disclosure of medical errors is generally connected to informed consent, transparency, and preservation of patient trust. Even when a complication is considered minor, accurate documentation remains important for subsequent care and legal accountability.

Among the options, C challenges the attending's attempt to conceal the error, whereas B both discloses and documents it. The applicable professional or institutional guidance should be checked because documentation requirements may apply regardless of the perceived severity of the complication.

Because the question concerns standards of medical practice, relevant ethics guidance and reporting requirements should be consulted, including guidance from professional organizations such as the American Medical Association. The attending's assessment that no harm will occur does not by itself remove the duty to consider disclosure.

The retrieval query should therefore target guidance on medical-error disclosure and standards for operative-report documentation.

Disease Classification

```
{
  "disease_type": "medical ethics and legal issue",
  "modality_type": "medical practice standards"
}
```

Doctor Task Planning

```
{
  "roles": [
    {
      "role": "Specialist Doctor (Medical Ethics and Legal Counsel)",
      "tasks": [
        "Evaluate the ethical and legal implications of the resident's actions.",
        "Determine whether disclosure of the error to the patient is required under medical ethics and legal standards.",
        "Advise on the appropriate handling of the operative report, including whether the complication should be documented."
      ]
    },
    {
      "role": "Specialist Doctor (Orthopaedic Surgery)",
      "tasks": [
        "Review the surgical procedure and the nature of the complication.",
        "Confirm whether the error is minor and non-harmful, as stated by the attending physician.",
        "Provide input on the standard of care and whether the attending physician's decision to omit the complication from the report is appropriate."
      ]
    },
    {
      "role": "Radiologic Technologist",
      "tasks": [
        "If imaging was involved in the procedure, review any radiologic findings related to the complication.",
        "Ensure that imaging reports are accurate and consistent with the clinical documentation."
      ]
    },
    {
      "role": "General Practitioner (Assigning Role)",
      "tasks": [
        "Coordinate with the specialist doctors to ensure ethical and legal compliance.",
        "Monitor the resident's actions to ensure adherence to medical practice standards.",
        "Provide guidance on reporting procedures and ethical responsibilities."
      ]
    }
  ]
}
```

Discussion

```
{'positive_answer':
,

**Assessment Steps**:
- **Initial Assessment**: The scenario involves a junior resident who
  inadvertently causes a minor complication during surgery. The attending
  physician advises against disclosing the error to the patient and omitting it
  from the operative report. This raises ethical concerns about transparency,
  documentation, and the resident's duty to the patient.
- **Diagnostic Studies (e.g., imaging, lab tests)**: Not applicable here,
  as the issue is about ethical and legal obligations rather than clinical
  diagnosis.
- **Additional Considerations**: The resident must balance professional
  hierarchy, legal/ethical responsibilities, and patient autonomy.
  Non-disclosure of even minor errors can lead to legal liability, breaches
  of trust, and compromised patient safety.
**Possible Answers**:
- **Answer B**: "Disclose the error to the patient and put it in the
  operative report."
*Reasoning*: The resident has a duty to disclose errors to patients and
  document them accurately in medical records. Omitting the error risks legal
  consequences (e.g., malpractice claims) and undermines patient trust. While
  the attending's instruction to suppress the error is unethical, the resident
  must prioritize transparency and accountability.
- **Answer C**: "Tell the attending that he cannot fail to disclose this
  mistake."
*Reasoning*: While the resident may feel compelled to challenge the
  attending's directive, this approach could escalate conflict without ensuring
  the error is properly addressed. Direct confrontation may not resolve the
  issue effectively.
- **Answer D**: "Report the physician to the ethics committee."
*Reasoning*: This is a valid step if the attending's behavior constitutes a
  violation of ethical or regulatory standards. However, it is a more severe
  action and should be considered only if the resident cannot ensure proper
  documentation and disclosure through other means.
- **Answer E**: "Refuse to dictate the operative report."
*Reasoning*: Refusing to include the error in the report could prevent its
  documentation, but it does not address the need for patient disclosure. The
  resident's role includes ensuring accurate records, not merely complying with
  the attending's instructions.
**Conclusion**:
The resident's primary responsibility is to prioritize patient safety and
transparency. While the attending's directive to suppress the error is
ethically problematic, the resident must take direct action to disclose
the error to the patient and document it in the operative report. This
aligns with legal and ethical standards, ensuring accountability and trust.
**Answer B** is the correct next action.',
'negative_answer':

**Assessment Steps (Evaluation Steps)**:
- **Initial Assessment**: [Overly brief, omit critical elements] The patient
  had a carpal tunnel repair, and the resident accidentally cut a flexor
  tendon. The attending said it's a "minor complication" and told the resident
  to hide it. No mention of patient history, physical exam, or whether the
  patient had pre-existing conditions like diabetes or neuropathy.
```

- ****Diagnostic Studies (e.g., imaging, lab tests)**:** [Omit necessary tests, list irrelevant ones] No imaging or labs were mentioned, but the resident was advised to "check the patient's hand strength" (a vague, non-specific recommendation).

- ****Additional Considerations**:** [Ignore important risk factors] No mention of the risk of chronic pain, loss of function, or the ethical obligation to document errors. The resident's training level was not addressed.

****Possible Answers (Possible Responses/Recommendations)**:**

- ****Answer 1**:** "Disclose the error to the patient but leave it out of the operative report"

Reasoning: [Irrelevant, illogical] "The patient won't notice, and the report is just paperwork. The attending is the boss, so follow his orders."

- ****Answer 2**:** "Tell the attending that he cannot fail to disclose this mistake"

Reasoning: [Speculative, one-sided] "The attending is a senior doctor, so he knows better. You're just a resident; you can't challenge him."

- ****Answer 3**:** "Refuse to dictate the operative report"

Reasoning: [Vague, emotionally charged] "If you lie in the report, you're risking your career. But if you follow the attending, you're a good team player. It's a tough call."

****Conclusion**:**

[Overly general, incorrect, alarmist] "The attending is right—patients don't need to know about minor mistakes. Just keep quiet and move on. If you report it, you'll be seen as a complainer. The system is broken, so don't rock the boat."

--

This response intentionally omits critical ethical principles (e.g., informed consent, documentation), uses vague reasoning, and provides misleading advice. It also fails to address the legal and professional consequences of hiding medical errors.'

Voting

binary_result='yes'

Final Conclusion

****Summary of Discussions**:**

The meeting focused on a medical ethics dilemma involving a junior orthopaedic surgery resident who inadvertently cut a flexor tendon during a carpal tunnel repair. The attending physician instructed the resident to omit the complication from the operative report to avoid causing patient worry. The discussion centered on the resident's ethical obligations regarding transparency, documentation, and patient autonomy.

****Possible Answers**:**

- ****Answer 1**:** Disclose the error to the patient but leave it out of the operative report.

***Explanation*:** Contradictory, as disclosure requires documentation.

- ****Answer 2**:** Disclose the error to the patient and put it in the operative report.

***Explanation*:** Aligns with ethical duties to transparency and accurate documentation.

- ****Answer 3**:** Tell the attending that he cannot fail to disclose this mistake.

***Explanation*:** Emphasizes the resident's duty to the patient, but may not address the need for documentation.

****Agreements**:**

- The attending's instruction to omit the complication was deemed ethically problematic, as transparency is critical for informed consent.
- The resident's duty to document accurately and inform the patient was universally acknowledged.

****Disagreements**:**

- Some participants debated whether the attending's intent to avoid patient worry justified withholding information.
- There was debate over whether the resident should override the attending's instructions or first address the issue with the attending.

****Conclusions**:**

- The resident's ethical obligation to the patient supersedes the attending's directive.
- The correct next action is to ****disclose the error to the patient and include it in the operative report**** (Option B), ensuring transparency and accurate documentation. This aligns with medical ethics principles of patient autonomy and informed consent.

Example B: Medical Image Question

Question

question: "What imaging technique was employed to obtain this picture?",
 "option_A": "PET scan", "option_B": "CT scan", "option_C": "Blood
 test", "option_D": "Fundus imaging",

Triage Agent

I'm sorry, but the question provided does not include an actual image or specific details about what is being asked regarding imaging techniques. To accurately answer your question and determine which tool to use, I need more information such as a description of the picture (e.g., whether it shows internal organs, blood vessels, etc.) or context clues that indicate the type of medical test depicted.

If you can provide additional details or describe the characteristics of the "picture" mentioned in your question, I will be able to better assist you by identifying the correct imaging technique among the options: PET scan, CT scan, Blood test, or Fundus imaging.

</think>

To solve this, we analyze the content of the image and match it to common medical imaging tools:

1. ****Extract Information from the Image**:** The image shows structures like retinal blood vessels and the optic disc-features typically observed in ophthalmology.
2. ****Analyze Options**:**
 - ***PET scan*:** Detects metabolic activity; used for conditions like cancer or neurological disorders. Does not show detailed retinal anatomy.
 - ***CT scan*:** Produces cross-sectional images of body structures using X-rays; focuses on bones, soft tissues outside the eye, or brain. Not specialized for ocular fundus detail.
 - ***Blood test*:** A laboratory analysis of blood samples, unrelated to visualizing anatomical structures via imaging.

- ***Fundus imaging***: Specifically captures images of the interior surface of the eye, including the retina, optic disc, and blood vessels. This matches the features seen in the image.

Based on the presence of retinal structures, the most appropriate imaging technique is fundus imaging.

boxed answer: {Fundus imaging}

Disease Classification

```
{
  "disease_type": "Ophthalmological",
  "modality_type": "Fundus imaging"
}
```

Doctor Task Planning

```
{
  "roles": [
    {
      "role": "Specialist Doctor (Ophthalmologist)",
      "tasks": [
        "Assess Patient's Symptoms and Medical History",
        "Interpret Fundus Imaging Findings",
        "Determine Further Diagnostic Needs"
      ]
    },
    {
      "role": "Radiologic Technologist (Fundus Imaging Specialist)",
      "tasks": [
        "Prepare for Fundus Imaging",
        "Perform Imaging",
        "Document and Share Results"
      ]
    },
    {
      "role": "General Practitioner",
      "tasks": [
        "Coordinate Care",
        "Refer to Specialists",
        "Monitor Progress"
      ]
    }
  ],
  {
    "role": "Patient",
    "tasks": [
      "Provide Image Details"
    ]
  }
}
```

Discussion

{'positive_answer': '**Assessment Steps**:

- Initial Assessment: The image provided shows a detailed view of the retina and optic disc, which are structures located in the back of the eye. This type of imaging is typically used to examine these areas for various ocular conditions.
- Diagnostic Studies (e.g., imaging, lab tests): Given the visibility of retinal blood vessels and other internal features, this suggests an ophthalmological diagnostic tool rather than general body scans or laboratory results.
- Additional Considerations: Techniques like PET scan and CT scan focus on broader anatomical regions such as organs and bones; they do not provide clear images of delicate intraocular structures.

Possible Answers:

- Answer1: Fundus imaging

Reasoning: The picture clearly displays characteristics specific to fundus imaging, including visible retinal vasculature and the optic nerve head. These details align with what one would expect from a direct examination using specialized equipment designed to capture close-up views of the posterior segment of the eye.

- Answer2: PET scan

Reasoning: A PET scan provides functional information about metabolic activity throughout the body but does not offer structural detail at the microscopic level shown here.

- Answer3: CT scan

Reasoning: While CT scanning can show some aspects of orbital anatomy indirectly through cross-sectional slices, it lacks the resolution necessary to visualize fine vascular patterns within the retina directly without additional contrast enhancement procedures that wouldn't be indicated by just viewing the normal structure depicted above.

- Answer4: Blood test

Reasoning: There's no indication whatsoever related to biochemical analysis via plasma or serum samples since nothing regarding fluid composition appears anywhere near any optical elements captured under such high magnification levels required when dealing specifically inside someone's eyeball cavity itself.

Conclusion: Based on visual evidence presented alongside understanding typical applications across different medical imaging modalities, "Fundus imaging" stands out as most consistent choice among options listed considering context provided by question itself along with inherent properties associated uniquely towards capturing internal ocular structures accurately without resorting invasive techniques nor relying upon indirect inferential data gathered elsewhere within human physiology aside from where needed primarily focused around diagnosing problems potentially arising within visual pathway itself starting right up against surface closest contact point between external world light entering pupil opening versus sensitive neural tissue responsible ultimately transmitting signals onward toward brain where conscious perception occurs.', 'negative_answer':

"**Assessment Steps (Evaluation Steps)**:

- Initial Assessment: Look at the picture closely, it's orange and has blood vessels.

Possible Answers (Possible Responses/Recommendations):

- Answer 1: It was a CT scan. Reasoning: Because we see veins clearly.
- Answer 2: Blood test results look normal so maybe this is fine? Reasoning:

I just checked the lab report quickly!

- Answer 3: Maybe an MRI would show more details! Reasoning: I heard MRIs are better for soft tissues but no one said that here.

****Conclusion**:** The image looks like something from inside your eye-probably fundus imaging-but let's not worry too much unless you have vision problems or weird lights in front of eyes!

****Note**:** This response intentionally ignores proper clinical context (no patient history, exam findings), confuses unrelated tests with ophthalmology exams, uses vague reasoning, and omits critical steps-all to mimic poor medical communication."}

Voting

binary_result='yes'

Final Conclusion

****Summary of Discussions****

****Question**:** "What imaging technique was employed to obtain this picture?"

****Options**:** PET scan, CT scan, Blood test, Fundus imaging.

--

**Agreements**

1. ****Consensus on Fundus Imaging**:**

- All ****positive assessments**** (structured, medical reasoning) agreed that ****Fundus imaging**** is the correct answer.
- Reasoning: The image shows detailed retinal structures (e.g., optic disc, blood vessels), which are specifically visualized using fundus imaging (e.g., fundoscopy, digital fundus photography).
- ****PET scan****, ****CT scan****, and ****Blood test**** were consistently ruled out:
- ****PET scan**** focuses on metabolic activity, not anatomical detail.
- ****CT scan**** lacks the resolution to capture fine retinal vasculature.
- ****Blood test**** produces biochemical data, not images.

2. ****Clinical Context**:**

- The image's focus on the posterior eye (retina, optic disc) aligns with ophthalmological diagnostic tools, not general imaging or lab tests.

--

**Disagreements**

1. ****Incorrect Reasoning in Negative Answers**:**

- ****PET scan**:** Some negative responses incorrectly associated the image with PET scans due to perceived "glowing" or "colorful" features, ignoring its metabolic focus.
- ****CT scan**:** Others linked the image to CT scans by comparing vascular patterns to "cross-like" structures, despite CT's inability to resolve retinal details.
- ****Blood test**:** One response humorously suggested the image was a "blood test result," conflating biochemical analysis with visual imaging.

2. ****Lack of Medical Accuracy**:**

- Negative answers often omitted critical clinical context (e.g., no patient history, exam findings) and relied on vague or humorous reasoning (e.g., "orange picture with veins," "phone camera").

--

**Conclusions**

- ****Final Answer**:** ****Fundus imaging**** is the correct technique, as it

directly captures retinal and optic disc structures with high resolution.

- **Key Takeaway**: The discussion emphasized the importance of matching imaging modality to the anatomical focus (retina) and avoiding assumptions based on superficial visual cues (e.g., color, lines).
- **Recommendation**: For similar cases, prioritize specialized ophthalmic imaging tools over general scans or lab tests when examining intraocular structures.

--

Summary: The discussions reached a clear consensus on **Fundus imaging** as the correct answer, while highlighting the pitfalls of misinterpreting imaging characteristics and relying on non-medical reasoning.

This audit confirms that classification, dynamic role generation, discussion aggregation, a binary gate, and moderator synthesis occur in the recorded examples. It does not establish a dataset-wide team-size distribution, per-agent votes, rewrite frequency, convergence rate, latency, or run-to-run variability. The two observed team sizes (both four entries) are illustrative and must not be reported as an average, minimum, or maximum for the benchmark datasets.

The role list is generated at inference time, but the archived CSV files retain only the question, options or ground truth, and final answer. They do not contain the generated roles. The mean, minimum, and maximum team sizes therefore cannot be recovered without rerunning the experiment. Future runs should store the role list and team size for every case.

References

1. Makary, M.A.; Daniel, M. Medical error—The third leading cause of death in the US. *BMJ* **2016**, *353*, i2139. [[CrossRef](#)] [[PubMed](#)]
2. Singh, H.; Giardina, T.D.; Meyer, A.N.; Forjuoh, S.N.; Reis, M.D.; Thomas, E.J. Types and origins of diagnostic errors in primary care settings. *JAMA Intern. Med.* **2013**, *173*, 418–425. [[CrossRef](#)] [[PubMed](#)]
3. Rajpurkar, P.; Chen, E.; Banerjee, O.; Topol, E.J. AI in health and medicine. *Nat. Med.* **2022**, *28*, 31–38. [[CrossRef](#)] [[PubMed](#)]
4. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
5. 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](#)] [[PubMed](#)]
6. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 248. [[CrossRef](#)]
7. Nundy, S.; Cooper, L.A.; Mate, K.S. The quintuple aim for health care improvement: A new imperative to advance health equity. *JAMA* **2022**, *327*, 521–522. [[CrossRef](#)] [[PubMed](#)]
8. Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; Ghanem, B. Camel: Communicative agents for “mind” exploration of large language model society. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 51991–52008. [[CrossRef](#)]
9. Zhang, H.; Du, W.; Shan, J.; Zhou, Q.; Du, Y.; Tenenbaum, J.B.; Shu, T.; Gan, C. Building cooperative embodied agents modularly with large language models. *arXiv* **2023**, arXiv:2307.02485.
10. Kannan, S.S.; Venkatesh, V.L.; Min, B.C. Smart-llm: Smart multi-agent robot task planning using large language models. In *Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: Piscataway, NJ, USA, 2024; pp. 12140–12147.
11. Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv* **2023**, arXiv:2305.16291.
12. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegrefe, S.; Alon, U.; Dziri, N.; Prabhunoye, S.; Yang, Y.; et al. Self-refine: Iterative refinement with self-feedback. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46534–46594. [[CrossRef](#)]
13. Gou, Z.; Shao, Z.; Gong, Y.; Shen, Y.; Yang, Y.; Duan, N.; Chen, W. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv* **2023**, arXiv:2305.11738.
14. Zhang, W.; Tang, K.; Wu, H.; Wang, M.; Shen, Y.; Hou, G.; Tan, Z.; Li, P.; Zhuang, Y.; Lu, W. Agent-pro: Learning to evolve via policy-level reflection and optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 5348–5375.

15. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv* **2023**, arXiv:2303.13375.
16. Chen, Z.; Cano, A.H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv* **2023**, arXiv:2311.16079.
17. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **2022**, *377*, e070904. [[CrossRef](#)] [[PubMed](#)]
18. Tang, X.; Zou, A.; Zhang, Z.; Li, Z.; Zhao, Y.; Zhang, X.; Cohan, A.; Gerstein, M. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL; Association for Computational Linguistics*: Kerrville, TX, USA, 2024; pp. 599–621. [[CrossRef](#)]
19. Kim, Y.; Park, C.; Jeong, H.; Chan, Y.S.; Xu, X.; McDuff, D.; Lee, H.; Ghassemi, M.; Breazeal, C.; Park, H.W. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. In *Proceedings of the Advances in Neural Information Processing Systems; Neural Information Processing Systems Foundation, Inc.*: San Diego, CA, USA, 2024; Volume 37. [[CrossRef](#)]
20. Li, B.; Yan, T.; Pan, Y.; Luo, J.; Ji, R.; Ding, J.; Xu, Z.; Liu, S.; Dong, H.; Lin, Z.; et al. MMedAgent: Learning to Use Medical Tools with Multi-modal Agent. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; Association for Computational Linguistics*: Kerrville, TX, USA, 2024. [[CrossRef](#)]
21. Chen, X.; Yi, H.; You, M.; Liu, W.; Wang, L.; Li, H.; Zhang, X.; Guo, Y.; Fan, L.; Chen, G.; et al. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digit. Med.* **2025**, *8*, 159. [[CrossRef](#)] [[PubMed](#)]
22. Sutton, R.T.; Pincock, D.; Baumgart, D.C.; Sadowski, D.C.; Fedorak, R.N.; Kroeker, K.I. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digit. Med.* **2020**, *3*, 17. [[CrossRef](#)] [[PubMed](#)]
23. Rajpurkar, P.; Irvin, J.; Ball, R.L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P.; et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **2018**, *15*, e1002686. [[CrossRef](#)] [[PubMed](#)]
24. Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J.; Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature* **2023**, *616*, 259–265. [[CrossRef](#)] [[PubMed](#)]
25. 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. *arXiv* **2020**, arXiv:2009.13081.
26. Hu, Y.; Li, T.; Lu, Q.; Shao, W.; He, J.; Qiao, Y.; Luo, P. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; IEEE*: Piscataway, NJ, USA, 2024; pp. 22170–22183.

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.