CURE: A Multimodal Benchmark for Clinical Understanding and Retrieval Evaluation

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) hold considerable promise for clinical diagnosis, a task that inherently couples the interpretation of complex visual and textual inputs with the consultation of authoritative medical literature. However, existing medical benchmarks evaluate MLLMs almost exclusively in end-to-end answering scenarios, without separating the contributions of perception, retrieval, and reasoning. This makes it impossible to determine whether an incorrect diagnosis stems from deficient multimodal understanding, unsuccessful evidence retrieval, or an inability to reason over evidence that was in fact available. We introduce the Clinical Understanding and Retrieval Evaluation (CURE) benchmark, comprising 500 multimodal clinical cases each paired with physician-cited reference literature, which evaluates models under controlled evidence conditions designed to reveal where performance degrades across these stages. Across these conditions, we find that the dominant performance gap is consistently localized to the input-to-evidence step rather than the evidence-to-diagnosis step: a single model can achieve 73.4% accuracy on differential diagnosis when supplied with physician-cited references yet drops to 25.4% under independent retrieval, and a parallel gap emerges when raw images replace textualized imaging findings. Critically, supplying the correct findings or the correct references largely recovers performance, indicating that the bottleneck of current clinical MLLMs lies in converting multimodal clinical input into precise evidence, not in downstream clinical reasoning. CURE will be made publicly available after acceptance.

1 Introduction

Multimodal large language models (MLLMs) have recently shown impressive capabilities in jointly modeling vision and language, enabling progress in tasks that require grounded interpretation across images and narratives (Li et al., 2023; Sellergren et al., 2025; Huang et al., 2025; Qin et al., 2026). In medicine, this progress is especially compelling: clinical diagnosis rarely depends on a single modality. Clinicians synthesize heterogeneous cues (e.g., imaging patterns, temporality, comorbidities, laboratory trends) to form differential diagnoses and decide next steps (Huang et al., 2020; Zhang et al., 2025). Critically, when cases are ambiguous or high-stakes, clinicians often consult authoritative diagnostic literature, such as textbooks, guidelines, or peer-reviewed case reports, to justify interpretations, refine differentials, and support downstream decisions such as treatment planning and prognosis estimation (Del Fiol et al., 2014). As a result, *tracing the supporting evidence* is not merely an auxiliary step, but a core component of reliable clinical reasoning.

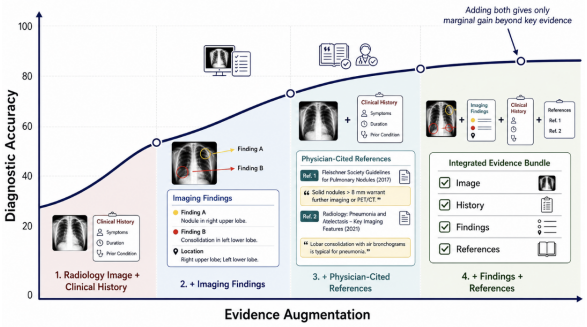

Figure 1: Diagnostic accuracy as clinical evidence is progressively completed. Supplying imaging findings and physician-cited references each yields large gains, whereas adding the final evidence component produces a negligible increase. The dominant performance gap thus lies in the input-to-evidence step, not in evidence-to-diagnosis reasoning.

This perspective has motivated a growing body of work that augments MLLMs with external knowledge through retrieval-augmented generation (RAG) (Lewis et al., 2020; Xiong et al., 2024) and, more recently, agent-based frameworks (Team

et al., 2025; Xu and Peng, 2025) that iteratively search, read, and refine. Despite rapid methodological iteration, evaluating these systems presents a dual challenge. On the *system level*, existing agent pipelines rely on heterogeneous search corpora (e.g., web sources, mixed medical repositories, proprietary databases) and entangled implementation choices, making results hard to interpret and reproduce across settings. On the *benchmark level*, established medical datasets such as MedXpertQA (Zuo et al., 2025) and MMMU (Yue et al., 2024) have advanced MLLM evaluation across visual question answering and multimodal reasoning, but they primarily assess end-to-end answering in isolation, scoring models solely on their final responses to given inputs. They lack a systematic testbed for evidence-grounded retrieval against a unified, authoritative reference set, leaving no principled way to determine whether an incorrect diagnosis stems from retrieval failure or reasoning failure. This leaves a key gap: developers lack an objective and clinically faithful way to *identify* whether an incorrect answer arises from deficient multimodal understanding, retrieval failure, or inability to leverage retrieved evidence.

To address this gap, we introduce the **Clinical Understanding and Retrieval Evaluation (CURE)**, a benchmark designed to evaluate MLLMs under a clinically motivated setting where models are expected to (i) interpret multimodal clinical inputs, (ii) retrieve relevant evidence from a unified, authoritative diagnostic literature source, and (iii) generate diagnoses that are both accurate and evidence-grounded. CURE is built to explicitly decouple model capability from retrieval behavior by providing standardized reference content and evaluation protocols that isolate where systems succeed or fail.

Using CURE, we benchmark state-of-the-art MLLMs which reveals a striking contrast: the same model can achieve up to 73.4% accuracy on open-ended diagnosis when supplied with physician reference evidence yet drops to 25.4% when tasked with independent retrieval. This massive performance gap highlights that the critical bottleneck in current clinical AI is not downstream reasoning but rather the ability to effectively translate multimodal visual cues into precise literature retrieval. In summary, our contributions are threefold:

- We introduce CURE, an evidence-grounded diagnostic benchmark of 500 multimodal clinical cases paired with physician-cited PubMed references, enabling controlled comparison across four evidence paradigms.
- We benchmark 16 MLLMs spanning proprietary and open-weight families and show that the dominant bottleneck is evidence retrieval rather than clinical reasoning: the same model that scores 73.4% with oracle references drops to 25.4% under independent retrieval.
- We provide a fine-grained analysis of failure modes, including faithfulness as a double-edged sword under noisy retrieval and the gap between image-derived and text-derived clinical cues, offering actionable directions for improving multimodal clinical AI.

2 Related Work

Recent medical AI benchmarks have evolved from perception-oriented question answering toward multimodal, retrieval-grounded, and increasingly process-aware evaluation of clinical reasoning.

2.1 Medical Multimodal Clinical Evaluation

Early multimodal medical benchmarks focused primarily on visual understanding and short-form clinical question answering. Representative datasets such as VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020) established foundational settings for medical visual question answering by combining clinical images with concise QA tasks. In parallel, text-centered medical reasoning benchmarks, including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022) and PubMedQA (Jin et al., 2019), advanced the evaluation of medical knowledge and reasoning abilities in language models. More recent benchmarks have moved toward multimodal clinical reasoning at the expert-level. Datasets such as MMMU-Medical (Yue et al., 2024), DiagnosisArena (Zhu et al., 2025), and MedXpertQA (Zuo et al., 2025) incorporate more complex diagnostic scenarios requiring integration of visual and textual evidence. However, these benchmarks evaluate models end-to-end without controlling the evidence available at inference time, making it impossible to attribute errors to perception, retrieval, or reasoning individually. CURE addresses this gap by introducing controlled evidence conditions that vary the information each model receives.

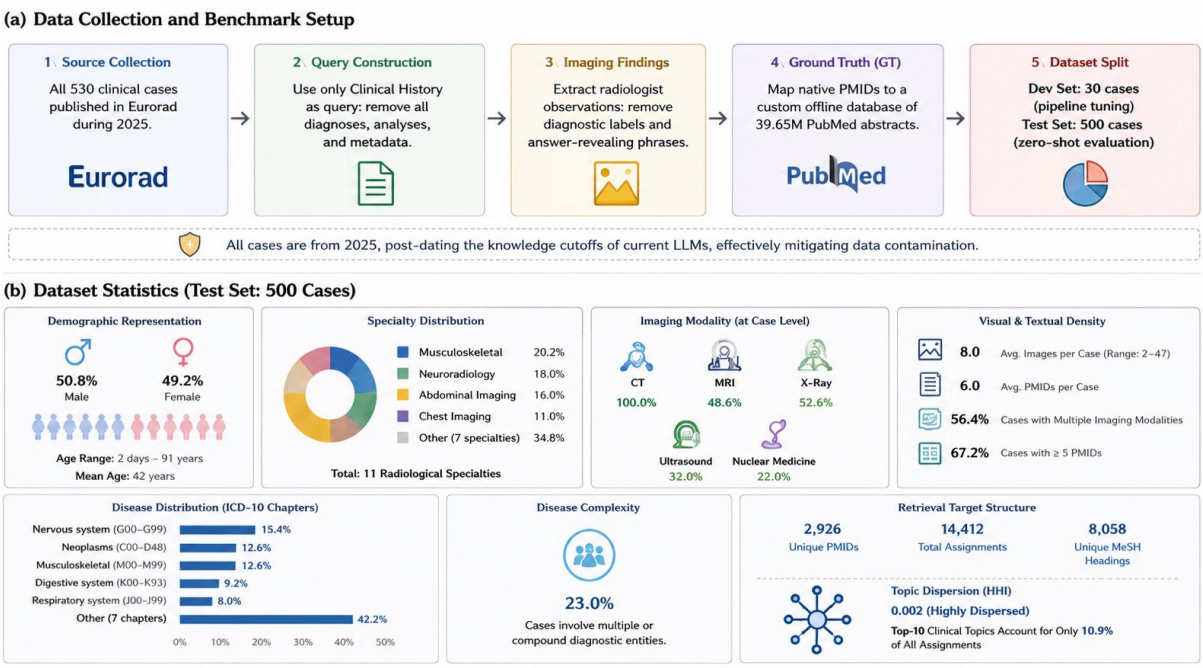

Figure 2: CURE benchmark. Cases are paired with PubMed reference evidence, and MLLMs are evaluated under different retrieval paradigms to isolate multimodal understanding from evidence-grounded diagnosis performance.

2.2 Retrieval-Augmented Medical QA

Benchmarks and systems such as MIRAGE (Dongre et al.), MedRAG (Xiong et al., 2024), and BiomedRAG (Li et al., 2025) demonstrate the value of incorporating external biomedical knowledge into medical reasoning pipelines. Related efforts based on MedCPT (Jin et al., 2023) retrieval models and physician-oriented retrieval systems further highlight the importance of evidence-grounded generation in medical AI. Recent agent-based and workflow-oriented systems, including Almanac (Zakka et al., 2024), MedAgents (Tang et al., 2024), and MiroFlow (Su et al., 2026), extend retrieval-augmented reasoning toward iterative and tool-augmented clinical workflows. Meanwhile, general-domain RAG benchmarks such as RGB (Chen et al., 2024), CRAG (Yang et al., 2024), and FRAMES (Krishna et al., 2025) increasingly emphasize separate evaluation of retrieval quality and downstream reasoning behavior. However, existing medical RAG benchmarks (e.g., MedRAG) focus on text-only question answering without multimodal inputs, while clinical diagnostic benchmarks (e.g., DiagnosisArena) do not incorporate controlled retrieval conditions. CURE bridges these two lines of work by pairing multimodal clinical cases with physician-cited references and evaluating models under standardized evidence paradigms.

3 The CURE Benchmark

We construct the **CURE** benchmark, which transitions the evaluation paradigm from simple visual question-answering to evidence-based diagnostic reasoning under controlled evidence conditions, bridging the gap between semantic matching and authoritative clinical inference.

3.1 Data Sourcing

We construct CURE from clinical cases published in Eurorad[1] by systematically collecting all cases released during 2025. Each Eurorad case includes anonymized radiological images together with accompanying educational reports. To simulate a realistic diagnostic workflow, we isolate only the information available at the initial interpretation stage and exclude all retrospective explanations or answer-revealing metadata. Images are retained in their original formats and native resolutions to preserve clinically relevant visual detail.

All collected cases were published in 2025, which post-dates the training cutoffs of current frontier multimodal language models. This temporal separation substantially reduces the likelihood of benchmark contamination through memorization of public benchmark instances or associated teaching materials. The benchmark is released under the

[1] https://www.eurorad.org/

same non-commercial educational usage principles as the original Eurorad content.

3.2 Annotation Protocol

Clinical History Extraction We extracted the Clinical History section from each Eurorad report as the primary textual query. To preserve the original clinical language distribution, we applied a deterministic preprocessing pipeline without paraphrasing or LLM-based rewriting. Specifically, we removed explicit diagnostic statements, retrospective commentary, and formatting artifacts, while retaining pre-diagnosis information, demographics, and symptoms. All extraction scripts were finalized on the development split. Two independent annotators reviewed the outputs, with disagreements resolved by a senior reviewer.

Imaging Findings Sanitization We additionally utilize the radiologist-authored Imaging Findings as an optional auxiliary modality. Unlike approaches using LLM-generated summaries, CURE preserves authentic physician-written observations. To prevent diagnostic leakage, we applied a lightweight sanitization procedure to eliminate direct mentions of the target diagnosis, definitive impressions, and retrospective explanations. The underlying radiological observations were retained without paraphrasing, preventing models from exploiting direct diagnostic cues while maintaining clinical authenticity.

PMID Extraction and Evidence Mapping CURE defines ground-truth evidence using physician-cited biomedical literature. We extracted all PubMed IDs (PMIDs) referenced within each report and deterministically mapped them against a locally indexed database containing 39.65 million PubMed abstracts. Only successfully resolved PMIDs were retained as valid retrieval targets, ensuring systems are evaluated on recovering the exact scientific evidence cited by human radiologists.

3.3 Quality Control

We specifically sourced cases from 2025 Eurorad publications to align the benchmark with contemporary clinical practice. To prevent the task from being solved through mere memorization or textual biases, we evaluated lightweight text-only baselines on the multiple-choice diagnosis task using only the sanitized clinical history. While these language-only models performed above random chance, they significantly underperformed compared to multimodal systems (see Table 4, History-only column). This confirms that successful performance necessitates genuine multimodal reasoning by meaningfully integrating imaging evidence and retrieval capabilities, rather than relying on superficial textual shortcuts.

3.4 Task Formulation

We evaluate end-to-end diagnostic reasoning under two complementary settings.

Differential Diagnosis Selection Models receive the multimodal query Q alongside 4–5 candidate diagnoses. Crucially, these candidates are directly extracted from the physician-annotated differential diagnoses natively present in the original Eurorad report, ensuring high clinical authenticity. GPT-5.2 is utilized strictly to standardize the terminology across the options; no diagnostic content is generated or altered by this step. Full prompting templates are provided in the appendix.

Open-ended Diagnosis Prediction Given solely Q, models generate a ranked list of diagnoses without predefined options, evaluated via Hit@1 and Hit@3. To accommodate synonymous clinical terminology, an LLM-as-a-judge ensemble (DeepSeek-V4, GPT-OSS-120B, and GLM-4.6) assesses semantic equivalence between predictions and ground-truth labels. This automated protocol demonstrates strong agreement with expert judgment ($\kappa = 0.86$) on a blinded 50-case subset.

4 Benchmark Characterization

We summarize key distributional properties of CURE to contextualize the experimental results that follow.

Demographic Representation. CURE comprises 500 clinical cases with a balanced gender distribution (50.8% male, 49.2% female). Patient ages span from 2 days to 91 years (mean: 42), including explicit non-numeric developmental stages (e.g., newborn, fetus) to ensure broad pediatric and adult coverage.

Specialty and Modality Coverage. The dataset spans 11 radiological specialties, predominantly musculoskeletal (20.2%), neuroradiology (18.0%), and abdominal imaging (16.0%). CURE is inherently multimodal: at the case level, CT, MRI, and ultrasound appear in 55.6%, 52.6%, and 22.0% of cases, respectively.

Visual and Textual Information Density. Cases average 8.0 validated medical images (range: 2–47) and 6.0 unique physician-vetted PubMed citations (PMIDs). Notably, 56.4% of cases feature multiple imaging modalities, and 67.2% cite $\geq$5 PMIDs, reflecting highly complex and context-rich examinations.

Disease Distribution and Complexity. A ICD-10 mapping reveals a challenging long-tail distribution across 12 chapters, led by nervous system (15.4%), neoplasms (12.6%), and musculoskeletal diseases (12.6%), with no single chapter exceeding 16%. Furthermore, 23.0% of cases involve multiple or compound diagnostic entities, necessitating ambiguity-aware evaluation.

Retrieval Target Structure. Evidence is rigorously grounded in 2,926 unique PMIDs, yielding 14,412 assignments across 8,058 unique clinical MeSH (Medical Subject Headings) headings. This topic distribution is highly dispersed (Herfindahl–Hirschman Index, HHI = 0.002, where values near zero indicate maximal diversity), with the top-10 clinical topics accounting for only 10.9% of all assignments.

Question-Type Distribution. Tags categorize the clinical reasoning tasks into identification (62.6%), differentiation (35.4%), and characterization (2.0%), demonstrating that CURE requires complex differential diagnosis beyond basic anomaly detection.

5 Experiments

5.1 Experimental Setup

We benchmark a diverse suite of models spanning three categories: (1) *SOTA & large-scale models* including GPT-5.5 (Singh et al., 2025), Gemini-3-Pro, Qwen3.5-397B, GLM-4.6V (Team, 2025), and Kimi-K2.6 (Team et al., 2026); (2) *general open-source models at comparable scales* including the Qwen3-VL (Bai et al., 2025) family (4B–32B) and Qwen2.5-VL-32B; and (3) *medical-specialized models* including Hulu-Med-32B, Lingshu-32B, HuatuoGPT-Vision-7B, and MedGemma (Sellergren et al., 2025) (4B/27B). All open-weight models are deployed via Sglang (Zheng et al., 2024). Evaluation is conducted on the 500-case test set using both open-ended diagnosis (Hit@1, Hit@3) and multiple-choice accuracy (MCQ).

We design two complementary experimental axes. The first isolates the impact of **evidence retrieval** by fixing the input modality (clinical history + images) and varying the external context:

- **Base:** No external evidence; measures intrinsic diagnostic capability.
- **Physician Reference (PR):** Oracle context from physician-cited abstracts, establishing an upper bound.
- **RAG:** Ensemble of bge-large-en-v1.5 (Xiao et al., 2024) and MedCPT (Jin et al., 2023) with top-5 retrieval, PMID-level deduplication, and re-ranking.
- **Agent Retrieval (AR):** MiroFlow (Team et al., 2025) agentic pipeline with unconstrained real-time web search.

The second axis isolates the impact of **input modalities** by fixing the retrieval condition (no external evidence) and progressively enriching the input: (1) clinical history only, (2) + raw images, (3) + physician-written imaging findings, and (4) + both images and findings. This design disentangles the contribution of visual understanding from text-based clinical reasoning.

5.2 The Impact of Evidence Retrieval Paradigms

Table 1 demonstrates that the effect of external evidence on multimodal clinical diagnostic reasoning is highly contingent on evidence quality. When physician reference context is provided, all models exhibit substantial gains in open-ended diagnosis (e.g., GLM-4.6V Hit@1: 0.216→1.232; Qwen3.5-397B Hit@1: 0.264→1.352), confirming that current systems already possess strong potential for medical logical deduction when given appropriate evidence. More strikingly, MCQ accuracy under PR reaches 93.8–95.0% for models with reliable format compliance (GPT-5.5, Gemini 3 Pro, Qwen3.5-397B, GLM-4.6V), suggesting that the major bottleneck in real-world deployment is not the models' reasoning capability per se, but the limited accessibility of high-quality, diagnostic-level pathological associations.

Medical-specialized models demonstrate superior evidence utilization. Under the PR condition, Hulu-Med-32B achieves 88.2% MCQ and Hit@3 = 1.648, substantially outperforming its architectural base Qwen3-VL-32B (74.0% MCQ, Hit@3 = 1.608) and all other medical-specialized

Model	History & Images			w/ Physician Reference			w/ RAG			w/ Agent Retrieval		
	Hit@1	Hit@3	MCQ	Hit@1	Hit@3	MCQ	Hit@1	Hit@3	MCQ	Hit@1	Hit@3	MCQ
SOTA & Large-scale Models												
GPT-5.5	0.508	0.700	82.8	**1.468**	1.636	**95.0**	0.308	0.472	81.4	0.570	0.884	83.2
Gemini 3 Pro	**0.684**	**0.948**	**88.4**	1.412	1.600	93.6	**0.568**	**0.820**	**88.0**	**0.738**	**1.010**	**87.6**
Qwen3.5-397B	0.264	0.496	71.0	1.352	1.588	94.4	0.204	0.388	60.4	0.392	0.646	71.4
GLM-4.6V	0.216	0.428	45.8	1.232	1.540	93.8	0.144	0.260	43.4	0.336	0.634	47.8
Kimi K2.6	0.464	0.672	31.8	1.376	1.548	23.8	0.336	0.504	23.2	0.638	0.870	30.8
General Open-source Models (Comparable Scale)												
Qwen3-VL-32B	0.224	0.392	74.8	1.404	1.608	74.0	0.184	0.352	76.2	0.348	0.626	74.7
Qwen3-VL-8B	0.100	0.196	72.0	1.352	1.524	71.6	0.136	0.244	69.2	0.316	0.618	68.3
Qwen3-VL-4B	0.068	0.144	64.5	1.250	1.436	63.2	0.096	0.192	58.4	0.248	0.492	61.2
Medical Specialized Models												
Hulu-Med-32B	0.290	0.504	62.4	1.390	1.648	88.2	0.256	0.428	60.4	0.448	0.772	63.5
Lingshu-32B	0.252	0.456	58.6	1.308	1.570	80.4	0.224	0.390	56.8	0.372	0.668	59.2
HuatuoGPT-Vision-7B	0.164	0.296	51.4	1.268	1.504	68.5	0.152	0.264	47.6	0.304	0.568	52.1
MedGemma-27b-it	0.316	0.536	75.2	1.284	**1.700**	74.0	0.292	0.572	72.4	0.352	0.626	69.9
MedGemma-4b-it	0.196	0.352	37.2	1.220	1.492	34.6	0.188	0.336	28.0	0.320	0.598	38.4

Table 1: Performance under different evidence retrieval paradigms on the test set ($n = 500$). The Base setting includes only clinical history and images; the other configurations augment with external context. Best results per column are **bolded**; best within each model category are underlined. Gray values indicate anomalous MCQ scores due to format non-compliance.

models. This indicates that domain-specific training enhances not only intrinsic diagnostic knowledge but also the capacity to integrate and reason over external evidence effectively.

RAG introduces retrieval noise that degrades performance. Under practical settings without human curation, the standard RAG paradigm leads to consistent degradation: GPT-5.5 Hit@1 drops from 0.508 to 0.308, and Hulu-Med-32B from 0.290 to 0.256. General-purpose retrieval tends to introduce redundant or weakly relevant literature that disrupts the model's reasoning chains. The degradation is most pronounced for open-ended diagnosis requiring deep cross-evidence integration, while MCQ accuracy shows more varied patterns—remaining stable for models with strong format compliance (Gemini 3 Pro: 88.4→88.0) but dropping substantially for others (Qwen3.5-397B: 71.0→60.4).

Agent Retrieval partially recovers performance. Compared with standard RAG, the Agent Retrieval paradigm shows promise in noise filtering via multi-step planning and iterative searching, recovering and in several cases surpassing the baseline (GPT-5.5: 0.570 vs. base 0.508; Hulu-Med-32B: 0.448 vs. base 0.290). However, a clear gap remains relative to physician reference context, suggesting that current agentic systems still struggle to reliably pinpoint the most discriminative diagnostic evidence from vast biomedical literature.

Format compliance remains a confound. Kimi K2.6 achieves competitive open-ended Hit@1 scores (0.464 base, 1.376 with PR) yet consistently produces anomalously low MCQ accuracy (23.8–31.8% across all conditions), indicating systematic format non-compliance rather than diagnostic deficits. MedGemma-4b-it exhibits similar patterns (MCQ 28.0–38.4%). These observations parallel the findings in Table 4 and reinforce the necessity of evaluating both open-ended and close-ended metrics jointly.

5.3 Direct Evaluation of Retriever Performance

To complement the end-to-end analysis above, we directly evaluate retriever performance using standard IR metrics against the physician-cited PMID set (Table 2). Across the same 500 cases with LLM-generated target queries, all three retrievers exhibit strikingly low absolute recall: even bge-large at top-10 recovers only 3.94% of the cited references

Retriever	Top-k	Retrieval Metrics			Downstream MCQ Acc.	
		Hit@k	Recall@k	MRR	Qwen3.5	Kimi-K2.6
bge-large	5	11.42	2.11	6.17	55.60	31.00
bge-large	10	19.04	3.94	7.17	57.00	27.80
MedCPT	5	3.61	0.71	1.89	51.60	21.80
MedCPT	10	5.81	1.08	2.23	53.40	21.60
Contriever	5	5.61	1.06	2.96	52.60	24.40
Contriever	10	9.22	1.90	3.46	52.60	23.60

Table 2: Retrieval quality and downstream diagnostic accuracy under the same 500-case test set and LLM-generated target queries. Retrieval metrics (Hit@k, Recall@k, MRR) are computed against physician-cited PMIDs and are retriever-dependent only; downstream accuracy reflects how each model utilizes the retrieved context.

(MRR = 7.17%), while MedCPT and Contriever remain below 2% Recall@10.

This direct measurement corroborates the end-to-end findings in Table 1: retrieval is the dominant bottleneck. The ranking among retrievers is consistent—bge-large achieves the highest citation recall and the best downstream accuracy for Qwen3.5—confirming that downstream diagnostic gains track retrieval quality when the reasoning model is held constant. However, the relationship is not purely mechanical: Kimi-K2.6 shows degraded accuracy at top-10 compared to top-5 despite higher recall, suggesting that models with lower noise tolerance may be harmed by additional retrieved documents even when coverage improves.

The low absolute recall reflects the inherent difficulty of matching LLM-generated queries to highly specialized physician-cited literature. Each CURE case cites an average of 6.0 PMIDs spanning rare radiological and pathological findings; surface-level lexical or embedding similarity is often insufficient to bridge the *clinical semantic gap* between patient-level descriptions and generalized biomedical knowledge. This pattern extends beyond dense retrievers: BM25 achieves only Hit@5 = 7.82% and Recall@5 = 1.49%, while adding a reranking stage improves Hit@3 to 12.42% (Recall@3 = 2.53%)—higher than raw sparse retrieval but still far below clinical utility.

Retrieval success predicts diagnostic accuracy. To establish a more direct link between retrieval quality and downstream performance, we stratify cases by whether the retrieved set contains at least one physician-cited PMID (Table 3). The results reveal a striking gap: for Qwen3.5, retrieval-hit cases achieve 84.6–87.8% accuracy versus 58.9–59.1% for retrieval-miss cases, a delta of +25 to +29 percentage points. GPT-5.5 shows a smaller but consistent advantage (+8.8pp under BM25 top-5). This stratified analysis provides the strongest evidence that retrieval quality causally drives diagnostic performance: holding the model and query constant, the sole variable is whether the retriever surfaced relevant evidence.

Model	Retrieval Setting	Hit Acc.	Miss Acc.	Δ
Qwen3.5	BM25 top-5	84.62	59.13	+25.49
Qwen3.5	AR top-3	87.76	58.89	+28.87
GPT-5.5	BM25 top-5	84.62	75.87	+8.75

Table 3: Diagnostic accuracy stratified by retrieval success (whether at least one physician-cited PMID was retrieved). The consistent accuracy gap confirms that retrieval quality is a direct predictor of diagnostic performance.

Having established that evidence quality is the primary driver of diagnostic accuracy (with further analysis of faithfulness and evidence similarity provided in Appendix C.6), we next examine how the *form* of the input—raw images versus textualized findings—affects model performance independently of retrieval.

5.4 Impact of Input Modalities on Diagnostic Performance

Table 4 reveals a consistent performance hierarchy across input conditions: clinical history alone yields the lowest scores, with progressive gains from adding images, imaging findings, or both. Among SOTA models, GPT-5.5 and Gemini 3 Pro dominate across all conditions, with GPT-5.5 achieving the highest overall scores (Hit@1 = 1.664, MCQ = 94.2% under the full multimodal setting). A clear gap separates these frontier models from the remaining systems.

Medical specialization provides consistent advantages over general-purpose counterparts. Among medical-specialized models, Hulu-Med-32B achieves the strongest performance, outperforming all comparable-scale general models by substantial margins. Compared with Qwen3-VL-32B (its architectural base), Hulu-Med-32B improves Hit@1 by 62% under clinical history only (0.420 vs. 0.260), 63% with images (0.520 vs. 0.320), and 39% with imaging findings (1.250 vs. 0.900). MCQ accuracy shows similar gains (+7.4, +6.6, +6.0, and +4.6 percentage points across the

Model	Clinical History			+ Image			+ Imaging Findings			+ Both		
	Hit@1	Hit@3	MCQ	Hit@1	Hit@3	MCQ	Hit@1	Hit@3	MCQ	Hit@1	Hit@3	MCQ
SOTA & Large-scale Models												
GPT-5.5	**0.667**	**0.777**	**69.6**	**1.047**	**1.126**	**81.4**	**1.661**	**1.717**	**92.8**	**1.664**	**1.713**	**94.2**
Gemini 3 Pro	0.612	0.724	64.8	0.956	1.042	78.2	1.580	1.645	91.4	1.610	1.685	93.6
Qwen3.5-397B	0.394	0.496	58.6	0.631	0.758	31.8	1.337	1.438	89.2	1.250	1.404	32.8
GLM-4.6V	0.145	0.194	41.0	0.126	0.168	45.6	0.260	0.324	82.2	0.166	0.208	76.0
Kimi K2.6	0.184	0.291	31.8	0.146	0.191	27.6	0.459	0.540	27.6	0.433	0.576	12.2
General Open-source Models (Comparable Scale)												
Qwen3-VL-32B	0.260	0.320	43.4	0.320	0.370	55.8	0.900	1.050	80.6	0.920	1.100	83.6
Qwen3-VL-8B	0.200	0.270	41.8	0.240	0.300	33.6	0.750	0.900	76.4	0.770	0.950	66.8
Qwen3-VL-4B	0.150	0.220	35.4	0.180	0.250	39.4	0.600	0.750	74.2	0.620	0.780	70.6
Medical Specialized Models												
Hulu-Med-32B	0.420	0.510	50.8	0.520	0.610	62.4	1.250	1.380	86.6	1.280	1.420	88.2
Lingshu-32B	0.350	0.420	45.0	0.400	0.480	52.0	1.100	1.250	80.0	1.150	1.300	82.0
HuatuoGPT-Vision-7B	0.290	0.340	43.2	0.334	0.362	51.6	0.996	1.110	76.2	0.994	1.100	76.4
MedGemma-27b-it	0.324	0.482	26.0	0.316	0.460	29.6	1.218	1.350	18.2	1.138	1.286	30.2
MedGemma-4b-it	0.286	0.420	28.6	0.318	0.432	36.0	1.032	1.178	27.4	1.020	1.182	49.4

Table 4: Effect of input modalities on diagnostic performance. Clinical History provides only the textual narrative; the other configurations progressively add imaging modalities. Best results per column are **bolded**; best within each model category are underlined. Gray values indicate anomalous MCQ scores due to format non-compliance.

four conditions). Lingshu-32B similarly outperforms its general-purpose counterparts, confirming that domain-specific training provides robust benefits for clinical diagnostic reasoning.

Imaging findings remain the dominant modality. Across all model categories, adding textual imaging findings produces the largest single-step improvement (averaging +0.6 Hit@1 over the image-only condition), consistent with the reasoning-friendly nature of structured textual evidence. Adding raw images yields smaller and more variable gains: for instance, GLM-4.6V and Kimi K2.6 show negligible or negative improvement from images alone, suggesting limited visual grounding capability in these models. The full multimodal setting (+ Both) generally achieves the best results, though the marginal gain over findings-only is modest for most models, indicating that current systems extract limited additional diagnostic signal when structured findings are already available.

Format compliance issues affect MCQ evaluation. Several models exhibit anomalously low MCQ scores that contrast with reasonable open-ended performance: Kimi K2.6 drops to 12.2% MCQ under the + Both condition despite Hit@1 = 0.433, and Qwen3.5-397B shows MCQ = 31.8% with images despite Hit@1 = 0.631. Similarly, MedGemma models achieve competitive open-ended Hit scores (e.g., MedGemma-27b Hit@3 = 1.350 under + Findings) but severely underperform on MCQ (18.2%), suggesting systematic format non-compliance rather than diagnostic capability deficits. These results highlight that MCQ evaluation alone can underestimate model capabilities when output formatting is unreliable, underscoring the importance of complementary open-ended evaluation.

6 Conclusion

We introduce CURE, a multimodal benchmark of 500 physician-annotated clinical cases that separates end-to-end diagnosis into evidence retrieval and clinical understanding. Across 16 models and four evidence paradigms, we find three key results: (1) retrieval rather than reasoning is the main deployment bottleneck, as models struggle to surface decisive literature at scale; (2) greater faithfulness improves performance with high-quality evidence but increases vulnerability to noisy or irrelevant retrieval; (3) textual findings consistently enhance diagnostic accuracy, whereas images provide only modest and variable gains, indicating substantial room for improvement in multimodal clinical.

Limitations

Prompt sensitivity. All models are evaluated using fixed, zero-shot prompt templates designed for consistency. We do not systematically explore alternative prompting strategies such as chain-of-thought, few-shot exemplars, or self-consistency decoding. Given that prompt design can significantly affect diagnostic reasoning behavior, the relative performance rankings reported here may shift under different elicitation protocols.

Retrieval strategy coverage. Our RAG experiments employ a specific pipeline (dual-encoder ensemble with reranking) and a single agentic framework (MiroFlow). More advanced retrieval strategies—such as iterative query decomposition, multi-hop retrieval, or hybrid sparse-dense pipelines with domain-adapted re-rankers—remain unexplored and may substantially narrow the gap between automated retrieval and physician reference performance.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Guilherme Del Fiol, T Elizabeth Workman, and Paul N Gorman. 2014. Clinical questions raised by clinicians at the point of care: a systematic review. *JAMA internal medicine*, 174(5):710–718.

Vardhan Dongre, Chi Gui, Shubham Garg, Hooshang Nayyeri, Gokhan Tur, Dilek Hakkani-Tür, and Vikram S Adve. Mirage: A benchmark for multimodal information-seeking and reasoning in agricultural expert-guided conversations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. 2020. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):136.

Zhongzhen Huang, Linjie Mu, Yakun Zhu, Xiangyu Zhao, Shaoting Zhang, and Xiaofan Zhang. 2025. Elicit and enhance: Advancing multimodal reasoning in medical scenarios. *arXiv preprint arXiv:2505.23118*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4745–4759.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2025. Biomedrag: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, 162:104769.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled

knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Yi Qin, Lehan Wang, Chenxu Zhao, Alex PW Lee, and Xiaomeng Li. 2026. Incentivizing cardiologist-like reasoning in mllms for interpretable echocardiographic diagnosis. *arXiv preprint arXiv:2601.08440*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Shiqian Su, Sen Xing, Xuan Dong, Muyan Zhong, Bin Wang, Xizhou Zhu, Yuntao Chen, Wenhai Wang, Yue Deng, Pengxiang Zhu, and 1 others. 2026. Miroflow: Towards high-performance and robust open-source agent framework for general deep research tasks. *arXiv preprint arXiv:2602.22808*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, and 1 others. 2025. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *arXiv preprint arXiv:2511.11793*.

V Team. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

Renjun Xu and Jingwen Peng. 2025. A comprehensive survey of deep research: Systems, methodologies, and applications. *arXiv preprint arXiv:2506.12594*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze D Gui, Ziran W Jiang, Ziyu Jiang, and 1 others. 2024. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, and 1 others. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

Duoer Zhang, Wenbo Xiao, Chen Jiang, Yuxuan Qiu, Zhan Feng, Hong Wang, Yefeng Zheng, and Wentao Zhu. 2025. A multimodal contrastive learning for detecting aortic dissection on 3d non-contrast ct with anatomy simplification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, and 1 others. 2024. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. Diagnosisarena: benchmarking diagnostic reasoning for large language models. *arXiv preprint arXiv:2505.14107*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *International Conference on Machine Learning*, pages 80961–80990. PMLR.

A Prompt Templates

VLM System Prompt. No system prompt was used for the selected local VLM runs. The request

was sent as a user message containing image parts followed by text.

MCQ User Prompt Template.

```
Clinical history:
{clinical_history}

Imaging findings:        [condition-dependent]
{imaging_findings}

References:              [condition-dependent]
{references}

Question:
What is the most likely diagnosis?
Choose the single best answer from the provided
options. Respond with only the option letter.
A: ... B: ... C: ... D: ...

Put your final answer in \boxed{}.
```

Notes:

- `Imaging findings` appears only under `clinical_history_imaging_findings` and `clinical_history_image_imaging_findings` conditions.
- `References` appears only under `clinical_history_image_references`.
- Images are supplied as separate image parts before the text for all image conditions.

Open-Ended User Prompt Template.

```
Clinical history:
{clinical_history}

Imaging findings:        [condition-dependent]
{imaging_findings}

References:              [condition-dependent]
{references}

Question:
What is the final diagnosis?
Respond with up to three diagnoses, ordered from
most likely to less likely. Use this exact format:
1. <most likely diagnosis>
2. <second diagnosis or N/A>
3. <third diagnosis or N/A>
Do not include explanations.
```

Open-Ended Judge Prompts.

The judge model receives the system prompt `"You are a precise medical diagnostic evaluator."` followed by a batch user prompt:

```
Evaluate each case independently. Compare only
the parsed prediction text with the gold
diagnosis; do not infer a new diagnosis from the
original case.

Scores:
  2 = fully correct or clinically equivalent
  1 = partially correct (right disease family,
      organ, pathology, mechanism, or imaging
     pattern but missing an important qualifier)
  0 = incorrect, unrelated, overly broad, empty,
      or non-diagnostic

For top-3, score each rank independently;
best_score_at_3 = max(rank1, rank2, rank3).

There are exactly {N} cases below. Return exactly
{N} result objects. Return compact JSON only:
{"results":[{"case_id":"...",
"scores":[s1,s2,s3]}]}

Cases:
{cases_json}
```

B Qualitative Case Analysis

We present three representative cases from EuroRad 2025 to illustrate characteristic failure patterns of current medical vision-language models. For each case, we show model predictions under both multiple-choice (MCQ) and open-ended diagnostic settings, across input conditions: text-only clinical history, clinical history with images, and clinical history with images plus structured imaging findings.

Case 19152: Emphysematous Hepatitis (55M, Abdominal Imaging). All models exhibit *frequency bias*, defaulting to emphysematous cholecystitis—a more prevalent gas-producing hepatobiliary condition—over the extremely rare emphysematous hepatitis (<30 CT-confirmed cases in literature). Notably, MedGemma-4B answers correctly with text-only input but fails when images are added, suggesting visual features introduce confounding bias toward the more familiar cholecystitis pattern. In the open-ended setting, Gemma-4-31B includes the correct diagnosis at rank-3, while all other models entirely miss the entity. Adding structured imaging findings text resolves the ambiguity for 3 out of 4 models.

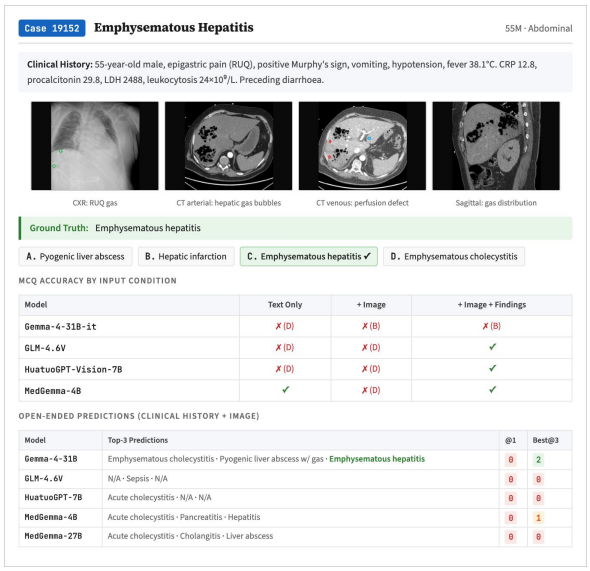

Figure 3: Case 19152 analysis. Models default to emphysematous cholecystitis (option D) across conditions. Structured imaging findings text provides stronger diagnostic signal than raw CT images.

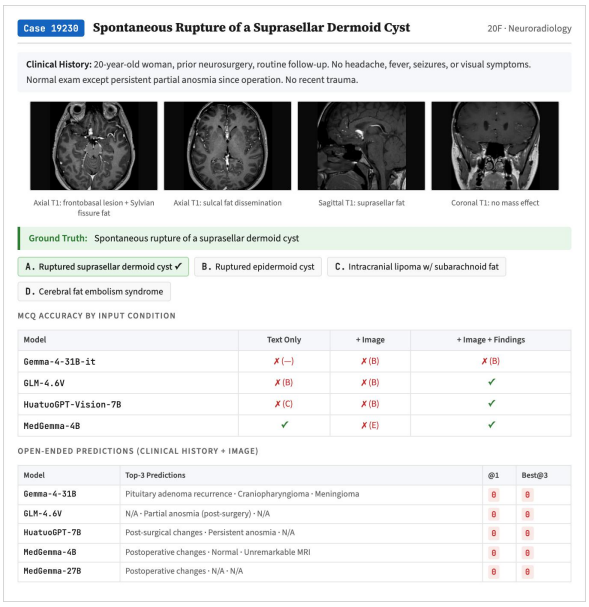

Figure 4: Case 19230 analysis. Open-ended predictions collapse to "post-surgical changes" across all models, demonstrating semantic anchoring on prior history rather than new imaging findings.

Case 19230: Ruptured Suprasellar Dermoid Cyst (20F, Neuroradiology). This case exposes two failure modes. First, *dermoid–epidermoid confusion*: in MCQ with images, all models select "ruptured epidermoid cyst" (B), failing to leverage T1-hyperintense signal pathognomonic for fat (dermoid). Second, *semantic anchoring*: in open-ended generation, models anchor on "prior neurosurgery" in the clinical history and predict "post-surgical changes," entirely ignoring the new finding of subarachnoid fat dissemination. This demonstrates that clinical context can paradoxically mislead models when they lack capacity to reason about imaging-specific signs.

Case 19277: Osteochondroma of the Acromion (27F, Musculoskeletal). This case reveals a striking *MCQ vs. open-ended gap*. Gemma-4-31B correctly identifies osteochondroma in constrained MCQ with images, yet produces entirely unrelated predictions (lipoma, liposarcoma) in free generation—demonstrating that elimination reasoning in MCQ does not generalize to unconstrained diagnostic settings. The universal text-only bias toward "chondrosarcoma" (D) reflects malignancy priors triggered by "accelerated growth" in the clinical history. With structured findings (mentioning corticomedullary continuity), all models correctly identify the benign entity.

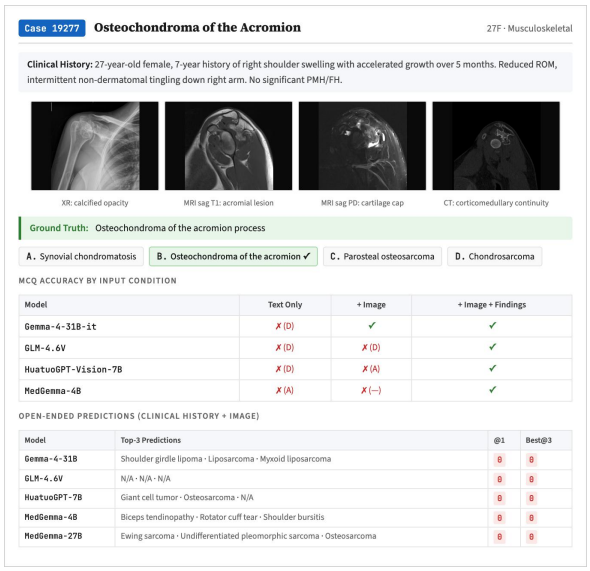

Figure 5: Case 19277 analysis. Gemma-4-31B succeeds in MCQ with images but fails completely in open-ended generation, revealing that constrained format inflates apparent diagnostic capability.

Condition	GLM-4.6V	Qwen3.5-397B	GPT-5.5
Clinical History + Image			
Base (no reference)	45.8	71.0	82.8
+ Gold ref top-5	93.0	92.6	94.4
+ Random ref top-5	47.4	67.0	82.0
+ Shuffled ref top-5	44.2	67.2	81.6
Clinical History + Image + Imaging Findings			
Base (no reference)	75.0	89.8	92.8
+ Gold ref top-5	96.8	96.6	97.2
+ Random ref top-5	80.2	87.4	93.8
+ Shuffled ref top-5	78.6	86.4	94.0
Clinical History only			
Base (no reference)	41.0	57.4	68.6
+ Random ref top-5	39.2	49.6	69.6
+ Shuffled ref top-5	36.4	48.6	67.4

Table 5: MCQ accuracy (%) with random and shuffled references vs. gold references. Random/shuffled references provide negligible or negative benefit, confirming that diagnostic improvement requires content-specific evidence rather than generic contextual augmentation.

C Ablation Studies on Evidence Retrieval

We conduct comprehensive ablation studies to further characterize how evidence quality, granularity, retrieval volume, and post-processing affect diagnostic performance. All experiments use the same 500-case test set described in Section 5.

C.1 Random and Shuffled Reference Controls

To confirm that diagnostic gains from physician references stem from content-specific information rather than mere context augmentation, we replace gold references with (a) randomly sampled PubMed articles unrelated to the case, and (b) shuffled references drawn from other cases in the dataset.

Table 5 shows that random and shuffled references provide no meaningful benefit—and in several cases degrade performance below the no-reference baseline (e.g., Qwen3.5-397B drops from 71.0% to 67.0–67.2% with irrelevant references). This contrasts sharply with gold references, which yield gains of +10 to +47 percentage points. The result confirms that the improvements observed with physician references in Table 1 are attributable to diagnostically relevant content, not to the mere presence of additional text in the prompt.

C.2 Reference Granularity Ablation

We investigate which component of the physician-cited references is most valuable by providing different granularities: title only, abstract only, top-5 most relevant references, filtered references (those directly cited in the discussion), and all cited references with full metadata.

Table 6 reveals a striking finding: **providing only the titles of relevant references achieves 88–97% of the full-reference performance across all conditions.** For GPT-5.5 under Clinical History + Image, title-only (92.8%) is within 1.6 percentage points of all references (92.6%). This implies that the primary challenge in evidence-based diagnosis is *locating* the correct literature—once the model knows which papers are relevant, even minimal information (a title acting as a semantic pointer) is sufficient to activate the correct diagnostic pathway. This finding has direct implications for retrieval system design: retrieval precision (finding the right papers) matters far more than extraction depth (how much of each paper is provided).

The one exception is the "filtered" condition, which consistently underperforms other granularities by 3–7 percentage points, suggesting that filtering to only discussion-cited references sometimes excludes complementary evidence that appears in the broader reference list.

C.3 Retrieval Post-processing Ablation

Standard RAG pipelines often incorporate reranking and context compression to improve relevance and reduce noise. We evaluate whether these post-processing steps benefit clinical diagnostic reasoning.

Table 7 demonstrates that standard RAG post-

Reference Granularity	GLM-4.6V	Qwen3.5-397B	GPT-5.5
Clinical History + Image			
Title only	88.0	93.0	92.8
Abstract only	93.0	93.6	92.4
Top-5 references	93.0	92.6	94.4
Filtered (discussion-cited)	86.8	90.8	91.8
All references	93.2	93.8	92.6
Clinical History + Image + Imaging Findings			
Title only	94.2	96.8	97.0
Abstract only	96.2	95.6	94.4
Top-5 references	96.8	96.6	97.2
Filtered (discussion-cited)	92.2	94.8	94.6
All references	96.6	96.6	95.2
Clinical History only			
Title only	91.2	89.8	93.0
Abstract only	91.0	91.6	92.2
Top-5 references	91.6	91.0	90.6
Filtered (discussion-cited)	85.4	87.2	90.6
All references	91.8	92.0	92.4

Table 6: MCQ accuracy (%) under different reference granularities. Title-only references achieve near-ceiling performance, indicating that identifying the correct literature is more important than extracting detailed content from it.

processing is **counterproductive** for clinical diagnostic reasoning:

- **Compression is consistently harmful:** Adding compression reduces GPT-5.5 accuracy by 7–12 percentage points across all settings. Context compression, designed to remove "irrelevant" content, inadvertently discards subtle diagnostic cues that are critical for rare disease identification.

- **Reranking has mixed effects:** Reranking occasionally improves performance marginally (e.g., References Top-1: 62.8→63.2 for Qwen3.5) but more often degrades it, particularly for GPT-5.5 (80.0→77.0 for Cases Top-1).

These results suggest that in the medical diagnostic domain, where relevant information is often distributed across subtle textual cues rather than concentrated in easily identifiable passages, aggressive post-processing destroys diagnostic signal. This contrasts with findings in general-domain QA where compression typically preserves or improves performance.

C.4 Retrieval Volume Sensitivity

We vary the number of retrieved documents (k) to assess the trade-off between coverage and noise.

Table 8 reveals a non-monotonic relationship between retrieval volume and diagnostic accuracy.

Source	Post-processing	Qwen3.5-397B	GPT-5.5
EuroRAD Cases BM25			
Top-1	None	61.4	80.0
Top-1	+ Rerank	59.0	77.0
Top-1	+ Rerank + Compress	55.0	68.4
Top-3	None	61.4	78.8
Top-3	+ Rerank	58.8	80.4
Top-3	+ Rerank + Compress	57.4	71.8
EuroRAD References BM25			
Top-1	None	62.8	75.8
Top-1	+ Rerank	63.2	71.4
Top-1	+ Rerank + Compress	57.8	67.8
Top-3	None	62.2	75.6
Top-3	+ Rerank	61.4	71.0
Top-3	+ Rerank + Compress	57.0	70.0

Table 7: MCQ accuracy (%) with retrieval post-processing ablation (Clinical History + Image setting). Compression consistently degrades performance; reranking shows mixed effects.

Top-k	EuroRAD Cases		EuroRAD References	
	Qwen3.5	GPT-5.5	Qwen3.5	GPT-5.5
Clinical History + Image (MCQ %)				
Top-1	61.4	80.0	62.8	75.8
Top-3	61.4	78.8	62.2	75.6
Top-5	60.8	83.2	60.0	76.4
Top-20	51.6	81.4	60.4	77.2
Clinical History only (MCQ %)				
Top-5 (history query)	44.4	81.2	48.6	82.6
Top-20 (LLM query)	48.6	81.4	48.6	77.2

Table 8: MCQ accuracy (%) as a function of retrieval volume (Clinical History + Image setting, BM25 with LLM-generated queries). The relationship between k and accuracy is non-monotonic and model-dependent.

GPT-5.5 shows relative robustness across k values (78.8–83.2% for Cases), while Qwen3.5-397B degrades substantially at Top-20 (51.6%, down from 61.4% at Top-1). This divergence in noise tolerance aligns with the findings in Section 5.2: models with stronger intrinsic capability (GPT-5.5) can better filter irrelevant retrieved content, while smaller or less robust models are more susceptible to retrieval noise. The practical implication is that optimal k should be tuned per-model rather than set uniformly.

C.5 ReAct-style Agentic Retrieval

We compare standard single-shot BM25 retrieval with a ReAct-style (**?**) agentic approach where the model iteratively formulates queries and decides when to stop retrieving.

Table 9 shows that ReAct-style agentic retrieval over BM25 provides marginal and inconsistent improvements compared with standard single-shot

Source	Method	Qwen3.5-397B	GPT-5.5
Clinical History + Image (MCQ %)			
Cases	BM25 Top-1	61.4	80.0
Cases	BM25 Top-3	61.4	78.8
Cases	ReAct Top-1	60.8	79.8
Cases	ReAct Top-3	59.2	80.2
References	BM25 Top-1	62.8	75.8
References	BM25 Top-3	62.2	75.6
References	ReAct Top-1	65.2	72.6
References	ReAct Top-3	62.8	73.8

Table 9: MCQ accuracy (%) comparing standard BM25 retrieval with ReAct-style agentic retrieval. ReAct shows marginal and inconsistent improvements over single-shot BM25.

retrieval. While Qwen3.5-397B benefits slightly from ReAct on references (62.8→65.2 at Top-1), GPT-5.5 shows no clear gain and occasionally degrades. This suggests that when the underlying retrieval index has low recall (as demonstrated in Section 5.3), iterative query reformulation alone cannot compensate for the fundamental limitation of the knowledge base. The stronger Agent Retrieval results observed in Table 1 (using MiroFlow with unconstrained web search) indicate that expanding the search space beyond a fixed corpus is more impactful than sophisticating the retrieval strategy within it.

C.6 Analysis of Retrieved Evidence Quality

Figure 6 shows that diagnostic behavior shifts dramatically with evidence quality. Under physician reference (coverage 92.1%), most diagnoses are *evidence-based correct* (64.5%), indicating effective use of high-quality evidence. Under RAG (coverage 23.0%), the dominant outcome becomes *evidence-misled* (57.3%) with high hallucination (31.6%); Agent Retrieval improves coverage to 38.2% and increases evidence-based correctness (7.0%→14.0%), but *evidence-misled* remains prevalent (53.7%) and hallucination stays substantial (26.8%). Meanwhile, *knowledge-based correct* is consistently rare (2.1%/4.1%/5.5% for PR/RAG/AR), suggesting models seldom override provided evidence using internal knowledge.

The core bottleneck in RAG/AR pipelines is retrieval quality and coverage rather than model unfaithfulness: models are often *too* faithful to the retrieved documents, so when retrieval fails to include diagnostically relevant literature, grounded but incorrect reasoning becomes the dominant failure mode. Per-model analysis further indicates faithfulness is a double-edged sword—high groundedness can yield strong performance under curated evidence yet increases susceptibility to being misled under low-quality retrieval, whereas lower groundedness does not improve accuracy and instead correlates with higher hallucination.

Figure 7 provides a mechanistic explanation: diagnostic correctness positively tracks the lexical alignment between retrieved evidence and the physician-annotated evidence (e.g., within the PR cohort, RAG–PR similarity drops from 0.28 to 0.17 as outcomes worsen), while similarity to the clinical history remains largely uninformative (0.14–0.18 across strata). The intrinsically low Hx–PR similarity of 0.06 to 0.09 highlights a *clinical semantic gap* between patient narratives and generalized medical knowledge, reinforcing that locating PR-like literature rather than matching Hx drives correct diagnosis.

Model	Keyword F1	Bigram F1	ROUGE-L F1
Qwen3.5-397B	0.211	0.043	0.123
Kimi K2.6	0.155	0.033	0.096

Table 10: Imaging finding generation quality (Clinical History + Images → Imaging Findings) evaluated against physician-written findings on the test set ($n = 500$). All metrics are token-level F1 scores.

C.7 Imaging Finding Generation from Raw Images

To quantify the visual grounding capability of current models, we evaluate their ability to generate structured imaging findings from clinical history and raw images—a task analogous to radiologist-level image interpretation. This experiment contextualizes the modality gap observed in Table 4, where providing gold imaging findings yields substantially larger diagnostic gains than providing raw images.

Table 10 shows that both models achieve low overlap with physician-written findings (Keyword F1 ≤ 0.211, ROUGE-L ≤ 0.123), indicating that current VLMs struggle to extract diagnostically relevant visual features from medical images at the level required for clinical reporting. This result provides a mechanistic explanation for the modality gap in Table 4: models benefit enormously from textualized findings for downstream reasoning, but cannot reliably generate such findings from raw pixels themselves. The visual grounding gap—translating pixel-level patterns

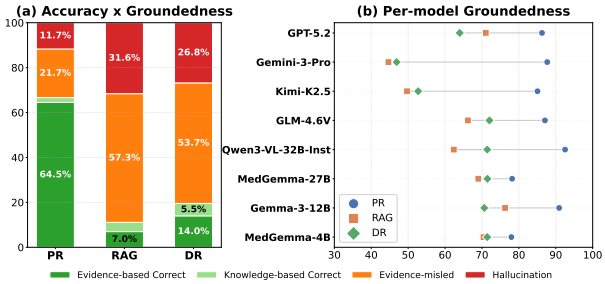

Figure 6: Faithfulness analysis across evidence conditions. (a) Distribution of accuracy × groundedness quadrants averaged over 16 models. (b) Per-model groundedness rate under each evidence condition.

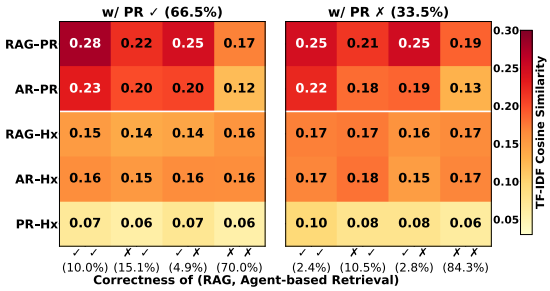

Figure 7: Heatmap of TF-IDF cosine similarities comparing retrieved evidence (RAG and AR), clinical history (Hx), and physician-annotated reference evidence (PR). Panels are stratified by baseline correctness.

into structured, diagnostically actionable observations—represents a distinct bottleneck from the retrieval gap identified in Sections 5.2–5.3, and suggests that improving medical image understanding is a complementary direction to improving evidence retrieval for advancing clinical diagnostic AI.