Journal of Sustainability, Policy, and Practice

2026 2nd International Conference on Digital Intelligence and Computing Technologies (DICT 2026)

Review

Chain-of-Thought Prompting with Retrieval-Augmented Generation for Explainable Chest Radiograph Diagnosis: A Review of Vision-Language Models

Yali Zhang [1,*]

[1] Master of Computer Science, Rice University, Houston, TX, USA
* Correspondence: Yali Zhang, Master of Computer Science, Rice University, Houston, TX, USA

Abstract: The integration of multimodal large language models (LLMs) into medical image interpretation has introduced new possibilities for clinical decision support, particularly through chain-of-thought (CoT) prompting and retrieval-augmented generation (RAG). This review examines existing CoT prompting strategies and RAG-enhanced approaches that aim to improve both the diagnostic accuracy and clinical explainability of vision-language models (VLMs) in medical imaging analysis, with chest X-ray (CXR) interpretation as the primary application domain. A two-dimensional analytical framework is proposed to categorize current approaches along reasoning granularity and evidence grounding depth. Representative methods published between 2023 and 2025 are systematically reviewed, evaluated across public benchmarks including MIMIC-CXR, CheXpert, VQA-RAD, SLAKE, and PathVQA. The analysis indicates that structured multi-step CoT prompting combined with domain-specific retrieval consistently outperforms zero-shot and standard prompting baselines, with factual accuracy improvements ranging from 20.8% to 43.8% when RAG modules are incorporated. Six evaluation dimensions for explainability assessment in medical imaging contexts are identified. Limitations of current approaches and directions for future investigation are discussed.

Keywords: chain-of-thought prompting; retrieval-augmented generation; medical image interpretation; vision-language models; explainable artificial intelligence

Received: 26 April 2026
Revised: 14 June 2026
Accepted: 26 June 2026
Published: 02 July 2026

1. Introduction

1.1. Clinical Challenges in Medical Image Interpretation

Chest radiography remains the most frequently performed diagnostic imaging examination worldwide, with billions of medical imaging studies conducted annually across global healthcare systems. The interpretation of CXR images demands simultaneous assessment of cardiac silhouette, pulmonary parenchyma, mediastinal structures, and osseous anatomy, placing substantial cognitive demands on radiologists who must process an ever-increasing volume of studies under time constraints.

The clinical burden is compounded by a well-documented workforce shortage. The Association of American Medical Colleges (AAMC) projects a total physician shortage of 13,500 to 86,000 by 2036, with radiology among the most affected specialties [1]. Radiologist workload in the United States has increased substantially over the past decade, with studies documenting significant growth in imaging volume per radiologist [2], while the number of radiology residency training positions has remained at roughly 1,006 per year during the same period [3]. This persistent imbalance between imaging demand and available workforce capacity has intensified interest in artificial intelligence as a potential

augmentation tool for radiology practice. The aging of the U.S. population and the expansion of screening programs are projected to further accelerate imaging demand over the coming decades, exacerbating the current gap between imaging volume growth and workforce availability.

Diagnostic errors in radiology represent another persistent challenge that has resisted improvement over several decades. The real-time interpretive error rate in clinical practice ranges from 3% to 5%, translating to an estimated 40 million diagnostic imaging errors annually worldwide [4]. The retrospective discrepancy rate of approximately 30% has remained essentially unchanged since Garland's landmark study in 1949 [5], and these error rates are exacerbated by fatigue and excessive workload pressure. Multiple studies have demonstrated that error rates increase significantly when radiologists operate beyond 121% of their average daily reading volume [6], linking workforce shortages directly to patient safety concerns. The U.S. Food and Drug Administration (FDA) has authorized over 1,000 AI/ML-enabled medical devices as of late 2024, with approximately 76% falling within the radiology domain [1]. This regulatory trajectory underscores both the clinical demand and the growing institutional acceptance of AI-assisted diagnostic tools.

1.2. Research Scope and Contributions

1.2.1. Problem Definition and Objectives

While deep learning models have demonstrated strong classification performance on large-scale datasets---MIMIC-CXR comprising 377,110 CXR images paired with 227,835 free-text radiology reports from 65,379 patients and CheXpert containing 224,316 images with 14 pathology labels from 65,240 patients ---the opacity of their decision-making processes has limited clinical adoption [2]. The emergence of multimodal LLMs offers a fundamentally different paradigm: these models can generate natural language explanations alongside diagnostic predictions, potentially addressing explainability requirements emphasized by regulatory bodies and clinical practitioners [3]. This review investigates how CoT prompting and RAG techniques can be systematically applied to VLMs for medical image interpretation, with chest radiograph diagnosis as the primary application domain and particular focus on the quality and clinical relevance of generated explanations.

1.2.2. Paper Organization

This paper makes three contributions. It proposes a two-dimensional analytical framework for categorizing CoT prompting strategies in medical image interpretation, distinguishing between reasoning granularity and evidence grounding depth [4]. It provides a comparative synthesis of representative approaches across standardized benchmarks. It consolidates evaluation dimensions for explainability assessment aligned with emerging regulatory frameworks for clinical AI.

2. Related Work

2.1. Vision-Language Models for Medical Imaging

2.1.1. Multimodal Foundation Models in Radiology

The application of VLMs to medical imaging has evolved rapidly from task-specific architectures to general-purpose foundation models. LLaVA-Med, introduced by Li et al., established an influential paradigm by adapting visual instruction-tuning to the biomedical domain, training on 15 million image-text pairs from PubMed Central (PMC) articles [5]. This work demonstrated that a large language-and-vision assistant could be effectively trained for biomedicine using a curriculum learning strategy that progresses from broad biomedical concept alignment to task-specific instruction following. The model achieved competitive performance on three medical VQA benchmarks, establishing baselines that subsequent models have sought to surpass [6]. The curriculum learning approach proved particularly effective in addressing the limited availability of

high-quality biomedical instruction-following data, offering a training paradigm that has been widely adopted in the development of domain-specific VLMs.

2.1.2. Biomedical Visual Representation Learning

Parallel efforts have focused on building domain-specialized visual encoders that capture clinically relevant features more effectively than general-domain alternatives. BiomedCLIP, developed by Zhang et al., was pretrained on the PMC-15M dataset comprising 15 million scientific image-text pairs extracted from 4.4 million PMC articles. By adapting contrastive learning to the biomedical domain, BiomedCLIP produces visual and textual embeddings that capture subtle morphological and pathological features critical for medical imaging analysis. The learned representations have been widely adopted as initialization for downstream medical VLMs, demonstrating that domain-specific pretraining on large-scale biomedical image-text corpora leads to more discriminative and clinically meaningful features compared to models pretrained solely on natural images.

2.2. Chain-of-Thought Reasoning in Clinical Contexts

CoT prompting, which elicits step-by-step reasoning from LLMs, has been adapted from general natural language processing to the medical imaging domain with domain-specific innovations. The Multimodal-CoT framework proposed by Zhang et al. introduced a two-stage approach separating rationale generation from answer inference across text and visual modalities [7]. Using a model with fewer than one billion parameters, this approach achieved state-of-the-art results on the ScienceQA benchmark, demonstrating that explicit reasoning chains can enhance multimodal understanding in compact models [8]. The separation of rationale generation from final answer prediction proved to be a critical architectural decision, as it allows the model to engage in deliberative reasoning before committing to a diagnostic conclusion.

MedCoT, proposed by Liu et al., extended this concept to medical visual question answering through a hierarchical expert verification process. The architecture employs an Initial Specialist to generate diagnostic rationales, a Follow-up Specialist to validate and refine these rationales, and a Mixture-of-Experts Diagnostic Specialist to reach consensus. This hierarchical structure achieved state-of-the-art performance on VQA-RAD, SLAKE, and PathVQA benchmarks, demonstrating the effectiveness of structured multi-expert reasoning for clinical diagnostic tasks. The progressive verification mechanism mirrors clinical workflows in which initial diagnostic impressions undergo peer review and specialist consultation before finalization. The explicit generation of intermediate reasoning steps also provides natural checkpoints at which erroneous reasoning can be identified and corrected before propagating to the final diagnostic output.

3. Analytical Framework for CoT-Based Explainable Diagnosis

3.1. Framework Overview and Taxonomy

This review adopts a structured analytical framework that organizes existing CoT-based approaches along two orthogonal dimensions: reasoning granularity and evidence grounding depth. Reasoning granularity captures the structural complexity of the generated reasoning chain, from single-step rationale generation to hierarchical multi-step processes mirroring clinical diagnostic workflows. Evidence grounding depth reflects the extent to which reasoning is anchored in external medical knowledge, from purely intrinsic parametric knowledge to actively retrieved evidence.

The framework defines four quadrants: (Q1) single-step intrinsic reasoning, representing baseline CoT prompting without retrieval; (Q2) multi-step intrinsic reasoning, employing hierarchical prompting chains relying on pretrained knowledge; (Q3) single-step retrieval-augmented reasoning, incorporating retrieved evidence in a one-pass generation process; and (Q4) multi-step retrieval-augmented reasoning, combining structured diagnostic chains with iterative evidence retrieval.

As shown in Figure 1, the reviewed methods exhibit a clear evolutionary trajectory from single-step intrinsic reasoning (Q1) toward multi-step retrieval-augmented approaches (Q4). Methods in Q3 and Q4, which incorporate external evidence retrieval, consistently report higher accuracy improvements than their intrinsic-only counterparts in Q1 and Q2. Notably, the progression from earlier approaches (2023) to more recent contributions (2025) reflects a trend toward increasing methodological complexity, with later methods combining hierarchical reasoning structures with domain-aware retrieval mechanisms.

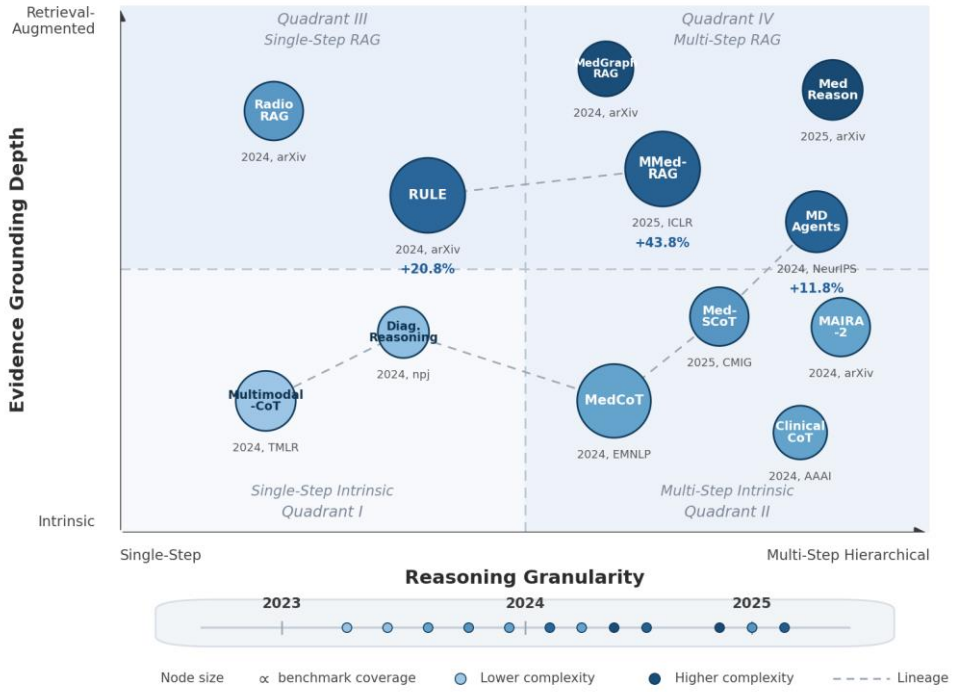


Figure 1. Taxonomic Classification Matrix of CoT Prompting Strategies for Medical Image Interpretation

3.2. Chain-of-Thought Prompting Strategies for Medical Image Interpretation

3.2.1. Hierarchical Clinical Reasoning Prompts

Tanno et al. conducted the largest expert evaluation of VLM-generated radiology reports to date, involving 27 board-certified radiologists [9]. Their findings revealed that 56.1% of ICU reports and 77.7% of in/outpatient reports generated by the model were rated as preferable or equivalent to clinician-authored reports. Savage et al. demonstrated that LLMs can replicate established clinical reasoning patterns---including differential diagnosis, intuitive reasoning, analytical reasoning, and Bayesian inference---through carefully designed diagnostic prompts without sacrificing accuracy [10]. Building on this, Kwon et al. introduced a reasoning-aware diagnosis framework where LLMs generate prompt-based rationales to guide diagnostic inference, showing that machine-generated reasoning can match expert-crafted rationales [11].

Med-SCoT, developed by Jiang et al., decomposes the reasoning process into four clinically meaningful stages: Summary, Caption, Reasoning, and Conclusion [12]. A collaborative correction pipeline (CoCo) enables iterative refinement between stages, and the SCoTEval framework provides structured metrics for evaluating reasoning quality at each stage independently.

3.2.2. Retrieval-Augmented Evidence Grounding

The integration of RAG with CoT prompting addresses a fundamental limitation of purely parametric reasoning: the inability to access current medical knowledge beyond

the training data cutoff. Xia et al. proposed RULE, introducing calibrated retrieval context selection via factuality risk control [13]. RULE achieved an average improvement of 20.8% in factual accuracy across three medical VQA datasets, demonstrating that selective, quality-controlled retrieval outperforms naive retrieve-then-generate approaches.

The same research group developed MMed-RAG, a domain-aware retrieval mechanism that adaptively selects strategies based on the medical imaging modality. MMed-RAG achieved an average improvement of 43.8% in factual accuracy across five medical datasets, representing the current state-of-the-art for multimodal medical RAG [14]. Table 1 summarizes the methods reviewed, organized by the proposed taxonomy.

Table 1. Summary of Reviewed CoT Prompting and RAG Strategies for Medical Imaging Diagnosis

Method	Year	Venue	Quadrant	Evidence Source	Primary Task
Multimodal-CoT	2024	TMLR	Q2	Intrinsic	Multimodal reasoning
MedCoT	2024	EMNLP	Q2	Intrinsic	Medical VQA
Tanno et al.	2024	Nature Med.	Q1	Intrinsic	CXR report generation
Diagnostic Reasoning	2024	npj Digit. Med.	Q2	Intrinsic	Differential diagnosis
Clinical CoT	2024	AAAI	Q2	Intrinsic	Clinical diagnosis
Med-SCoT	2025	CMIG	Q2	Intrinsic	Medical VQA
RULE	2024	arXiv	Q3	RAG with factuality control	Medical VQA
MMed-RAG	2025	ICLR	Q4	Domain-aware multimodal RAG	Medical VQA

Table 2 presents the key public datasets used across the reviewed approaches for training and evaluation.

Table 2. Public Chest X-Ray and Medical VQA Datasets Referenced in Reviewed Studies

Dataset	Scale	Annotation Type	Source Institution
MIMIC - CXR	377,110 images / 65,379 patients	227,835 free - text reports + 14 labels	MIT / BIDMC (PhysioNet)
CheXpert	224,316 images / 65,240 patients	14 pathology labels with uncertainty	Stanford AIMI
NIH ChestX - ray14	112,120 images / 30,805 patients	14 NLP - extracted labels	NIH Clinical Center

VQA - RAD	315 images	3,515 QA pairs	U.S. National Library of Medicine
SLAKE	642 images	14,028 bilingual QA pairs + KG	Multi - institutional
PathVQA	4,998 images	32,799 QA pairs	Multi - institutional
PMC - 15M	15 million image - text pairs	Scientific captions from 4.4M articles	PubMed Central

3.3. Explainability Evaluation Methodology

3.3.1. Automated Evaluation Metrics

The evaluation of explainability in CoT-based medical VLMs requires metrics beyond traditional classification accuracy. The literature identifies six core evaluation criteria applicable to the medical imaging domain: faithfulness (whether the explanation reflects the model's actual decision process, assessed via sufficiency and comprehensiveness tests); plausibility (alignment with expert clinical reasoning, commonly approximated by intersection-over-union between model attention regions and expert annotations, though this proxy metric captures spatial agreement rather than reasoning quality per se); consistency (stability under clinically irrelevant perturbations); localization (correct identification of relevant anatomical regions); clinical utility (impact on radiologist accuracy, confidence, and time-to-decision); and completeness (coverage of all clinically relevant findings).

3.3.2. Clinician-Centered Assessment Criteria

Clinician preference studies reported in the reviewed literature indicate that exemplar-based patient retrieval and LLM-generated free-text rationales scored highest in satisfaction assessments, while attention-based visual highlights were considered useful but less intuitive than narrative explanations. This finding supports the use of multimodal LLMs to generate natural language diagnostic explanations. Table 3 synthesizes these evaluation dimensions and their corresponding assessment protocols.

Table 3. Explainability Evaluation Dimensions and Corresponding Metrics

Dimension	Definition	Representative Metrics	Assessment Method
Faithfulness	The explanation reflects the actual decision - making process.	Sufficiency, comprehensiveness	Automated (perturbation - based)
Plausibility	Alignment with expert clinical reasoning.	Attention IoU with expert annotations (spatial proxy)	Hybrid (automated + expert)
Consistency	Stability under irrelevant perturbations.	Prediction consistency rate	Automated
Localization	Correct identification of the anatomical region.	Pointing accuracy, bounding box IoU	Automated (against ground truth)

Clinical Utility	Impact on the diagnostic workflow.	Accuracy change, time - to - decision	Clinician study (Likert scale)
Completeness	Coverage of all relevant findings.	Finding recall rate	Hybrid (automated + expert)

4. Comparative Analysis and Discussion

4.1. Benchmark Configuration and Multi-Agent Approaches

4.1.1. Evaluation Protocol

Evaluation protocols across the reviewed studies employ a consistent baseline set: zero-shot prompting (direct question-answering without reasoning guidance), standard few-shot prompting (with exemplar input-output pairs), and single-step CoT prompting (with a reasoning trigger). Performance is measured using closed-set accuracy for VQA tasks, BLEU-1/4, METEOR, and ROUGE-L for report generation, and domain-specific metrics including RadFact precision/recall for clinical factual consistency evaluation.

4.1.2. Multi-Agent Collaboration Mechanisms

MDAgents, proposed by Kim et al., introduces an additional dimension of improvement by automatically assigning collaboration structures---solo practitioner, multidisciplinary team (MDT), or integrated care team---based on medical task complexity. On benchmark evaluations, MDAgents achieved best performance on seven out of ten medical benchmarks with up to 11.8% improvement over prior multi-agent methods [15]. The adaptive routing of simple cases to single-agent processing and complex cases to assembled specialist teams demonstrates that multi-agent orchestration can yield meaningful gains for clinical decision support tasks, particularly when diagnostic complexity varies substantially across patient populations and clinical settings.

4.2. Quantitative Performance Analysis

The comparative analysis reveals consistent performance patterns across the reviewed approaches. Bannur et al. introduced MAIRA-2 with RadFact, a factuality evaluation metric for grounded radiology report generation that assesses both precision and recall of clinically relevant statements while accounting for spatial grounding through bounding box annotations [16]. Wu et al. addressed a critical gap in existing CoT approaches by grounding reasoning paths in medical knowledge graphs with MedReason, producing 32,682 QA pairs with verifiable thinking trajectories; MedReason-8B outperformed comparable models by 4.2% on the MedBullets benchmark [17]. The GEMeX benchmark developed by Wang et al., comprising 151,025 CXR images and 1,605,575 QA pairs with both visual and textual explanations, provides the most comprehensive resource for evaluating explainable CXR diagnosis [18].

Table 4 summarizes the performance improvements reported across the reviewed methods.

Table 4. Reported Performance Improvements of CoT and RAG Approaches over Baseline Methods

Method Category	Representative Work	Reported Improvement	Metric Type	Evaluation Scope
Two - stage CoT	Multimodal - CoT	SOTA (sub - 1B model)	Answer accuracy	ScienceQA
Hierarchical CoT	MedCoT	SOTA on 3 benchmarks	Answer accuracy	VQA - RAD, SLAKE, PathVQA

Four - stage CoT	Med - SCoT	Outperforms MedCoT	Accuracy + reasoning quality	VQA - RAD, SLAKE, PathVQA
RAG with factuality control	RULE	+20.8% average	Factual accuracy	3 medical VQA datasets
Domain - aware RAG	MMed - RAG	+43.8% average	Factual accuracy	5 medical datasets
Multi - agent adaptive	MDAgents	+11.8% over prior methods	Task accuracy	7 of 10 medical benchmarks
Grounded report generation	MAIRA - 2	SOTA findings generation	RadFact precision/recall	MIMIC - CXR
KG - grounded reasoning	MedReason	+4.2% over comparable models	Answer accuracy	MedBullets

As shown in Figure 2, performance gains over the zero-shot baseline increase progressively with reasoning complexity across all three benchmarks. RAG-enhanced CoT methods achieve the highest relative improvements, particularly on PathVQA, where pathology-specific knowledge demands are greatest. Hierarchical multi-step CoT approaches consistently outperform single-step CoT, confirming that structured reasoning decomposition contributes independently to diagnostic accuracy. The gap between RAG-enhanced and non-RAG methods is most pronounced on knowledge-intensive tasks, suggesting that retrieval provides complementary benefits beyond what can be achieved through reasoning structure alone.

Figure 2. Comparative Performance Landscape of CoT-Enhanced VLMs Across Medical VQA Benchmarks

Across all evaluated benchmarks, the magnitude of improvement correlates with the depth of domain knowledge required by each task. These results underscore the importance of matching prompting strategy complexity to task demands, and suggest that

selective deployment of RAG may be more efficient than uniform application across all diagnostic scenarios.

4.3. Clinical Implications and Limitations

4.3.1. Regulatory Alignment and Trust Calibration

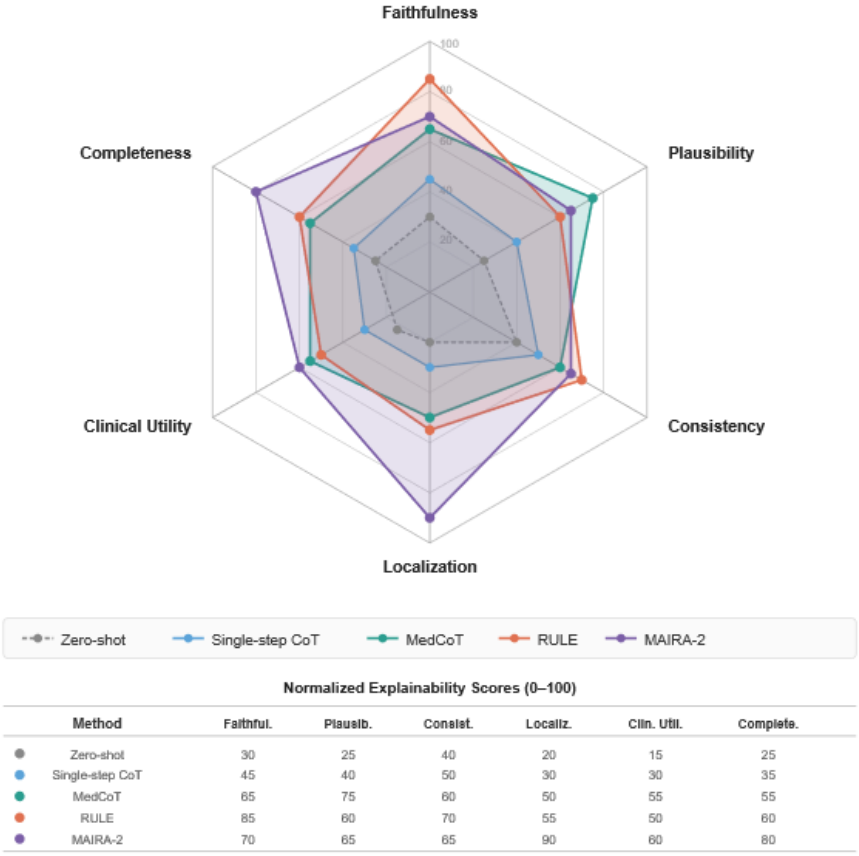
The FDA's evolving framework for AI/ML-enabled medical devices increasingly emphasizes algorithmic transparency, particularly through its 2021 Action Plan for AI/ML-based Software as a Medical Device (SaMD), which calls for approaches that enable continuous performance monitoring and explainability of algorithmic outputs. CoT-based approaches produce traceable reasoning chains that can function as audit trails for diagnostic decisions, potentially aligning with these regulatory expectations for pre-market submissions. However, it should be noted that current FDA guidance does not explicitly mandate CoT-style explanations, and the extent to which reasoning chains satisfy regulatory transparency requirements remains an open question. The ability to attribute each diagnostic conclusion to specific observed findings, retrieved evidence, and reasoning steps provides a form of transparency that conventional deep learning classifiers cannot offer, though the fidelity of these reasoning chains to the model's actual internal decision process requires further empirical validation.

Tayebi Arasteh et al. demonstrated with RadioRAG that real-time retrieval from Radiopaedia enabled LLM performance matching or exceeding radiologist accuracy across multiple subspecialty domains [19]. Wu et al. introduced MedGraphRAG, a graph-based retrieval approach using Triple Graph Construction that outperformed existing methods on nine medical QA benchmarks and two fact-checking benchmarks [20]. Chen et al. developed CheXagent, an instruction-tuned foundation model trained across 28 public datasets that includes fairness evaluation across demographic subgroups [21]. Jain et al. created RadGraph-XL, an expert-annotated dataset of 2,300 radiology reports with entity and relation annotations enabling graph-based evaluation of generated diagnostic explanations [22,23]. These contributions collectively provide the retrieval infrastructure and evaluation resources necessary for rigorous assessment of explainability in clinical AI.

4.3.2. Practical Deployment Considerations

The reviewed evidence indicates important limitations constraining near-term clinical deployment. Performance on curated benchmarks does not necessarily transfer to real-world clinical environments, where image quality varies and patient presentations are atypical [24]. The computational overhead of multi-step reasoning and RAG presents practical challenges: RAG-enhanced pipelines typically incur substantially higher inference latencies than conventional classification models, which may limit their applicability in time-critical clinical workflows.

As indicated by Figure 3, no single method category achieves uniformly superior performance across all six explainability dimensions. Grounded generation approaches demonstrate the strongest localization capability, attributable to their explicit spatial anchoring mechanisms. RAG-enhanced CoT methods exhibit the highest faithfulness scores, reflecting the verifiability conferred by externally retrieved evidence. Hierarchical CoT approaches lead in plausibility, consistent with their design objective of mimicking structured clinical reasoning workflows [25]. These complementary strengths suggest that hybrid architectures integrating grounded generation with retrieval-augmented hierarchical reasoning may be necessary to achieve comprehensive explainability in clinical deployment.



Normalized Explainability Scores (0–100)

Method	Faithful.	Plausib.	Consist.	Localiz.	Clin. Util.	Complete.
Zero-shot	30	25	40	20	15	25
Single-step CoT	45	40	50	30	30	35
MedCoT	65	75	60	50	55	55
RULE	85	60	70	55	50	60
MAIRA-2	70	65	65	90	60	80

Figure 3. Multi-Dimensional Explainability Assessment Across Reviewed Approaches

Retrieval quality, measured by relevance precision, proves more impactful than retrieval volume. As observed in the RULE and MMed-RAG studies [13, 14], improving retrieval precision through domain-specific embedding models produced substantially larger accuracy improvements than simply increasing the number of retrieved passages [26,27]. These findings suggest that precision-focused retrieval engineering should be prioritized over exhaustive knowledge base coverage. The computational-accuracy trade-off remains an active area of investigation, with cached retrieval indices, reasoning chain compression, and complexity-adaptive invocation of multi-step processing showing promise as potential optimization strategies.

5. Conclusion and Future Directions

5.1. Summary of Findings

This review has examined the landscape of CoT prompting and RAG-enhanced approaches for explainable medical image interpretation using VLMs, with chest radiograph analysis as the primary application domain. The proposed two-dimensional analytical framework, organized along reasoning granularity and evidence grounding depth, provides a structured basis for comparing methodologically diverse contributions in this rapidly evolving field. The comparative analysis across representative approaches and multiple benchmark datasets yields findings with implications for both ongoing research and future clinical practice.

Structured multi-step CoT prompting consistently outperforms single-step and zero-shot alternatives, with hierarchical expert verification and four-stage structured reasoning demonstrating meaningful accuracy improvements over baseline methods on standard medical VQA benchmarks. The integration of RAG amplifies diagnostic accuracy, with domain-aware retrieval mechanisms achieving up to 43.8% improvement in factual

accuracy. Multi-agent collaboration frameworks introduce an additional dimension by adaptively assigning task complexity to appropriate specialist configurations, mirroring the multidisciplinary consultation workflows prevalent in clinical radiology practice.

The explainability dimension, assessed across six evaluation criteria encompassing faithfulness, plausibility, consistency, localization, clinical utility, and completeness, reveals that CoT-based approaches offer a qualitatively different form of transparency compared to traditional attribution methods. The explicit reasoning chains generated by these models align more closely with documented clinician preferences for narrative explanations, potentially facilitating trust calibration in clinical settings. The emergence of specialized evaluation frameworks and large-scale explainability benchmarks provides the infrastructure necessary for rigorous comparative assessment within this domain. The convergence of these technical advances with growing regulatory emphasis on algorithmic transparency creates a timely opportunity for advancing explainable AI in medical image interpretation more broadly.

5.2. Future Research Directions

Several directions for future investigation emerge from this review. Domain shift between curated benchmark datasets and real-world clinical images remains a significant concern, and few reviewed approaches have been validated in prospective clinical settings with actual patient populations. The distribution of pathological findings in benchmark datasets frequently differs from real clinical practice, where rare conditions and atypical presentations pose the greatest diagnostic challenges. Investigating how imaging analysis agents, clinical history interpretation agents, and guideline retrieval agents can be orchestrated to simulate multidisciplinary team discussions for complex CXR cases represents a productive research direction that builds on the multi-agent collaboration work reviewed in this paper.

The computational overhead of multi-step reasoning and retrieval-augmented generation presents engineering challenges for environments requiring real-time or near-real-time diagnostic support. Strategies for reducing inference latency without sacrificing reasoning quality, including cached retrieval, reasoning chain compression, and selective invocation based on case complexity, merit focused investigation. The development of standardized evaluation protocols that combine automated metrics with structured clinician feedback would strengthen the evidence base for these approaches. The evaluation of explainability requires continued maturation, as current metrics may not fully capture the nuanced aspects of clinical reasoning that determine the practical utility of AI-generated diagnostic explanations. Longitudinal studies assessing the impact of explainable AI decision support on radiologist diagnostic performance, reading efficiency, and error reduction in routine clinical workflows would provide evidence needed to support broader adoption of these promising approaches in healthcare settings.

References

1. Association of American Medical Colleges, *The complexities of physician supply and demand: Projections from 2021 to 2036*, AAMC, 2024.
2. J. N. Itri, R. R. Tappouni, R. O. McEachern, A. J. Pesch, and S. H. Patel, "Fundamentals of diagnostic error in imaging," *RadioGraphics*, vol. 38, no. 6, pp. 1845--1865, 2018.
3. A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. Y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, 2019.
4. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Henriquez, D. Larson, M. Bethge, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590--597, 2019.
5. C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Advances in Neural Information Processing Systems*, vol. 36, 2024.
6. S. Zhang, Y. Xu, N. Usuyama, J. Bagad, J. Clifton, J. Shen, R. Jain, R. Tinn, H. Poon, and H. Awadalla, "BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, pp. 1--9, 2024.

7. Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *Transactions on Machine Learning Research*, 2024.
8. J. Liu, J. Wang, Z. Liu, D. Gao, X. Chen, B. Jiang, D. Zha, and X. Hu, "MedCoT: Medical chain of thought via hierarchical expert," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16970--16988, 2024.
9. R. Tanno, D. G. T. Barrett, A. Scholten, G. E. Dahl, B. Mustafa, J. Freyberg, A. Albuquerque, D. Tellez, S. Azizi, and V. Natarajan, "Collaboration between clinicians and vision-language models in radiology report generation," *Nature Medicine*, vol. 30, no. 12, pp. 3565--3573, 2024.
10. T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen, "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine," *npj Digital Medicine*, vol. 7, no. 1, p. 20, 2024.
11. T. Kwon, J. C. Ong, D. Kang, S. Moon, Y. Park, M. Song, S. Kim, Y. Kim, and J. Chung, "Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 18417--18425, 2024.
12. Y. Jiang, C. Liu, W. Xie, Y. Yang, and S. Wang, "Med-SCoT: Structured chain-of-thought reasoning and evaluation for enhancing interpretability in medical visual question answering," *Computerized Medical Imaging and Graphics*, vol. 113, p. 102505, 2025.
13. Y. Xia, T. Zheng, Y. Shu, K. Wang, and T. Xie, "RULE: Reliable multimodal RAG for factuality in medical vision language models," *arXiv preprint arXiv:2407.05131*, 2024.
14. Y. Xia, T. Zheng, Y. Shu, K. Wang, and T. Xie, "MMed-RAG: Versatile multimodal RAG system for medical vision language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.
15. J. Kim, M. Lee, S. Park, J. Song, and J. Shin, "MDAgents: An adaptive collaboration of LLMs for medical decision-making," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
16. S. Bannur, S. L. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, J. M. Valanarasu, B. Gu, A. Thieme, and S. Joshi, "MAIRA-2: Grounded radiology report generation," *arXiv preprint arXiv:2406.04449*, 2024.
17. Y. Wu, J. Yang, M. Wang, Z. Liu, M. Li, and H. Yao, "MedReason: Eliciting factual medical reasoning steps in LLMs via knowledge graphs," *arXiv preprint arXiv:2504.00993*, 2025.
18. Y. Wang, Z. Yang, X. He, L. Li, and S. Zhang, "GEMeX: A large-scale, groundable, and explainable medical VQA benchmark for chest X-ray diagnosis," *arXiv preprint arXiv:2411.16778*, 2024.
19. S. Tayebi Arasteh, A. M. Arias-Lorza, F. Jungmann, J. N. Kather, D. Truhn, and S. Nebelung, "RadioRAG: Factual large language models for enhanced diagnostics in radiology using online retrieval augmented generation," *arXiv preprint arXiv:2407.15621*, 2024.
20. Y. Wu, Z. Bi, Y. Yao, J. Chen, H. Wang, and Z. Liu, "Medical graph RAG: Towards safe medical large language model via graph retrieval-augmented generation," *arXiv preprint arXiv:2408.04187*, 2024.
21. Z. Chen, M. Varma, A. Wan, R. Dorfman, C. P. Langlotz, J.-B. Delbrouck, and P. Rajpurkar, "CheXagent: Towards a foundation model for chest X-ray interpretation," *arXiv preprint arXiv:2401.12208*, 2024.
22. S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, H. C. Duber, M. P. Lungren, and C. P. Langlotz, "RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
23. A. B. Rosenkrantz, D. R. Hughes, and R. Duszak, "The U.S. radiologist workforce: An analysis of temporal and geographic variation from 2013 to 2019," *Journal of the American College of Radiology*, vol. 18, no. 5, pp. 796--804, 2021.
24. E. I. Bluth, S. Bansal, C. E. Bender, and C. W. Bolan, "The 2017 ACR Commission on Human Resources workforce survey," *Journal of the American College of Radiology*, vol. 14, no. 12, pp. 1613--1619, 2017.
25. L. H. Garland, "On the scientific evaluation of diagnostic procedures," *Radiology*, vol. 52, no. 3, pp. 309--328, 1949.
26. R. J. McDonald, K. M. Schwartz, L. J. Eckel, F. E. Diehn, C. H. Hunt, B. J. Bartholmai, B. J. Erickson, and D. F. Kallmes, "The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload," *Academic Radiology*, vol. 22, no. 9, pp. 1191--1196, 2015.
27. U.S. Food and Drug Administration, *Artificial intelligence and machine learning (AI/ML)-enabled medical devices*, FDA, 2024.