MedPAO-Assist: Edge-Deployable, Modality-Aware Agentic AI for Radiology Report Structuring

Shrish Shrinath Vaidya[1][0009-0006-5438-9500], Minmini Selvam[2], and Ganapathy Krishnamurthi[1]

[1] Department of Data Science and Artificial Intelligence, Indian Institute of Technology Madras, Chennai, India
shrish.s.vaidya@dsai.iitm.ac.in
[2] TANUH – The AI Centre of Excellence in Healthcare, Indian Institute of Science, Bengaluru, India

Abstract. The adoption of Large Language Models for radiology report structuring has been largely dominated by either proprietary cloud-based models or large open models requiring high-end GPUs for deployment, which is a critical bottleneck for clinical adoption in radiology workflows that demand data privacy and offline operation. To address this, we present MedPAO-Assist, an edge-deployable agentic framework that generates structured radiology reports directly from radiologist dictation and voice commands, requiring only $\approx$ 3.3 GB peak RAM and achieving an average end-to-end latency of $\approx$ 31 seconds on a standard laptop CPU through an asynchronous, multi-threaded execution pipeline. The system currently supports four major imaging modalities – Chest X-ray, Chest CT, Abdomen CT, and Brain MRI. To our knowledge, it is among the earliest agentic systems to run its entire pipeline with a Small Language Model (SLM) in the 0.6B-parameter range. This multi-agent architecture, built around an Extract-Ground-Generate paradigm, performs concept extraction, SNOMED-CT grounding, and structured report generation, followed by a speech-driven human-in-the-loop correction phase, together forming a fully "no-touch" reporting workflow that requires no keyboard input, external API calls, or internet connectivity. We evaluate each pipeline stage individually – concept extraction, ontology grounding, and concept categorization and report hardware profiling results demonstrating the system's feasibility for real-time, CPU-only clinical deployment. The complete code and setup instructions are here: https://github.com/MiRL-IITM/MedPAO-Assist.

Keywords: Edge processing · Agentic Systems · Small Language Models.

1 Introduction

Radiology reports are formal documents that institutions rely on for billing, quality auditing, and regulatory compliance [6]. Professional societies now recommend structured reporting be integrated directly into radiologists' workflows

[7, 3]. Yet manually structuring a report, or correcting an automatically generated one, remains time-consuming on top of an already demanding clinical workload [1], motivating systems that perform accurate, protocol-aware report structuring directly from dictation without disrupting existing workflow.

Automated structured reporting first emerged through rule-based methods [17] and traditional ML models [27], both of which depend on heavy manual annotation and generalize poorly beyond the templates they were built for. Building on this, an early speech-driven NLP dialogue system demonstrated automated structuring for abdominal CT in suspected urolithiasis [16], but, like its rule-based predecessors, remained tied to a single fixed protocol rather than a modality-agnostic design. Commercial products such as MedicAI [21], RadAI [25], and Quillr [24] have since brought speech-driven structuring into clinical use at scale, yet all three are exclusively cloud-hosted, with no on-premise deployment option – a critical gap for institutions where patient dictations and imaging-derived data cannot leave the local clinical environment. More recently, LLM-based approaches [18, 13, 2] have shown that GPT-3.5/4 can structure reports well given sufficient clinical context, and RadExtract [8] similarly applies Google's LangExtract library with Gemini to the task; however, these approaches rely on proprietary, cloud-hosted models, raising data privacy and hospital-compliance concerns. Efforts more attuned to deployment constraints have instead trained dedicated pipelines for CXR-specific structuring [23] or lightweight, edge-oriented models [22]; the latter shows that sub-300M-parameter models can rival LLMs in performance, but still requires substantial supervised training data and remains confined to a single modality. Across this entire trajectory, methods are further framed as fully automated pipelines rather than human-AI co-adaptive tools, offering no structured, memory-informed pathway for a radiologist to correct and refine model output as part of the reporting workflow itself.

Beyond the workflow gaps above, realizing such a system safely is constrained by two further factors: patient dictations constitute protected health information, making cloud-hosted large language models (LLMs) a poor fit for real-time clinical deployment, while on-device Small Language Models (SLMs) that avoid this risk are known to suffer degraded reasoning on long-context inputs [20, 12] – directly at odds with the length of typical radiology dictations.

To address these gaps, we propose **MedPAO-Assist**, an agentic AI framework with the following contributions:

- A **Multi-Agent framework** (MAS) for protocol-driven, multi-modality report structuring.
- Efficient use of **SLMs and PEFT methods** for medical concept extraction and categorization.
- **Low-resource CPU deployment** with minimal latency and memory footprint.
- Integration of MAS with **ASR** to provide a multi-modality, protocol-compliant, **touch-free** reporting interface.

2 Methodology

2.1 System Architecture

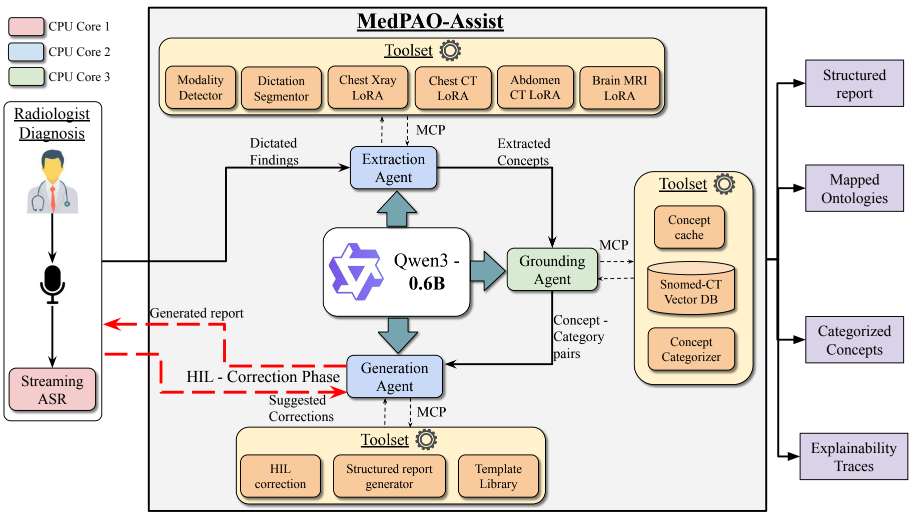


Fig. 1: Proposed System Architecture

MedPAO-Assist employs a multi-threaded architecture that achieves low end-to-end latency by enabling real-time speech streaming through three parallel threads (Fig. 1). All agents share a Qwen3-0.6B Small Language Model (SLM) [30] with task-specific LoRA adapters for specialized reasoning. Streaming transcription is performed using INT8-quantized MedASR [29], while the SLM and LoRA adapters are deployed in GGUF format with FP16 inference on the target CPU. The system assumes a clinically familiar dictation workflow in which the radiologist states the imaging modality, followed by the findings and impression, allowing the Modality Detector to identify the appropriate protocol template early in the stream. Inspired by MedPAO [28], the framework follows an **EGG (Extract–Ground–Generate)** paradigm, where the ASR, Extraction, and Grounding modules operate asynchronously on incoming speech segments to minimize latency. The Generation Agent then aggregates the extracted and grounded information to produce the final structured report, followed by a human-in-the-loop correction stage.

2.2 Extraction Agent

Small Language Models (SLMs) exhibit degraded performance on long-context inputs [20, 12]. To address this, the Extraction Agent identifies and retains only clinically relevant findings(i.e., concepts), transforming downstream processing from sentence-level to concept-level reasoning while preserving essential diagnostic information. The same SLM is used by the Modality Detector to identify the imaging modality and retrieve the corresponding protocol template. A rule-based Dictation Segmentor further separates the report into *Findings* and *Impression* sections using regular expressions, enabling independent downstream processing.

Concept Extraction Tools: Separate LoRA adapters are trained for each imaging modality using a three-stage pipeline of Supervised Fine-Tuning (SFT), Knowledge Distillation(KD), and Instruction Tuning(IT) [10, 9, 30]. First, the base model is LoRA fine-tuned on 500K MIMIC-IV-Note [15] radiology reports to adapt it to radiology language. Next, concept extraction knowledge is distilled from Qwen3-32B, which performs zero-shot concept extraction on another 500K MIMIC-IV-Note reports, with its predictions and output logits used to continue training the same adapter. Finally, the adapter is instruction-tuned on the RadGraph-XL [4] training set, containing report–concept pairs from Chest X-ray, Chest CT, Abdomen CT, and Brain MRI.

2.3 Grounding agent

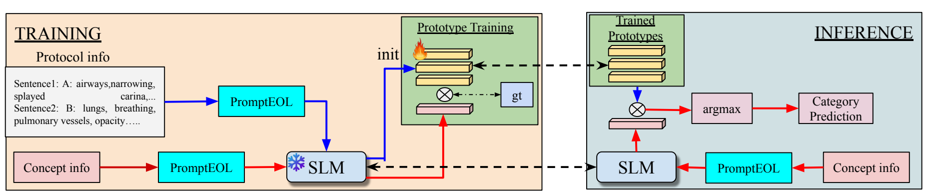


Fig. 2: Proposed Strategy for concept categorization using PromptEOL[14]

Ontology Mapping: The Ontology Mapping module is responsible for grounding the extracted medical concepts in structured clinical knowledge. SNOMED-CT is adopted as the reference ontology due to its comprehensive coverage of radiological terminology. Extracted concepts are mapped to their semantically corresponding SNOMED-CT terms using SapBERT [19] embeddings and embedding-based similarity search. To enable efficient CPU inference and reduce memory consumption, the similarity search employs IVF indexing and SQ8 scalar quantization of embeddings, as implemented in FAISS [5]. For each matched SNOMED-CT concept, its corresponding description and ancestor concepts are also retrieved to facilitate the subsequent categorization stage.

Concept Categorization: Unlike previous approaches [28], our framework cannot rely on the reasoning capabilities of larger language models to directly assign extracted concepts to protocol-specific categories. Instead, we leverage the inherent sentence embedding capabilities of Small Language Models (SLMs), following the PromptEOL framework [14]. Building upon this approach, we introduce a learnable embedding layer to optimize protocol prototypes(Fig. 2). Experimental evaluation demonstrates that the proposed method achieves strong performance with as few as 20-50 annotated concept–category training pairs, indicating that the SLM-derived prototype embeddings provide an effective initialization for low-resource learning. Protocol templates for the different imaging modalities were selected in consultation with a radiologist and based on the RSNA Reporting Template Library [26]. Synthetic concept–category pairs were subsequently generated using Claude Sonnet 5 according to the protocol definitions and used for training and evaluation of the proposed concept categorization framework.

2.4 Generation agent

Structured Generation: The Generation Agent is responsible for producing the final structured report by aggregating the extracted concepts, their corresponding source sentences, and the concept categorization results. As the source sentences remain in spoken dictation form (e.g., "I could see...", "I am looking at...") until this stage, they must be transformed into clinically appropriate reporting language. To achieve this, the structured report generation module employs a LoRA adapter over the base SLM to convert the dictated phrases into their corresponding medical report equivalents. This LoRA adapter is fine-tuned on a synthetic dataset generated using Claude Sonnet 5, GPT-OSS-120B, Llama-3.3-70B-Versatile, and DeepSeek V3.

Human-in-Loop Correction: Since the proposed framework operates using a 0.6B-parameter Small Language Model (SLM), occasional prediction errors are inevitable. To address this, MedPAO-Assist incorporates a human-in-the-loop correction stage following the initial structured report generation. During this phase, the ASR model and the base SLM operate synchronously to process verbal corrections provided by the clinician or radiologist. The spoken feedback is transcribed and used to update the generated report, enabling efficient post-generation refinement while preserving a speech-driven workflow.

3 Results and Analysis

3.1 Concept Extraction

The proposed concept extraction pipeline described in Section 2.2 is evaluated on the RadGraph-XL test set, comprising approximately 290 radiology reports spanning Chest X-ray, Chest CT, Abdomen CT, and Brain MRI modalities.

Table 1: Evaluation of Concept Extraction task with various phases of finetuning.

Method	**F1-score ↑**	**Precision ↑**	**Recall ↑**	**Hamming Loss ↓**
Base Model	0.08	0.19	0.06	0.0048
+ SFT	0.08	0.15	0.07	0.0047
+ KD	0.55	0.66	0.50	0.0026
+ IT	0.78	0.82	0.76	0.0012

Since concept extraction is performed using a generative Small Language Model (SLM), the predictions are evaluated using fuzzy string matching with a similarity threshold of 80. The resulting classification metrics across the different stages of fine-tuning are presented in Table 1. The results demonstrate a consistent improvement in performance across successive training stages, with the final model achieving an F1-score of 0.78 achieving a boost of 0.7 as compared to the base model. This substantial improvement highlights the effectiveness of the proposed three-stage fine-tuning strategy.

3.2 Ontology mapping

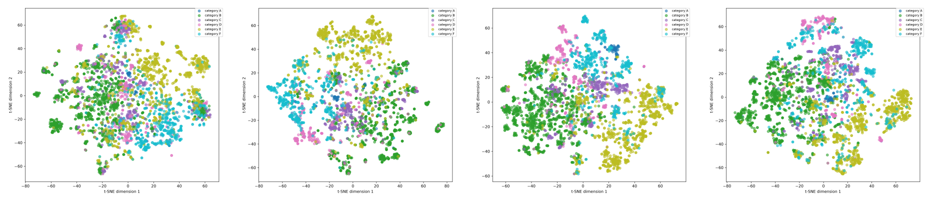
Table 2: Evaluation of SNOMED CT grounding under various FAISS index configurations.

Method	**Top1-Acc. ↑**	**Top3-Acc. ↑**	**Latency(in sec) ↓**	**Peak RAM(in MB) ↓**
Baseline[1]	56.74	66.00	0.173	3500
+ IVF	56.28	66.02	0.072	510
+ SQ8	56.14	66.01	0.039	400
On CPU[2]	55.53	65.54	0.3	538

The ontology mapping module of the Grounding Agent (Section 2.3) is evaluated on the SNOMED CT Entity Linking Challenge [11] test set. As shown in Table 2, brute-force search over the complete SNOMED CT embedding space incurs the highest latency and memory usage. IVF indexing (`nlist=256`, `nprobe=64`) substantially improves efficiency with negligible impact on retrieval accuracy, while SQ8 vector quantization further reduces the memory footprint. The final deployment combines an INT8-quantized SapBERT model with IVF and SQ8 on CPU, reducing peak memory from 3500 MB to 538 MB while maintaining comparable Top-1/Top-3 accuracy, demonstrating that aggressive model and index compression can be achieved with minimal loss in ontology mapping performance.

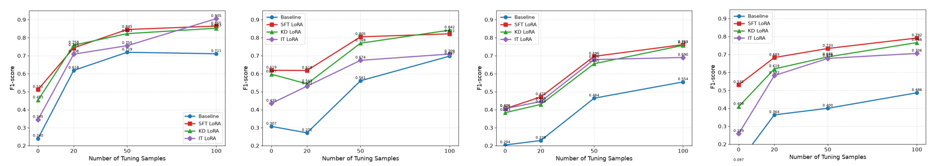
[1] Brute-force search by loading the complete SNOMED CT embedding database into memory.
[2] Q8-quantized SapBERT + IVF + SQ8 indexing on CPU.

3.3 Concept Categorization



(a) Base model (b) +SFT (c) +KD (d) +IT

Fig. 3: t-SNE plot of embedding of various models on CXR protcol categories.



(a) CXR (b) ChestCT (c) AbdomenCT (d) BrainMRI

Fig. 4: Performance of all the models in the proposed categorization framework across various modality protocol categories vs number of finetuning samples.

The proposed concept categorization framework (Section 2.3) is evaluated qualitatively using t-SNE visualizations of sentence embeddings generated from enriched CXR concept–ontology text, and quantitatively on a synthetic dataset spanning the four imaging modalities. Figure 3 compares embeddings from the base model and LoRA adapters trained at different stages (Section 2.2) on CXR concepts alone, showing that the combination of the base model and SFT LoRA produces the most discriminative clusters. For quantitative evaluation, 200 concept–category pairs per modality were synthesized using Claude Sonnet 5 from protocol definitions, with 100 stratified samples used for training and 100 for testing. As shown in Figure 4, protocol prototypes learned with varying training sizes ([0, 20, 50, 100]) consistently achieve the highest F1-score when using the base model with the SFT LoRA, while the remaining LoRA adapters improve progressively with additional training data, indicating that the SFT LoRA learns more effective sentence embeddings for low-resource concept categorization.

Table 3: Hardware profiling of MedPAO-Assist on the targeted CPU, per stage and end-to-end under sequential vs. asynchronous execution. Throughput units differ by stage; all other columns share consistent units (TTFT/Total Latency in seconds, Peak RAM in GB).

Stage	**TTFT ↓**	**Throughput ↑**	**Total Latency ↓**	**Peak RAM ↓**
MedASR	0.708	79.425	18.171	2.68
Concept Extraction	3.551	4.699	11.652	3.04
Ontology Mapping	0.754	3.155	4.599	2.95
Concept Categorization	5.355	0.549	26.352	3.22
Structured Report Generator	7.947	4.928	18.545	3.23
Total System (Sequential)	21.722	–	67.498	3.34
Total System (Asynchronous)	**4.259**	–	**31.002**	3.20

3.4 Performance on Edge

All experiments were performed on an Intel i7-1165G7 quad-core CPU with 8 GB RAM, the target deployment platform, with results in Table 3 averaged over five reports from each imaging modality. The complete pipeline requires only 3.0–3.5 GB of peak memory and achieves an average end-to-end latency of 31 s for reports averaging 260 words under asynchronous execution. Table 3 further compares sequential execution, intended for offline conversion of existing reports, with asynchronous (pipelined) execution that overlaps processing stages to support live microphone-based reporting. Throughput is reported in stage-specific units (chars/s for MedASR, tokens/s for Concept Extraction and Structured Report Generation, and concepts/s for Ontology Mapping and Concept Categorization). Peak RAM remains similar across LLM-driven stages because all models are preloaded, so per-stage values represent resident memory rather than incremental memory usage.

4 Conclusion

In this work, we introduced MedPAO-Assist, an edge-deployable agentic AI system that generates protocol-compliant structured radiology reports from dictation. The framework uses cooperating small language models for concept extraction, ontology grounding, and report generation, with a speech-driven human-in-the-loop correction stage to keep the radiologist in control. By coupling MedASR streaming transcription with this agentic pipeline, the system is designed to reduce manual reporting effort while preserving modality-specific protocol adherence, without relying on external APIs, subscriptions, or internet connectivity - an essential property for deployment in privacy-sensitive clinical settings. This study is an initial step: evaluation is currently limited to hardware profiling on a single target CPU, and future work will include end-to-end workflow validation with practicing radiologists, along with further compression and CPU-level optimization to reduce latency and memory usage. By enabling privacy-preserving, offline, voice-driven structured reporting on commodity hardware, MedPAO-Assist provides a practical pathway toward integrating agentic AI into routine radiology workflows, particularly in settings where cloud-based solutions or dedicated GPU infrastructure are not feasible.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bailey, C.R., Bailey, A.M., McKenney, A.S., Weiss, C.R.: Understanding and appreciating burnout in radiologists. RadioGraphics **42**(5), E137–E139 (2022), https://doi.org/10.1148/rg.220037
2. Busch, F., Hoffmann, L., Dos Santos, D.P., Makowski, M.R., Saba, L., Prucker, P., Hadamitzky, M., Navab, N., Kather, J.N., Truhn, D., et al.: Large language models for structured reporting in radiology: past, present, and future. European Radiology **35**(5), 2589–2602 (2025)
3. Chung, C.Y., Makeeva, V., Yan, J., Prater, A.B., Duszak, R., Safdar, N.M., Heilbrun, M.E.: Improving billing accuracy through enterprise-wide standardized structured reporting with cross-divisional shared templates. Journal of the American College of Radiology **17**(1, Part B), 157–164 (2020). https://doi.org/10.1016/j.jacr.2019.08.034
4. Delbrouck, J.B.: RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports. PhysioNet (Sep 2025). https://doi.org/10.13026/j8e7-pr22, version 1.0.0
5. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2024)
6. Durack, J.C.: The value proposition of structured reporting in interventional radiology. American Journal of Roentgenology **203**(4), 734–738 (2014), https://doi.org/10.2214/AJR.14.13112
7. European Society of Radiology (ESR): Esr paper on structured reporting in radiology. Insights into Imaging **9**(1), 1–7 (2018). https://doi.org/10.1007/s13244-017-0588-8
8. Goel, A.: LangExtract (2026). https://doi.org/10.5281/zenodo.21126643, https://github.com/google/langextract
9. Gu, Y., Dong, L., Wei, F., Huang, M.: Minillm: Knowledge distillation of large language models. In: Proceedings of ICLR (2024)
10. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (2025)
11. Hardman, W., Banks, M., Davidson, R., Truran, D., Ayuningtyas, N.W., Ngo, H., Johnson, A., Pollard, T.: SNOMED CT Entity Linking Challenge. PhysioNet (Feb 2026), https://doi.org/10.13026/az93-vm52, version 1.2.1
12. Hsieh, C.P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Zhang, Y., Ginsburg, B.: Ruler: What's the real context size of your long-context language models? arXiv preprint arXiv:2404.06654 (2024)
13. Jiang, H., Xia, S., Yang, Y., Xu, J., Hua, Q., Mei, Z., Hou, Y., Wei, M., Lai, L., Li, N., et al.: Transforming free-text radiology reports into structured reports using chatgpt: A study on thyroid ultrasonography. European journal of radiology **175**, 111458 (2024)

14. Jiang, T., Huang, S., Luan, Z., Wang, D., Zhuang, F.: Scaling sentence embeddings with large language models. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 3182–3196. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.181, https://aclanthology.org/2024.findings-emnlp.181/
15. Johnson, A., Pollard, T., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV-Note: Deidentified free-text clinical notes. PhysioNet (Jan 2023). https://doi.org/10.13026/1n74-ne17, version 2.2
16. Jorg, T., Kämpgen, B., Feiler, D., Müller, L., Düber, C., Mildenberger, P., Jungmann, F.: Efficient structured reporting in radiology using an intelligent dialogue system based on speech recognition and natural language processing. Insights into imaging **14**(1), 47 (2023)
17. Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S.F., Forshee, R., Walderhaug, M., Botsis, T.: Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. Journal of biomedical informatics **73**, 14–29 (2017)
18. Li, H., Wang, H., Sun, X., He, H., Feng, J.: Prompt-guided generation of structured chest x-ray report using a pre-trained llm. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
19. Liu, F., Shareghi, E., Meng, Z., Basaldella, M., Collier, N.: Self-alignment pretraining for biomedical entity representations. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4228–4238. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.334, https://aclanthology.org/2021.naacl-main.334/
20. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics **12**, 157–173 (2024). https://doi.org/10.1162/tacl_a_00638, https://aclanthology.org/2024.tacl-1.9/
21. Medicai: Structured reporting for radiology, with templates and AI. https://www.medicai.io/products/structured-reporting-for-radiology (2026)
22. Moll, J., Fay, L., Azhar, A., Ostmeier, S., Gatidis, S., Lueth, T.C., Langlotz, C., Delbrouck, J.B.: Structuring radiology reports: Challenging llms with lightweight models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 7718–7735 (2025)
23. Pellegrini, C., Delchev, A., Özsoy, E., Navab, N., Keicher, M.: Prototype-based knowledge guidance for fine-grained structured radiology reporting. arXiv preprint arXiv:2603.11938 (2026)
24. Quillr AI: Quillr AI – AI assistant for radiologists. https://quillr.ai/ (2026)
25. Rad AI: Radiology reporting software. https://www.radai.com/reporting (2026)
26. Radiological Society of North America: RadReport radiology reporting template library. https://radreport.org/
27. Steinkamp, J.M., Chambers, C., Lalevic, D., Zafar, H.M., Cook, T.S.: Toward complete structured information extraction from radiology reports using machine learning. Journal of digital imaging **32**(4), 554–564 (2019)
28. Vaidya, S.S., Palani, G., Ramesh, S., Balasubramanian, V., Selvam, M., Srinivasaraja, G., Krishnamurthi, G.: Medpao: A protocol-driven agent for structuring

medical reports. In: Qiu, J., Wu, J., Langlotz, C., Huang, B., Lei, Z., Wu, H., Liu, H., Xie, W. (eds.) AI for Clinical Applications. pp. 33–45. Springer Nature Switzerland, Cham (2026)
29. Wu, K., Variani, E., Bagby, T., Reddy, S., Pilgrim, R.: Medasr: An open-source model for high-accuracy medical dictation (2026), https://arxiv.org/abs/2605.16555
30. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)