MARCH: Multi-Agent Radiology Clinical Hierarchy for CT Report Generation

Yi Lin[1], Yihao Ding[2], Yonghui Wu[3], Yifan Peng[1]
[1]Weill Cornell Medicine, New York, USA
[2]University of Western Australia, Crawley, Australia
[3]University of Florida, Florida, USA
Correspondence: yip4002@med.cornell.edu

Abstract

Automated 3D radiology report generation often suffers from clinical hallucinations and a lack of the iterative verification found in human practice. While recent Vision-Language Models (VLMs) have advanced the field, they typically operate as monolithic "black-box" systems without the collaborative oversight characteristic of clinical workflows. To address these challenges, we propose MARCH (**M**ulti-**A**gent **R**adiology **C**linical **H**ierarchy), a multi-agent framework that emulates the professional hierarchy of radiology departments and assigns specialized roles to distinct agents. MARCH utilizes a *Resident Agent* for initial drafting with multi-scale CT feature extraction, multiple *Fellow Agents* for retrieval-augmented revision, and an *Attending Agent* that orchestrates an iterative, stance-based consensus discourse to resolve diagnostic discrepancies. On the RadGenome-ChestCT dataset, MARCH significantly outperforms state-of-the-art baselines in both clinical fidelity and linguistic accuracy. Our work demonstrates that modeling human-like organizational structures enhances the reliability of AI in high-stakes medical domains.

1 Introduction

The interpretation of medical imaging, particularly three-dimensional (3D) volumetric data like chest Computed Tomography (CT), remains a cornerstone of modern diagnostic medicine (Moor et al., 2023). Despite its importance, generating radiology reports that are accurate, comprehensive, and clinically valid remains cognitively demanding and represents a primary bottleneck in clinical workflows. While Large Language Models (LLMs) and Vision-Language Models have shown promise in automating radiology report generation (Ma et al., 2025), these approaches often exhibit clinical *hallucinations*, struggle to detect subtle pathological findings in sparse 3D data, and lack the iterative verification and cross-checking (Zhu et al., 2025).

In radiology, a well-established strategy to reduce such interpretive errors is to provide structured cognitive support via a *devil's advocate* role, commonly implemented through *overread* or *read-out* sessions (Seah et al., 2021; Waite et al., 2017). Typically, a *Resident* performs the initial interpretation and drafts a report, which is then independently reviewed and, if necessary, reinterpreted by a *Fellow*. When discrepancies persist, an *Attending* radiologist conducts a parallel read to adjudicate and finalize the report. This hierarchical, verification-driven workflow improves diagnostic accuracy of the final report, supports continuous learning through targeted feedback, and enables timely updates to patient management (Hill et al., 2017). By contrast, most existing automated report generation systems do not model this multi-agent review process, and instead rely on end-to-end, black-box generation (Wang et al., 2023a).

To bridge this gap, we propose MARCH (**M**ulti-**A**gent **R**adiology **C**linical **H**ierarchy), a consensus-driven multi-agent framework that explicitly models the hierarchical and collaborative structure of radiology *read-out* sessions (Du et al., 2025; Liao et al., 2025; Dang et al., 2025). Specifically, MARCH framework is organized into three layers: (1) *Resident Agent* and *Fellow Agent* that generates a detailed initial draft from volumetric scans utilizing a 3D vision encoder coupled with a multi-region segmentation module; (2) *Retrieval Agents* and *Fellow Agents* that jointly refine the draft by grounding it in evidence-based findings retrieved from an expansive clinical database through multi-modal similarity search; and (3) *Attending Agent* that moderates an iterative consensus-driven discourse.

Unlike conventional sequence-to-sequence approaches that merely average outputs, MARCH simulates multi-round "clinical meetings" in which agents explicitly agree, correct, or refine interpretations until a clinically coherent consensus is reached, thereby naturally aligning with the cogni-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 273–285

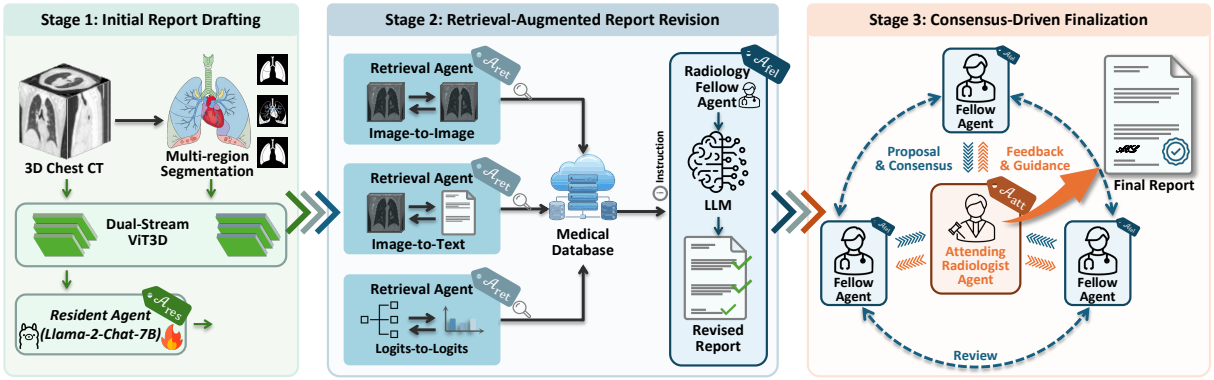

Figure 1: Overview of the MARCH framework. It consists of three main stages: 1) Initial Report Drafting, 2) Retrieval-Augmented Report Revision, and 3) Consensus-Driven Finalization.

tive structure of decision-making and verification in radiology (Goold and Stern, 2006; Bolin, 2006).

Our primary contributions are summarized as follows: (i) We present MARCH, a hierarchical multi-agent framework that explicitly models the Resident-Fellow-Attending workflow for automated radiology report generation. (ii) We introduce a retrieval-augmented revision stage that utilizes image-, text-, and logit-based indices to improve diagnostic grounding, coherence, and clinical fidelity. (iii) We design a consensus-driven finalization protocol in which multiple agents iteratively resolve diagnostic discrepancies through stance-based discourse. (iv) Comprehensive empirical analyses and ablation studies demonstrate that MARCH significantly outperforms state-of-the-art baselines on both reference-based and clinical validity metrics.

2 Related Work

LLMs have demonstrated remarkable capabilities in various medical applications (Singhal et al., 2025; Goh et al., 2024; Wang et al., 2025; Yamamoto et al., 2024; Ma et al., 2025; Liu et al., 2024a). Notably, general-purpose models such as GPT-4 (OpenAI et al., 2024) have been shown to outperform medical students on standardized medical exams. Building on this success, recent studies have explored the use of LLMs for generating medical reports from imaging data (Sloan et al., 2024). These approaches typically involve fine-tuning LLMs with paired image-text datasets (Johnson et al., 2019) or employing prompt engineering techniques to employ large pre-trained models (Wang et al., 2023b). These methods have achieved promising results in zero-shot or few-shot settings (Liu et al., 2024b). However, these methods continue to face challenges such as diverse medical image modalities, complex medical terminology, and clinically unsafe hallucinations (Liu et al., 2025; Jiang et al., 2025).

Multi-agent systems extend LLMs by decomposing complex tasks into smaller, manageable subtasks (Khan et al., 2025; Plaat et al., 2025). In the medical domain, such systems have been applied to medical diagnosis (Fan et al., 2025; Kim et al., 2024), treatment recommendation (Chen et al., 2025a), and medical image analysis (Li et al., 2024). These systems typically integrate task planning, knowledge retrieval, and response generation components to enable structured reasoning and improved reliability (Zhong et al., 2025; Yi et al., 2025). Despite these advances, the application of multi-agent systems for 3D radiology report generation remains underexplored.

In contrast to these approaches, MARCH addresses the inherent complexity of 3D radiology report generation by distributing heterogeneous tasks across a multi-agent hierarchy equipped with both specialized medical tools (e.g., trainable models) and reasoning capabilities (LLMs). We pivot from monolithic models to a modular, multi-agent framework in which agents adjudicate conflicting findings through dynamic, multi-round negotiation. This design not only enhances interpretability and reliability but also enables seamless integration of domain-specific knowledge, thereby setting a new benchmark in 3D report generation.

3 Methodology

The proposed MARCH framework operates through three interrelated phases: (1) Initial Report Drafting, (2) Retrieval-Augmented Report Revision, and (3) Consensus-Driven Finalization (Figure 1).

3.1 Initial Report Drafting

In this stage, we instantiate a *Resident Agent* $\mathcal{A}_{\text{res}}$ to generate a preliminary radiology report draft $\mathcal{T}$ from chest CT scans $\mathcal{I}$. The agent is trained on a large-scale corpus of paired volumetric CT scans and reports to learn the cross-modal alignment between visual pathology and textual descriptions.

To mitigate the sparsity of abnormal findings in volumetric data, $\mathcal{A}_{\text{res}}$ embeds a multi-region segmentation module based on the SAT (Segment Anything with Text) model (Zhao et al., 2025). This module partitions $\mathcal{I}$ into ten anatomical subregions (e.g., bone, breast), allowing the encoder to attend to localized anatomical and pathological entities.

Formally, the report is generated as $\mathcal{T} = \mathcal{A}_{\text{res}}(\mathcal{I}; \theta_{\text{res}})$, where θ_{res} are the learned parameters. Our implementation utilizes a frozen dual-stream ViT3D backbone pre-trained on RadFM (Wu et al., 2025) for spatial feature extraction and LlaMA-2-Chat-7B (Touvron et al., 2023) optimized via LoRA (Hu et al., 2021) for text generation.

3.2 Retrieval-Augmented Report Revision

To mitigate omissions and hallucinations, the *Retrieval Agent* $\mathcal{A}_{\text{ret}}$ identifies relevant clinical context from a training database $\mathcal{D}$. We propose three retrieval paradigms: (i) **Image-to-Image** & **Image–to-Text retrieval**, which uses a 3D vision encoder to retrieve visually similar CT volumes and their corresponding reports from $\mathcal{D}$. (ii) **Logits-based retrieval**, where a classification head atop $\mathcal{A}_{\text{res}}$ predicts 18 canonical clinical abnormalities (e.g., pleural effusion, atelectasis), and these logits are used to retrieve reports with similar diagnostic profiles. In this paper, each retrieval agent retrieves top-3 cases and concatenates them into a structured retrieved evidence $\mathcal{R} = \mathcal{A}_{\text{ret}}(\mathcal{I}, \mathcal{D})$ and then provided to a *Fellow Agent* $\mathcal{A}_{\text{fel}}$, which refines the initial draft by validating findings and resolving inconsistencies, producing an enhanced report $\mathcal{T}'$: $\mathcal{T}' = \mathcal{A}_{\text{fel}}(\mathcal{T}, \mathcal{R})$.

3.3 Consensus-Driven Finalization

The final stage employs a multi-round collaborative protocol orchestrated by an *Attending Agent* $\mathcal{A}_{\text{att}}$ to ensure that the final report reaches a clinical consensus among multiple specialized fellow agents $\{\mathcal{A}_{\text{fel},i}\}_{i=1}^{N}$, where N is the number of fellows.

Round 1: Consensus Synthesis. $\mathcal{A}_{\text{att}}$ first aggregates the enhanced reports from all fellows to generate an initial consensus report $\mathcal{T}^{(0)}$ and identifies potential clinical conflicts:

$$\mathcal{T}^{(0)} = \mathcal{A}_{\text{att}}(\{\mathcal{T}'_i\}_{i=1}^{N}). \tag{1}$$

Round $t+1$: Iterative Refinement. In subsequent rounds, each fellow $\mathcal{A}_{\text{fel},i}$ reviews the current consensus $\mathcal{T}^{(t)}$ and provides a stance $S_i^{(t)}$, indicating agreement, proposing corrections, or adding supplementary observations:

$$S_i^{(t)} = \mathcal{A}_{\text{fel},i}(\mathcal{T}'_i, \mathcal{T}^{(t)}). \tag{2}$$

The attending agent $\mathcal{A}_{\text{att}}$ then ensembles these stances to update the report:

$$\mathcal{T}^{(t+1)} = \mathcal{A}_{\text{att}}(\mathcal{T}^{(t)}, \{S_i^{(t)}\}_{i=1}^{N}). \tag{3}$$

This iteration continues until the attending agent determines that a stable consensus or a predefined maximum number of rounds T has been reached.

4 Experimental Setting

We evaluate MARCH on the RadGenome-ChestCT dataset (Wu et al., 2025), which contains 25,692 chest CT scans from 21,304 patients. Each CT scan is accompanied by a detailed radiology report authored by experienced radiologists. We adhere to the official data split, using 24,128 scans for training and 1,564 scans for testing. Statistics of the dataset are summarized in Appendix B.

We use GPT-4.1 and GPT-4o as the LLM backbones for *Fellow* and *Attending Agent*, respectively, with a temperature of 0 to ensure deterministic outputs. The *Resident* and *Retrieval Agent* are implemented using the HuggingFace Transformers library (Wolf et al., 2020) and trained on a single NVIDIA H100 GPU. The training process employs the AdamW optimizer with a learning rate of $1e-5$ and a batch size of 1. We train MARCH for 10 epochs, which takes approximately 40 hours. Appendix A lists the template prompts for each agent.

We assess MARCH using standard natural language generation metrics that evaluate both lexical and semantic alignment with reference texts, including BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). In addition, we assess clinical validity of the generated reports using the Clinical Efficacy (CE) score, which measures the accuracy of 18 predefined clinical abnormalities by computing precision, recall, and F1 scores with a pretrained RadBERT-RoBERTa-4m model (Yan et al., 2022).

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CE-Precision	CE-Recall	CE-F1
R2GenPT (2023b)	0.433	0.341	0.282	0.242	0.399	0.323	0.340	0.066	0.110
MedVInT (2024)	0.443	0.349	0.288	0.246	0.404	0.326	0.377	0.148	0.212
CT2Rep (2024)	0.444	0.344	0.279	0.236	0.402	0.310	0.317	0.089	0.139
M3D (2024)	0.436	0.345	0.285	0.245	0.400	0.326	0.407	0.090	0.148
RadFM (2025)	0.442	0.345	0.281	0.237	0.399	0.315	0.382	0.131	0.195
Reg2RG (2025b)	0.473	0.365	0.296	0.249	0.441	0.367	0.423	0.181	0.253
MARCH (**Ours**)	**0.482**	**0.375**	**0.305**	**0.257**	**0.456**	**0.383**	**0.495**	**0.335**	**0.399**

Table 1: Comparison of MARCH against state-of-the-art methods on RadGenome-ChestCT.

Method	BLEU-1	BLEU-4	METEOR	CE-F1
Resident-only	0.469	0.246	0.435	0.219
SR-SA	0.476	0.250	0.447	0.332
SR-MA	0.475	0.251	0.454	0.352
MR-MA	0.479	0.255	0.456	0.362
Ours	**0.482**	**0.257**	**0.456**	**0.399**

Table 2: Ablation study of components in MARCH. SR-SA: Single Round Single Agent; SR-MA: Single Round Multi-Agent; MR-MA: Multi-Round Multi-Agent.

LLM	BLEU-1	BLEU-4	METEOR	CE-F1
Resident-only	0.469	0.246	0.435	0.219
GPT-4.1-mini	0.480	0.255	0.454	0.393
GPT-4.1	0.482	0.257	0.456	0.399
GPT-4o	0.479	0.255	0.454	0.392
GPT-5	0.480	0.255	0.454	0.391

Table 3: Performance comparison of different LLMs.

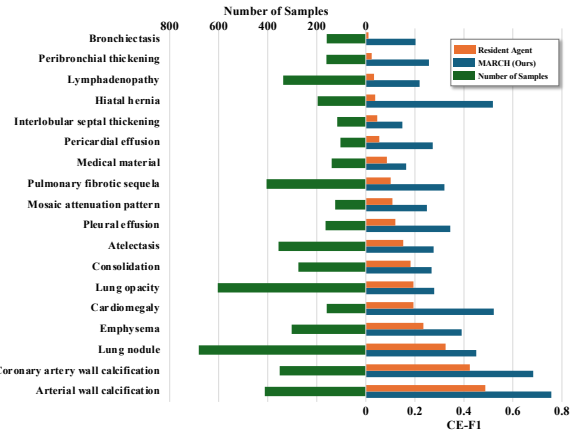

Figure 2: Clinical efficacy across various abnormalities.

5 Results and Discussions

Comparison with State-of-the-Art Methods. We benchmark MARCH against several state-of-the-art approaches for medical report generation (Table 1). Across all evaluation metrics, MARCH consistently outperforms all baseline methods, demonstrating superior performance in generating high-quality and clinically accurate radiology reports.

Ablation Study. To assess the contribution of each component in MARCH, we conduct an ablation study by removing or modifying key elements of the model. Table 2 indicates that each component significantly contributes to the overall performance, with the most notable drop observed when removing the consensus-driven Finalization.

Sensitivity across LLMs. To investigate the impact of LLM size on report generation, we evaluate MARCH using different LLM backbones. Table 3 shows all variants of MARCH consistently outperform baseline methods, with marginal performance improvements observed across advanced LLMs.

Clinical Efficacy. Figure 2 presents the clinical efficacy of MARCH in F1-score across 18 abnormalities. Compared with the baseline method, Resident Agent, MARCH demonstrates superior performance across most abnormalities, particularly in detecting minor abnormalities such as "hiatal hernia" and "pericardial effusion", indicating its enhanced ability to identify subtle clinical findings. Detailed analysis is provided in Appendix E.

Case studies. To further elucidate MARCH's capacity for generating interpretable and clinically grounded reports, we present a case study in Figure 3 (Appendix C), which traces the hierarchical, multi-agent workflow intrinsic to our approach, with findings organized by anatomical region.

Additional Results. We further provide additional experimental results, including the effectiveness of agent number (Appendix D) and examples of generated reports (Appendix F).

6 Conclusions

This paper presents MARCH, a consensus-driven, multi-agent framework to reduce cognitive errors in interpreting abnormal CT findings. In contrast to prior work, MARCH coordinates agents for multi-scale 3D feature extraction, evidence-based retrieval augmented generation, and iterative consensus discourse. Empirical evaluations demonstrate that MARCH significantly outperforms state-of-the-

art models on both linguistic metrics and clinical fidelity, generates reports that reduce the risk of single-reader misinterpretation, and supports transparent, collaborative report generation.

Limitations

This method has demonstrated substantial improvements in medical image report generation. However, several limitations remain to be addressed in future work. First, our current evaluation primarily utilizes the GPT series of large language models for multi-agent reasoning. Exploring the generalizability of this clinical hierarchy using diverse open-source or domain-specific medical LLMs is a critical next step. Second, MARCH currently lacks a long-term memory mechanism, which limits its ability to incorporate longitudinal patient history or learn from past diagnostic errors across different cases. Finally, while the framework emulates human clinical workflows, it operates as a fully autonomous system without a hybrid human-agent interface. Future iterations should investigate "human-in-the-loop" configurations where agents provide preliminary consensus reports for radiologist review and incorporate real-time clinical feedback to further bridge the gap between AI assistance and professional practice.

Acknowledgements

This work was supported by the National Institutes of Health [grant numbers R01CA289249], the National Science Foundation (NSF) [grant numbers 2145640], the Patient-Centered Outcomes Research Institute (PCORI) [grant numbers ME-2023C3-35934], and the Advanced Research Projects Agency for Health (ARPA-H) [grant name PARADIGM]. We gratefully acknowledge the support of NVIDIA Corporation and the NVIDIA AI Technology Center (NVAITC) UF program.

References

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. 2024. M3D: Advancing 3D medical image analysis with multi-modal large language models. *arXiv [cs.CV]*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Jane Nelson Bolin. 2006. Strategies for incorporating professional ethics education in graduate medical programs. *The American Journal of Bioethics*, 6(4):35–36.

Kai Chen, Ji Qi, Jing Huo, Pinzhuo Tian, Fanyu Meng, Xi Yang, and Yang Gao. 2025a. A self-evolving framework for multi-agent medical consultation based on large language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. 2025b. Large language model with region-guided referring and grounding for CT report generation. *IEEE transactions on Medical Imaging*, 44(8):3139–3150.

Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. 2025. Multi-agent collaboration via evolving orchestration. In *NeurIPS*.

Zhuoyun Du, Chen Qian, Wei Liu, Zihao Xie, Yifei Wang, Rennai Qiu, Yufan Dang, Weize Chen, Cheng Yang, Ye Tian, Xuantang Xiong, and Lei Han. 2025. Multi-agent collaboration via cross-team orchestration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10386–10406, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, and 1 others. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 7(10):e2440969–e2440969.

Susan Dorr Goold and David T Stern. 2006. Ethics and professionalism: what does a resident need to learn? *The American Journal of Bioethics*, 6(4):9–17.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. 2024. CT2Rep: Automated radiology report generation for 3D medical imaging. In *Medical Image Computing and Computer Assisted Intervention*, Lecture notes in computer science, pages 476–486. Springer Nature Switzerland, Cham.

Katherine A Hill, Mohini Dasari, Eliza B Littleton, and Giselle G Hamad. 2017. How can surgeons facilitate resident intraoperative decision-making? *The American Journal of Surgery*, 214(4):583–588.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen,

and Others. 2021. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, pages 1–26.

Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. 2025. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. 2025. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and 1 others. 2024. MMedAgent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760.

Callie C Liao, Duoduo Liao, and Sai Surya Gadiraju. 2025. AgentMaster: A multi-agent conversational framework using A2A and MCP protocols for multimodal information retrieval and analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 52–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8, pages 1–8.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024a. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18635–18643.

Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. 2025. Argus: benchmarking and enhancing vision-language models for 3d radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16448–16460.

Rui Liu, Mingjie Li, Shen Zhao, Ling Chen, Xiaojun Chang, and Lina Yao. 2024b. In-context learning for zero-shot medical report generation. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 8721–8730.

DongAo Ma, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. 2025. A fully open AI foundation model applied to chest radiography. *Nature*, pages 1–11.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, USA. Association for Computational Linguistics.

Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. 2025. Agentic large language models, a survey. *arXiv preprint arXiv:2503.23037*.

Jarrel C Y Seah, Cyril H M Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, Ben Hachey, Stephen J F Hogg, Benjamin P Johnston, Christine Bennett, Luke Oakden-Rayner, Peter Brotchie, and Catherine M Jones. 2021. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit. Health*, 3(8):e496–e506.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. 2024. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton

Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv [cs.CL]*.

Stephen Waite, Jinel Scott, Brian Gale, Travis Fuchs, Srinivas Kolla, and Deborah Reede. 2017. Interpretive error in radiology. *AJR Am. J. Roentgenol.*, 208(4):739–749.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023a. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023b. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1(3):100033.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2025. Accelerating clinical evidence synthesis with large language models. *npj Digital Medicine*, 8(1):509.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications*, 16(1):7866.

Akira Yamamoto, Masahide Koda, Hiroko Ogawa, Tomoko Miyoshi, Yoshinobu Maeda, Fumio Otsuka, Hideo Ino, and 1 others. 2024. Enhancing medical interview skills through ai-simulated patient interactions: nonrandomized controlled trial. *JMIR medical education*, 10(1):e58753.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. RadBERT: Adapting transformer-based language models to radiology. *Radiol. Artif. Intell.*, 4(4):e210258.

Ziruo Yi, Ting Xiao, and Mark V Albert. 2025. A multimodal multi-agent framework for radiology report generation. *arXiv preprint arXiv:2505.09787*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Development of a large-scale medical visual question-answering dataset. *Communications Medicine volume*, 4(1):277.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Large-vocabulary segmentation for medical images with text prompts. *NPJ Digital Medicine*, 8(1):566.

Zhusi Zhong, Yuli Wang, Jing Wu, Wen-Chi Hsu, Vin Somasundaram, Lulu Bi, Shreyas Kulkarni, Zhuoqi Ma, Scott Collins, Grayson Baird, and 1 others. 2025. Vision-language model for report generation and outcome prediction in CT pulmonary angiogram. *NPJ Digital Medicine*, 8(1):432.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. 2025. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769.

Appendix

A Example Prompt Template

Here we present an example of the prompt template used for the Resident Agent $\mathcal{A}_{\text{res}}$ in the Initial Report Drafting, Retrieval-Augmented Revision, and Consensus-Driven Finalization phases of MARCH.

Prompt 1: Resident Agent ($\mathcal{A}_{\text{res}}$) for Initial Report Drafting

```
The global information is provided as the context:
<image_1> <image_2> ... <image_n>.
The region 1 is
<region_1> <region_2> ... <region_n>.
The region 2 is
<region_(n+1)> <region_(n+2)> ... <region_2n>.
...
The region 10 is
<region_(9n+1)> <region_(9n+2)> ... <region_10n>.

Given the provided global and regional information from this CT scan, please generate a comprehensive medical report for each region. First, identify the anatomical area corresponding to each region, then provide detailed information about these anatomical structures and any abnormalities that are essential. You can refer to the global information as the context and take it as a supplement.
```

Prompt 2: Fellow Agent ($\mathcal{A}_{\text{fel}}$) for Consensus Synthesis

```
You are an experienced medical doctor. Your task is to review and analyze a patient's medical report based on the retrieved relevant medical reports from the database. You need to carefully compare the initial medical report with the retrieved relevant medical reports, identify any discrepancies or inconsistencies, and make necessary modifications to ensure the accuracy and coherence of the final medical report.

Please follow these steps to complete your task:
1. Carefully read and understand both the initial medical report and the retrieved relevant medical reports.
2. Identify any discrepancies, conflicts, or inconsistencies between the two reports.
3. Make necessary modifications to the initial medical report to resolve any identified issues, ensuring that the final report is clinically accurate and coherent.
4. If no changes are necessary, output the original initial medical report as is.

Here is the initial medical report: <init_report>

Here are the retrieved relevant medical reports from the database: <retrieved_report>

Here is an example of the format you should output: {"report": "The region 0 is abdomen: ... The region 1 is bone: ... The region 2 is breast: ..."}.
Respond in JSON format without any additional content.
```

Prompt 3: Attending Agent ($\mathcal{A}_{\text{att}}$) for Consensus Synthesis

```
You are an authoritative expert in the medical field. You are organizing a collaborative consultation. Now several doctors have made patient's medical report based on the retrieved relevant medical reports from the database. Your task is to analyze the rationality of each doctor's opinion, summarize the opinions to obtain a synthesized report for the patient and give your final medical report.

First, please read the patient's initial medical report, as follows: <init_report>

Then, all doctors make a revised medical report based on the retrieved relevant medical reports from the database. The following are their opinions: <doctor_info> You need to read all doctors' opinions carefully and analyze whether their opinions make sense.

Next, please write a synthesized report including the following:
1. Final medical report after your analysis.
2. A List of supporting evidence represented as a string. Please output detailed content from some of the evidence provided by the doctors or some analysis results of the doctors. Do not just list the doctor's name.

Please follow the guidelines below to complete your task:
1. Carefully read and understand both the initial medical report and the retrieved relevant medical reports.
2. Identify any discrepancies, conflicts, or inconsistencies between the two reports.
3. Make necessary modifications to the initial medical report to resolve any identified issues, ensuring that the final report is clinically accurate and coherent.
4. If no changes are necessary, output the original initial medical report as is.

Here is an example of the format you should output: {"report": "The region 0 is abdomen: ... The region 1 is bone: ... The region 2 is breast: ...", "reasons": ["...", "..."]}.
Respond in JSON format without any additional content.
```

Prompt 4: Fellow Agent for Iterative Refinement

```
You are an experienced radiology fellow participating in a consultation with several other medical doctors for a patient. The attending radiologist of this consultation has generated a revised report based on all doctors' analysis of the patient. Please provide your viewpoint on his opinion.

Here is the relevant medical knowledge: retrieved_report

Here is the initial medical report: init_report

Here is your last analysis of the patient, which is not completely reasonable, and you may need to adjust it based on the attending radiologist's opinion: fellow_report

Here is the revised report generated by the attending radiologist: attending_report

Here are the reasons provided by the attending radiologist: attending_reason

You need to consider the attending radiologist's opinion carefully and provide your opinions on the revised report generated by the attending radiologist. Please output your opinions including the following content:
1. Your viewpoint on the opinion of the attending radiologist, i.e., respond with "agree" or "disagree".
2. The confidence score of your opinion, respond with an integer between 1 and 3. The meaning of the confidence score is as follows:
3 for High - You are an expert in the subject area and have extensive knowledge in the medical domain. You are highly confident in your ability to provide an accurate and thorough assessment. Your evaluation is based on deep expertise and a comprehensive understanding of the work.
2 for Moderate - You have a good understanding of the subject area and is familiar with the medical domain. You feel confident in your ability to accurately assess the quality and significance of the work. Your evaluation is based on a solid grasp of the content and context.
1 for Low - You have some knowledge of the subject area and is somewhat familiar with the medical domain. You understand the main points but may lack depth in certain areas. You are reasonably confident in your assessment but acknowledges some limitations in your expertise.
3. The reason for your opinion. If you change your opinion, for example, you agree with the attending radiologist's opinion which is different from your last analysis, please respond that you have changed in your response and provide detailed reasons for the change. You need to point out the parts where you got the wrong conclusion or the important parts you ignored in your last analysis, and the new key features that you think are important and the impact of these features on the patient's clinical abnormalities.
4. The evidence you use to support your opinion. Please choose from the relevant medical knowledge I provide as your evidence, and must output important content from the evidence.

Here are examples of the format you should output:
"answer": "agree", "confidence": 3, "reason": "The reason for your opinion.", "evidences": ["Evidence 1 ...", "Evidence 2 ..."],
"answer": "disagree", "confidence": 1, "reason": "The reason
```

for your opinion.", "evidences": ["Evidence 1 ...", "Evidence 2 ..."]
Respond in JSON format without any additional content.

Prompt 5: Attending Agent ($\mathcal{A}_{\text{att}}$) for Iterative Refinement

You are an authoritative expert in the medical field. You are organizing a collaborative consultation. Now several doctors have made analysis and judgments on a your previous report. Your task is to judge whether everyone has reached a consensus on the medical report based on the analysis statements of each doctor and then analyze the rationality of each doctor's opinion and give your final medical report.

In the previous discussion, you took into account the opinions of all the doctors and obtained a report about the patient, which is listed as follows: <current_report>

In response to the patient's synthesized report, several doctors put forward their own opinions and reasons. In each doctor's statement, they first express whether they agree with your statement in the previous synthesized report and give the confidence level on their own judgment. Then, they further elaborate on their views by stating reasons and listing relevant references. The following are their opinions: <fellow_info>

Now, you need to judge whether the next round of discussion is needed based on each doctor's statement. Considering the following four cases:
1. If all doctors agree with the previous synthesized report, there is no need to continue the discussion.
2. If some doctors disagree with the previous report, but they are not confident in their judgment and have not listed convincing evidence, there is no need to continue the discussion.
3. If some doctors strongly oppose the previous report and you think their evidence is worth discussing, please continue the discussion.
4. If most doctors disagree with the previous report, please continue the discussion.

If you think the discussion should continue, you need to analyze the rationality of each doctor's opinions and summarize the opinions you think are reasonable to obtain a synthesized report for the patient. You should follow these cases:
1. If a doctor expresses strong opposition to your previous report, you need to focus on hisher reasons and arguments and think carefully about whether you need to reconsider the diagnosis of the patient and modify your synthesized report accordingly.
2. If a doctor expresses opposition but also has some doubts about hisher own opinion, you need to consider hisher opinion, but you can stick to your original opinion. 3. If a doctor expresses agreement, then you do not need to modify your original synthesized report based on hisher opinion.
Please output the following four contents: 1. Whether to continue the discussion or not. Please respond with 'Yes' or 'No'.
2. Your revised report.
3. Your reasons for revision. The format of the reason for revision is: which doctor's opinion or relevant literature you refer to, and which original opinions you modify. Please output detailed content, don't just list the doctor's name.
4. The instructions for each doctor to follow in the next round of discussion. The instruction should be specific and actionable, guiding each doctor on how to adjust their analysis or what aspects to focus on based on the previous round's discussion.

When you output the revised report, please follow the guidelines below to complete your task:
1. Carefully read and understand both the initial medical report and the retrieved relevant medical reports.
2. Identify any discrepancies, conflicts, or inconsistencies between the two reports.
3. Make necessary modifications to the initial medical report to resolve any identified issues, ensuring that the final report is clinically accurate and coherent.
4. If no changes are necessary, output the original initial medical report as is.

Here are two examples of the format you should output:
{"action": "No", "report": "The region 0 is abdomen: ... The region 1 is bone: ... The region 2 is breast: ...", "reasons": ["reason1...", "reason2..."], "instructions": ["instruction1...", "instruction2..."]}.
{"action": "Yes", "report": "The region 0 is abdomen: ... The region 1 is bone: ... The region 2 is breast: ...", "reasons": ["reason1...", "reason2..."], "instructions": ["instruction1...", "instruction2..."]}.
Respond in JSON format without any additional content.

Characteristics	Train	Test
Number of CT scans	24,128	1,564
Number of patients	20,000	1,304
Age (mean±std, years)	48.74±17.28	48.39±16.87
Sex (M/F)	14,097/10,028	910/654
Regions		
Abdomen	23,553	1,518
Bone	23,479	1,509
Breast	1,080	58
Heart	23,289	1,433
Esophagus	20,693	1,328
Lung	23,741	1,514
Mediastinum	23,684	1,523
Pleura	18,156	1,172
Thyroid	1,093	51
Trachea/bronchi	21,951	1,417
Clinical Abnormalities		
Arterial wall calcification	6,607	423
Atelectasis	6,005	359
Bronchiectasis	2,341	163
Cardiomegaly	2,533	159
Consolidation	4,066	280
Coronary artery wall calcification	5,747	348
Emphysema	4,558	304
Hiatal hernia	3,391	197
Interlobular septal thickening	1,887	119
Lung nodule	10,999	697
Lung opacity	8,944	607
Lymphadenopathy	5,839	343
Medical material	2,846	149
Mosaic attenuation pattern	1,788	124
Peribronchial thickening	2,566	178
Pericardial effusion	1,641	106
Pleural effusion	2,628	179
Pulmonary fibrotic sequela	6,175	399

Table 4: Statistics of the RadGenome-Chest CT dataset.

B Statistics of Datasets

In Table 4, we provide detailed statistics of the RadGenome-ChestCT dataset used in our experiments, including the number of samples, patient demographics, and the distribution of region-specific reports and clinical abnormalities.

The dataset contains a total of 25,692 3D chest CT scans from 21,304 unique patients. Following the dataset's standard split, 24,128 scans are allocated for training and 1,564 for testing. Demographics are consistent across both sets, with a mean age of approximately 48.7 years (±17.2). The cohort includes 15,007 male and 10,682 female cases. This dataset is de-identified and publicly available, and its use has been approved by the

Institutional Review Board (IRB) of the institutions involved in its creation.

The dataset provides high-granularity reports across 10 anatomical regions, ensuring comprehensive spatial coverage. The most prevalent regions are the mediastinum, lung, and abdomen, each appearing in more than 25,000 scans. Additionally, the reports are annotated via RadBERT-RoBERTa-4m (Yan et al., 2022) with 18 distinct abnormalities of varying clinical prevalence. Lung nodules are the most frequent pathology ($n = 11,696$), whereas cardiovascular findings such as arterial wall calcification ($n = 7,030$) and pulmonary abnormalities such as lung opacity ($n = 9,551$) provide a diverse range of diagnostic targets.

C Case Study

To demonstrate MARCH's capacity for generating interpretable, clinically grounded reports, we present a case study in Figure 3, which follows our hierarchical, multi-agent workflow and organizes findings by anatomical region.

Stage 1: Initial Report Drafting. The Resident Agent $\mathcal{A}_{\text{res}}$ first generates a draft report directly from the CT images. It describes each predefined region. In this example, $\mathcal{A}_{\text{res}}$ reports largely normal findings across the abdomen (e.g., normal adrenal glands and upper abdominal organs), bones (preserved vertebral body heights), esophagus (no abnormal wall thickening), heart and mediastinum (normal contours and vessels, no effusion), lungs (no nodular/infiltrative lesion), pleura (no effusion/thickening), and airway (trachea and main bronchi patent), while also noting gynecomastia and recommending thyroid ultrasound correlation.

Stage 2: Retrieval-Augmented Report Revision. Next, the Retrieval Agent $\mathcal{A}_{\text{ret}}$ query a reference database for visually and semantically similar studies using complementary strategies, including *Image-to-Image retrieval*, *Image-to-Text retrieval*, and *Logits-based retrieval*. Conditioned on this retrieved evidence, multiple Fellow Agents $\mathcal{A}_{\text{fel}}$ refine the draft. This step encourages diagnostic diversity and completeness. In the shown case, fellows add or sharpen clinically relevant details (e.g., specifying no focal liver lesion within the imaged field, identifying a few millimetric nonspecific pulmonary nodules, describing pleuroparenchymal sequelae densities, refining pleural phrasing, and noting diffuse mild bronchial ectasia with peribronchial thickening).

Number	BLEU-1	BLEU-4	METEOR	CE-F1
1	0.473	0.253	0.451	0.323
3	0.470	0.255	0.456	0.330
5	0.476	0.257	0.455	0.335
10	0.473	0.254	0.455	0.337
20	0.475	0.255	0.454	0.327

Table 5: Effectiveness of different *Fellow Agent* $\mathcal{A}_{\text{fel}}$ numbers on the subset of RadGenome-ChestCT.

Stage 3: Consensus-Driven Finalization.
Round 1: Consensus Synthesis. The Attending Agent $\mathcal{A}_{\text{att}}$ subsequently consolidates these revisions into a single report and produces an explicit feedback report comprising: (i) key findings, (ii) diagnostic rationales that trace which fellows introduced each modification (e.g., additional lung and bronchial findings not present in the initial draft), and (iii) targeted suggestions to guide the next round of improvement.
Round $t+1$: Iterative Refinement. Finally, fellows update their reports based on the attending's critique, specifically addressing localized findings such as bronchial wall thickening or hepatic textures. Each fellow articulates a formal stance, including agreement confidence, reasons, and evidentiary support. The attending adjudicates any remaining discrepancies, and once consensus is reached, terminates the discussion and releases the finalized report. In this example, the final report preserves the resident's normal baseline findings while incorporating consensus additions (e.g., nonspecific millimetric lung nodules, pleuroparenchymal sequelae, and bronchial ectasia/peribronchial thickening) and maintaining recommendations such as thyroid ultrasound correlation.

D Effectiveness of Agent Number

We conduct an ablation study to evaluate the sensitivity of MARCH to the number of participating Fellow Agents $\mathcal{A}_{\text{fel}}$. As detailed in Table 5, we vary the agent count from 1 to 20. Due to the budget constraints for LLM usage, the evaluation is conducted on a subset of 100 samples from the RadGenome-ChestCT test set.

The experimental results indicate that increasing the number of Fellow Agents initially enhances report quality. An ensemble of 5 agents achieved the highest linguistic quality, reaching a BLEU-1 of 0.476 and a METEOR score of 0.455. 10-agent ensembles slightly improved the F1 score to 0.337.

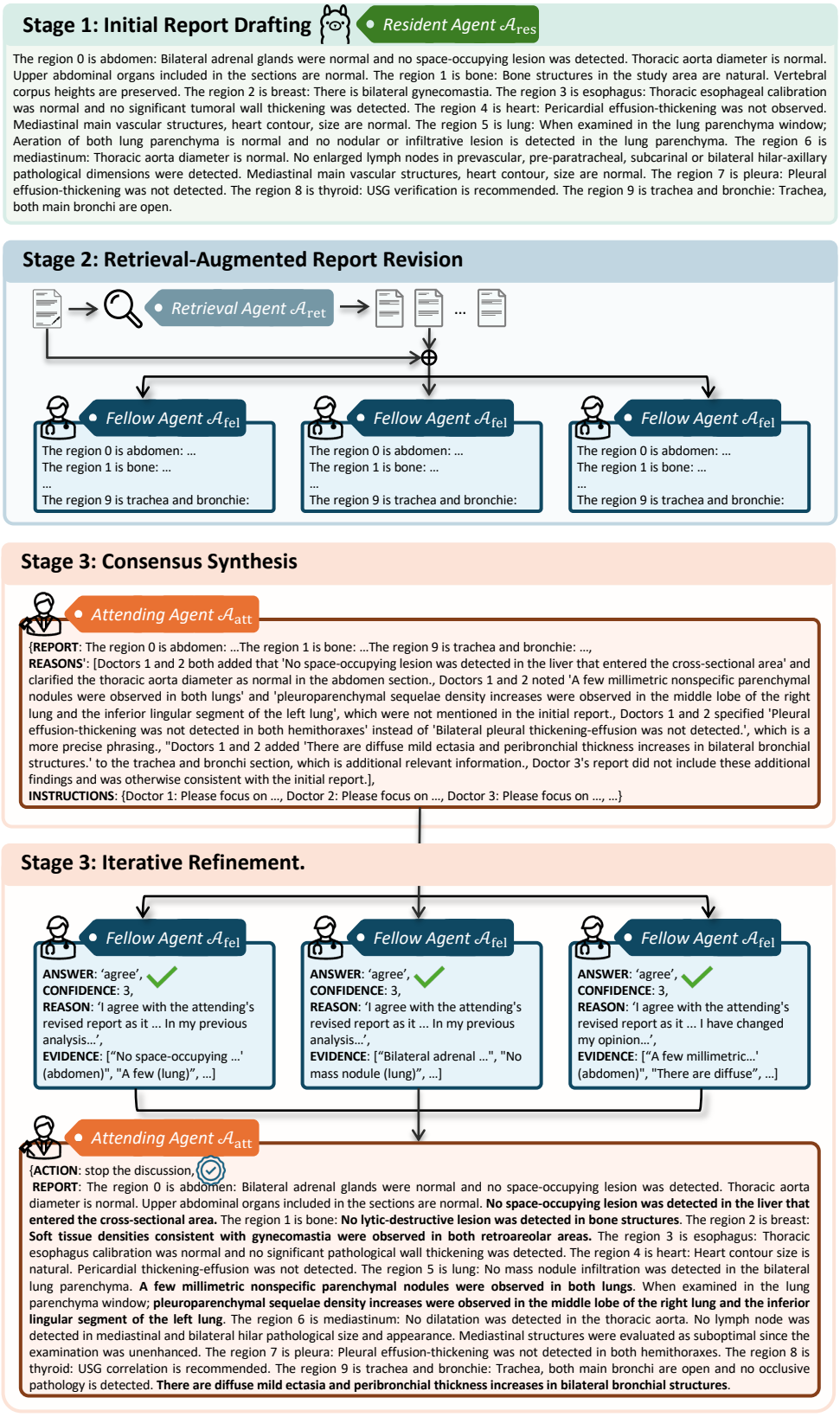

Figure 3: Case study.

However, increasing the ensemble size to 20 agents led to a marginal decline in performance, suggesting that excessive agent density may introduce redundant information or discursive noise. These findings indicate that a moderate number of agents effectively balance diverse diagnostic perspectives with collaborative stability. To optimize the trade-off between report fidelity and inference cost, we utilized 3 Fellow Agents as the default configuration for experiments on the whole dataset.

E Clinical Efficacy across Clinical Abnormalities

In Figure 4, we present a detailed analysis of MARCH's clinical efficacy across 18 clinical abnormalities in the RadGenome-ChestCT dataset, including *arterial wall calcification, coronary artery wall calcification, lung nodule, emphysema, cardiomegaly, lung opacity, consolidation, atelectasis, pleural effusion, mosaic attenuation pattern, pulmonary fibrotic sequela, medical material, pericardial effusion, interlobular septal thickening, hiatal hernia, lymphadenopathy, peribronchial thickening, and bronchiectasis*. We compare MARCH against the Resident Agent $\mathcal{A}_{\text{res}}$ baseline that generates reports without multi-agent collaboration. Results demonstrate that MARCH consistently outperforms the baseline across all abnormalities in terms of Precision, Recall, and F1-Score. Specifically, MARCH achieves high recall for abnormalities such as Hiatal hernia, Coronary artery wall calcification, and Arterial wall calcification, with scores exceeding 0.8. In terms of the overall F1-Score, MARCH shows significant gains in identifying complex findings such as Arterial wall calcification, Coronary artery wall calcification, and Cardiomegaly, while maintaining a robust balance between precision and sensitivity.

F Examples of the Generated Report

In Figure 5, we provide examples of radiology reports generated by MARCH compared to those produced by the Resident Agent $\mathcal{A}_{\text{res}}$ baseline. $\mathcal{A}_{\text{res}}$'s initial drafts often include extraneous or uncertain observations, such as suspected gynecomastia or small nonspecific nodules. In contrast, MARCH delivers refined reports that more closely align with the reference reports by effectively filtering these potential hallucinations and emphasizing clinically relevant findings through hierarchical collaboration. For instance, MARCH accurately identifies the thyroid's enlarged and nodular appearance while maintaining a concise summary of normal findings in the abdomen and bone structures. In Case 2, while the Resident Agent focuses on standard anatomical observations, MARCH demonstrates superior clinical fidelity by correctly identifying specific anatomical variants, such as an accessory spleen, and incorporating critical recommendations for USG verification. These examples underscore MARCH's capability to leverage multi-agent collaboration to resolve diagnostic ambiguities and generate more reliable, clinically validated radiology reports.

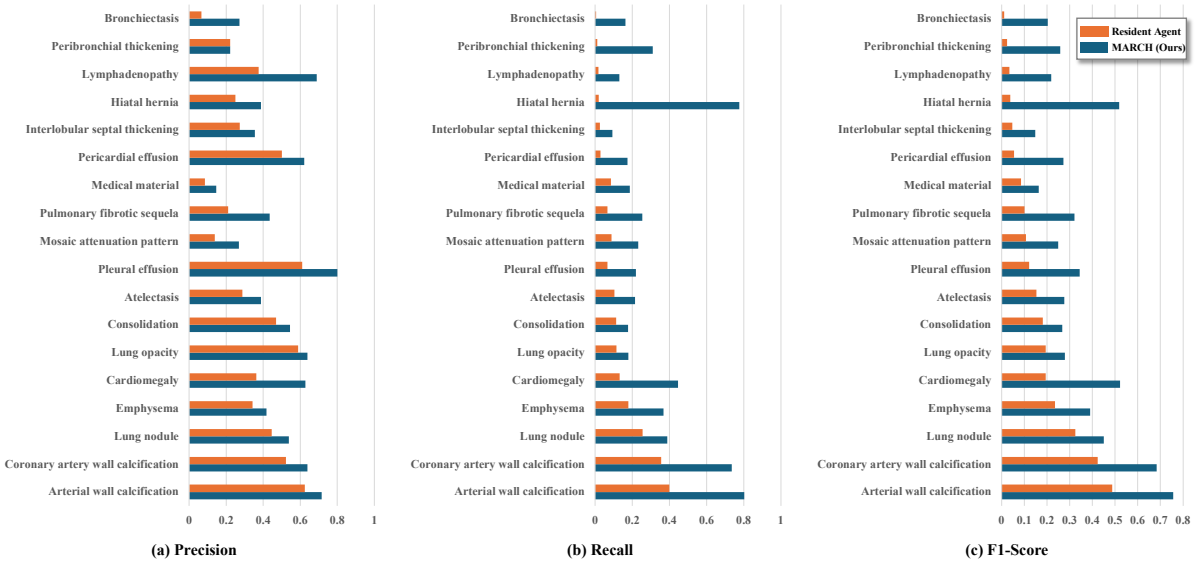

Figure 4: Clinical efficacy in terms of (a) Precision, (b) Recall, and (c) F1-score.

	Reference Report	Resident Agent	ARCH (Ours)
Case 1	Bilateral adrenal glands were normal and no space-occupying lesion was detected. Thoracic aorta diameter is normal. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bone structures in the study area are natural. Vertebral corpus heights are preserved. Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. Pericardial effusion-thickening was not observed. Mediastinal main vascular structures, heart contour, size are normal. When examined in the lung parenchyma window; Aeration of both lung parenchyma is normal and no nodular or infiltrative lesion is detected in the lung parenchyma. Pleural effusion-thickening was not detected. Trachea, both main bronchi are open.	Bilateral adrenal glands were normal and no space-occupying lesion was detected. Thoracic aorta diameter is normal. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bone structures in the study area are natural. Vertebral corpus heights are preserved. There is a 20x10 mm nodular density at the level of the right breast areola (gynecomastia?). Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. Pericardial effusion-thickening was not observed. Mediastinal main vascular structures, heart contour, size are normal. Aeration of both lung parenchyma is normal and no nodular or infiltrative lesion is detected in the lung parenchyma. Pleural effusion-thickening was not detected. When examined in the lung parenchyma window; A subpleural 4 mm nonspecific nodule is observed in the lateral aspect of the right lung middle lobe. The thyroid is larger than normal and nodular in appearance. Trachea, both main bronchi are open.	Bilateral adrenal glands were normal and no space-occupying lesion was detected. Thoracic aorta diameter is normal. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bone structures in the study area are natural. Vertebral corpus heights are preserved. Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. Pericardial effusion-thickening was not observed. Mediastinal main vascular structures, heart contour, size are normal. Aeration of both lung parenchyma is normal and no nodular or infiltrative lesion is detected in the lung parenchyma. Pleural effusion-thickening was not detected. When examined in the lung parenchyma window; The thyroid is larger than normal and nodular in appearance. Trachea, both main bronchi are open.
Case 2	There is an accessory spleen view in the spleen hilum. Thoracic aorta diameter is normal. Changes in the gallbladder bed related to possible cholestectomy are observed. There is a decrease in density consistent with hepatosteatosis in the sections passing through the upper abdomen. Bone structures in the study area are natural. Vertebral corpus heights are preserved. Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. Pericardial effusion-thickening was not observed. Mediastinal main vascular structures, heart contour, size are normal. Sequelae changes are observed in the inferior lingular segment. There is a ground-glass-like focal density increase at the apical level of the upper lobe. When examined in the lung parenchyma window; Sequela changes are observed in the middle lobe of the right lung. Thoracic aorta diameter is normal. There was no pleural effusion or obvious sign of pneumonia. Trachea, both main bronchi are open.	Bilateral adrenal glands were normal and no space-occupying lesion was detected. Thoracic aorta diameter is normal. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bone structures in the study area are natural. Vertebral corpus heights are preserved. There is bilateral gynecomastia. Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. Pericardial effusion-thickening was not observed. Mediastinal main vascular structures, heart contour, size are normal. When examined in the lung parenchyma window; Aeration of both lung parenchyma is normal and no nodular or infiltrative lesion is detected in the lung parenchyma. Thoracic aorta diameter is normal. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. Pleural effusion-thickening was not detected. USG verification is recommended. Trachea, both main bronchi are open.	There is an accessory spleen view in the spleen hilum. Thoracic aorta diameter is normal. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bone structures in the study area are natural. Vertebral corpus heights are preserved. Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. Pericardial effusion-thickening was not observed. Mediastinal main vascular structures, heart contour, size are normal. When examined in the lung parenchyma window; Aeration of both lung parenchyma is normal and no nodular or infiltrative lesion is detected in the lung parenchyma. Thoracic aorta diameter is normal. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. Pleural effusion-thickening was not detected. USG verification is recommended. Trachea, both main bronchi are open.

Figure 5: Examples of the generated reports.