MedTutor: A Retrieval-Augmented LLM System for Case-Based Medical Education

Dongsuk Jang[1,3] **Ziyao Shangguan**[1] **Kyle Tegtmeyer**[2]
Anurag Gupta[2] **Jan Czerminski**[2] **Sophie Chheang**[2] **Arman Cohan**[1]
[1]Department of Computer Science, Yale University
[2]Department of Radiology and Biomedical Imaging, Yale School of Medicine
[3]Interdisciplinary Program for Bioengineering, Seoul National University
{james.jang,ziyao.shangguan,arman.cohan}@yale.edu

Code Demo Video Dataset

Abstract

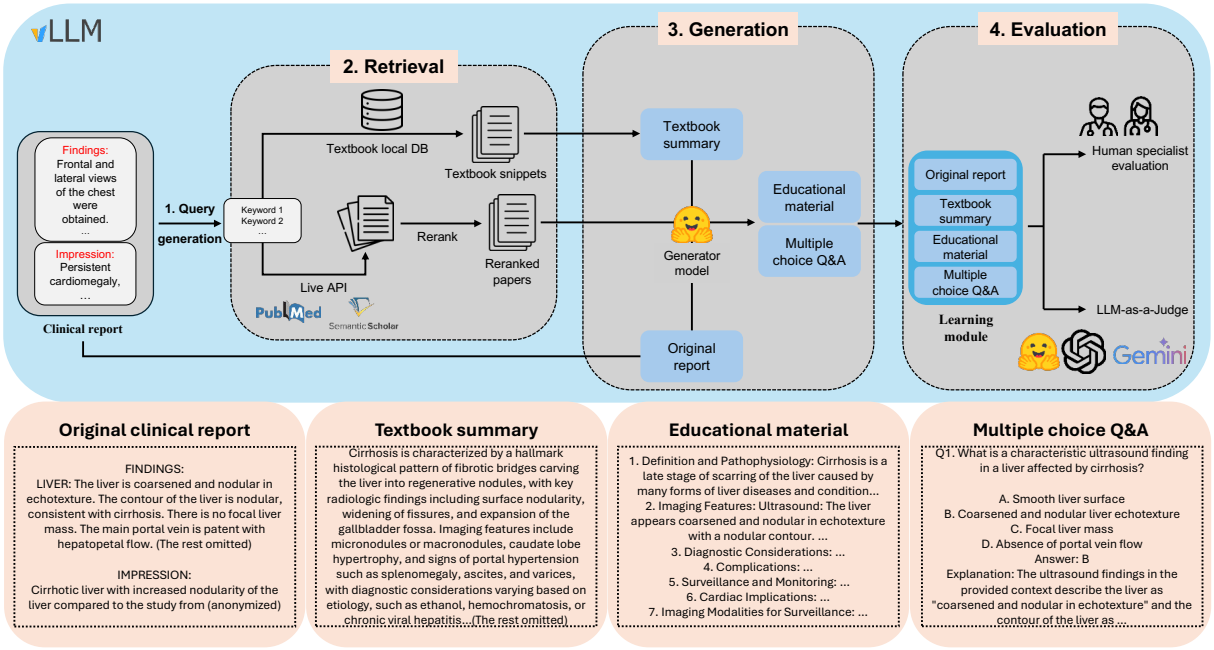
The learning process for medical residents presents significant challenges, demanding both the ability to interpret complex case reports and the rapid acquisition of accurate medical knowledge from reliable sources. Residents typically study case reports and engage in discussions with peers and mentors, but finding relevant educational materials and evidence to support their learning from these cases is often time-consuming and challenging. To address this, we introduce **MedTutor**, a novel system designed to augment resident training by automatically generating evidence-based educational content and multiple-choice questions from clinical case reports. MedTutor leverages a Retrieval-Augmented Generation (RAG) pipeline that takes clinical case reports as input and produces targeted educational materials. The system's architecture features a hybrid retrieval mechanism that synergistically queries a local knowledge base of medical textbooks and academic literature (using PubMed, Semantic Scholar APIs) for the latest related research, ensuring the generated content is both foundationally sound and current. The retrieved evidence is filtered and ordered using a state-of-the-art reranking model and then an LLM generates the final long-form output describing the main educational content regarding the case-report. We conduct a rigorous evaluation of the system. First, three radiologists assessed the quality of outputs, finding them to be of high clinical and educational value. Second, we perform a large-scale evaluation using an LLM-as-a Judge to understand if LLMs can be used to evaluate the output of the system. Our analysis using correlation between LLMs outputs and human expert judgments reveals a moderate alignment and highlights the continued necessity of expert oversight.

1 Introduction

The training of medical residents is an intensive learning process, built upon the foundation of studying and interpreting thousands of case reports. Residents routinely engage with clinical cases through discussions with peers and mentors, analyzing findings and differential diagnoses to deepen their understanding. However, while direct feedback from attending physicians is invaluable, the process of finding relevant educational material and supporting evidence for specific cases is often time-consuming and inconsistent (Rogers et al., 2019; Daniel et al., 2020). The sheer volume of medical literature and the challenge of identifying pertinent resources for each case can limit the depth of learning that residents achieve from their clinical experiences (Anderson and Anderson, 2019; Bednarczyk et al., 2014). There exists a significant opportunity to augment this traditional learning process with AI tools that can efficiently retrieve and synthesize educational content from clinical cases, drawing upon vast archives of medical knowledge. LLMs present a promising avenue for this augmentation, but their application in high-stakes medical domains is fraught with challenges, most notably the risk of factual inaccuracy (or hallucination) and the use of outdated knowledge (Abd-alrazaq et al., 2023; Li et al., 2023; Xie et al., 2023).

To overcome these challenges, we develop MedTutor, a system that grounds LLM generation in verifiable, contextually relevant medical knowledge to case reports through a RAG pipeline. Our primary goal is to provide medical residents with a reliable tool that transforms any given clinical report into a concise, and highly relevant educational module. We focus on radiology as the domain of study, although, our techniques are generalizable to other domains. The system begins by decomposing clinical reports into actionable diagnostic queries and keywords that can be effectively issued to a search index, enabling targeted retrieval of relevant educational material. It then initiates a hybrid retrieval process that simultaneously queries a curated database of medical textbooks, and per-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 319–353
November 4-9, 2025



Original clinical report

FINDINGS:
LIVER: The liver is coarsened and nodular in echotexture. The contour of the liver is nodular, consistent with cirrhosis. There is no focal liver mass. The main portal vein is patent with hepatopetal flow. (The rest omitted)

IMPRESSION:
Cirrhotic liver with increased nodularity of the liver compared to the study from (anonymized)

Textbook summary

Cirrhosis is characterized by a hallmark histological pattern of fibrotic bridges carving the liver into regenerative nodules, with key radiologic findings including surface nodularity, widening of fissures, and expansion of the gallbladder fossa. Imaging features include micronodules or macronodules, caudate lobe hypertrophy, and signs of portal hypertension such as splenomegaly, ascites, and varices, with diagnostic considerations varying based on etiology, such as ethanol, hemochromatosis, or chronic viral hepatitis...(The rest omitted)

Educational material

1. Definition and Pathophysiology: Cirrhosis is a late stage of scarring of the liver caused by many forms of liver diseases and condition...
2. Imaging Features: Ultrasound: The liver appears coarsened and nodular in echotexture with a nodular contour. ...
3. Diagnostic Considerations: ...
4. Complications: ...
5. Surveillance and Monitoring: ...
6. Cardiac Implications: ...
7. Imaging Modalities for Surveillance: ...

Multiple choice Q&A

Q1. What is a characteristic ultrasound finding in a liver affected by cirrhosis?

A. Smooth liver surface
B. Coarsened and nodular liver echotexture
C. Focal liver mass
D. Absence of portal vein flow
Answer: B
Explanation: The ultrasound findings in the provided context describe the liver as "coarsened and nodular in echotexture" and the contour of the liver as ...

Figure 1: The overall architecture of the MedTutor system.

forms live searches on academic search engines (i.e., PubMed and Semantic Scholar) for current published literature related to the case.

The retrieved evidence undergoes a multi-faceted processing step: academic articles and textbook snippets are reranked for relevance to the case, using a state-of-the-art reranking model, Qwen3-Reranker-8B (Zhang et al., 2025b). Finally, all processed evidence—the original report, keywords, top-ranked articles, and textbook summaries—is synthesized by a generator LLM into two distinct outputs: a comprehensive set of educational material and a set of multiple-choice questions (MCQs) designed to test understanding. The overview of the system is illustrated in Figure 1.

This work makes three primary contributions:

• We detail the design and implementation of the MedTutor system, a scalable and efficient architecture that leverages asynchronous I/O, parallel multi-GPU inference with vLLM (Kwon et al., 2023), and optimized batch processing to handle large workloads.

• We introduce a new, expert annotated benchmark dataset for evaluating the quality of AI-generated educational content. We run our pipeline with 6 LLMs(see Appendix A for details) across 2,000 clinical reports per each 5 major radiology datasets (*i.e.,* Yale Hospital Internal, MIMIC-CXR (Johnson et al., 2019, 2024), MIMIC-IV-note (Johnson et al., 2023), CheXpert Plus (Chambon et al., 2024), and ReXGradient-160K (Zhang et al., 2025a)).

• We collected comprehensive evaluations from three radiologists, alongside a LLM-as-a-Judge evaluation with four models for all experiments. This dataset, which we are planning to publicly release, will be a valuable resource for evaluating the quality and clinical utility of generative models in medicine.

Our analysis provides insights about the usefulness of our system to users and highlight the strengths and weaknesses of LLMs in evaluating educational content in our setting.

2 Related Work

Our work is situated at the intersection of RAG, the application of LLMs in medicine, and the critical need for trustworthy medical AI systems. We structure our review accordingly.

2.1 LLMs and RAG in the Medical Domain

The application of LLMs in medicine has shown immense promise. General-purpose foundation models have demonstrated impressive capabilities on standardized medical exams and complex diagnostic problems (Nori et al., 2023; Singhal et al., 2023). This has fueled a broader vision for generalist biomedical AI that can assist with a wide range of clinical tasks (Tu et al., 2023). However, the "black-box" nature of these models and their potential for factual errors or "hallucinations" remain

significant barriers to clinical adoption, necessitating robust evaluation frameworks (Huang et al., 2024; Li et al., 2023; Xie et al., 2023).

To mitigate these risks, RAG has emerged as a key paradigm for building dependable clinical tools (Lewis et al., 2021). By grounding LLM outputs in external, verifiable evidence from reliable medical literature, RAG provides a pathway to trustworthy AI. A recent perspective in *Nature Medicine* strongly advocates for RAG as a prerequisite for the responsible deployment of generative AI in healthcare (Yang et al., 2024). The field is now maturing to a point where standardized benchmarks for medical RAG are being established, allowing for more rigorous evaluation of these systems (Xiong et al., 2024a). Our work contributes to this growing body of literature by presenting a novel RAG-based system specifically designed for medical education, a domain where accuracy and reliability are paramount.

2.2 LLMs for Medical Education

LLMs show promise in generating medical exam content, though concerns about accuracy necessitate expert oversight (Zhu et al., 2024). Integrating RAG improves reliability by grounding output in external sources, with studies reporting notable gains in question-answering accuracy using medical textbooks (Chen et al., 2025; Wang et al., 2023). Benchmarks like MIRAGE further validate RAG's role in medical QA tasks (Xiong et al., 2024b).

For resident training, LLMs can assess skills and provide feedback, but expert review remains vital (Atsukawa et al., 2024). Systems enabling citation generation enhance factuality (Wang et al., 2025), while evaluation frameworks like LLM-as-a-Judge offer scalability despite only moderate alignment with human judgment (Zheng et al., 2025). New approaches continue to embed evidence-based medicine principles into RAG pipelines for clinically accurate educational content (Lu et al., 2025).

Our system, MedTutor, is distinct in its focus on transforming a single clinical report into a comprehensive educational module, featuring synthesized educational material and MCQs grounded in a hybrid retrieval from both medical textbooks and the latest academic literature. This approach is designed not to replace expert judgment but to augment it, fostering the self-directed learning skills that are crucial for lifelong professional development (Bravata et al., 2003; Williams and Ntiri, 2018).

3 MedTutor

MedTutor is a RAG system designed to support medical residents on case-based education. It involves a pipeline approach in retrieving highly relevant educational content from both textbooks and literature and produces a coherent educational material as well as multiple-choice questions related to a case. While MedTutor's design is general and can be applicable to many clinical practices, we focus our domain on radiology due to availability of public datasets and our access to domain experts.

3.1 The MedTutor Pipeline Stages

The input to MedTutor is a case report, which will be processed through a sequence of automated stages, each designed for parallel execution.

Case decomposition into search queries: The process begins with a source radiology report. Then we use an LLM (`Llama-3.3-70B-Instruct`) to process the radiology report and decompose it into multiple keyword based queries that will be used for retrieval. These queries are key diagnostic terms and findings. Prompts for case decomposition into search queries are shown in Appendix D.1.

Hybrid Evidence Retrieval: For each search query, the system performs a hybrid retrieval process in parallel described below: *(1) Local retrieval for textbook material:* Textbooks and notes are essential resources for medical education. In our MedTutor system, we first apply OCR to a radiology textbook (Dahnert and Ovid Technologies, 2017) using the `mistral-ocr-2503` model, then segment and index the material by page. We generate dense embeddings for these materials with the `Qwen3-Embedding-8B` model, which has demonstrated state-of-the-art performance in embedding and retrieval tasks on the MTEB benchmark (Muennighoff et al., 2022) among models of comparable size. These embeddings are stored in a precomputed vector database for subsequent queries. For local database search, we employ a bi-encoder architecture to generate dense vector representations for both the query and the pre-indexed textbook pages, subsequently identifying the most relevant page using cosine similarity. *(2) Retrieval using academic APIs:* Some case reports are more specialized or rare, requiring retrieving knowledge from latest academic literature. Therefore, we also employ retrieval from academic search engines. We use PubMed and Semantic Scholar APIs, two commonly used and freely available scholarly sys-

tems, to fetch the latest relevant research papers. To prevent rate-limiting, API calls are managed by an `asyncio.Semaphore`. If pre-fetched results for the queries are available, this step is skipped to improve efficiency.

Evidence Processing: The retrieved evidence is then processed through two concurrent tasks: *(1) Reranking:* As the search engine results using keyword queries can be noisy, we employ a reranking stage to prioritize the most relevant (top 2) documents to the case report. This is handled by a dedicated service running the `Qwen3-Reranker-8B` model, a strong reranker according to the MTEB benchmark. The reranker is given a contextualized query containing both the original report's text and the specific search keyword to improve relevance. *(2) Query-focused Summarization:* Concurrently, the content retrieved from the local textbook database is summarized with respect to the query by a generator LLM to distill key information related to the keywords into a concise way.

Generating Learning Modules: Finally, the original case report, the top retrieved content including the textbook snippets and abstracts of related papers, and the search keywords are passed to a generator LLM to generate a concise learning module. These learning modules contain comprehensive explanatory material contextualizing the case within broader medical knowledge, followed by multiple-choice questions designed to test understanding of key concepts. Prompts used for generating learning modules are in Appendix D.

Optimized Multi-Task Generation: The generation step is heavily optimized for efficiency. Instead of generating outputs sequentially, the system first constructs prompts for all cases received, and all sub-tasks.

Batch Construction: Two distinct batches of input prompts to LLMs are created: one for generating the final educational modules and another for generating multiple-choice questions. These input prompts are long-context (3530 tokens for MCQ, 3463 tokens for Educational module in average), containing the original report, the list of keywords, the abstracts of the top-ranked papers after reranking, and the generated textbook summaries. *Concurrent Batch Inference:* The two batches are sent concurrently to the generation service. The `generate_text_batch` method in our `VLLMHandler` passes the entire list of prompts to the vLLM engine in a single call. This fully leverages vLLM's continuous batching capability, allowing the GPU to process multiple requests simultaneously without padding, dramatically increasing throughput and reducing overall processing time. This architecture, particularly the use of batch generation with vLLM, allows MedTutor to process hundreds of complex reports far more efficiently than a naive, sequential approach, making it a practical tool for large-scale educational content creation.

Local Deployment: We deploy MedTutor completely locally using locally served open-source LLMs, without reliance on any cloud-based LLM APIs. This allows responsible and private handling of medical data.

3.2 System Design Details

The MedTutor pipeline is an asynchronous, multi-stage system designed for efficiency, scalability, and modularity. The architecture leverages parallel processing across multiple GPUs and optimized batching to handle large-scale report generation. The entire workflow is orchestrated by a central `asyncio` event loop, which communicates with dedicated `ModelWorker` processes via multiprocessing queues. A conceptual overview of the architecture is shown in Figure 1.

3.3 Architecture for Scalability

At the core of our system is a hybrid concurrency model designed to maximize throughput and resource utilization.

Asynchronous Orchestration: The main process runs on an `asyncio` event loop, managing I/O-bound tasks such as live API calls for literature retrieval and orchestrating the overall pipeline. This allows the system to handle thousands of concurrent operations efficiently without being blocked by network latency.

Parallel Multi-GPU Inference: To handle the computationally intensive model inference, we spawn separate `ModelWorker` processes for each required service (e.g., reranking, generation). Each worker is pinned to a specific GPU or set of GPUs as defined in the `configs.json` file. Within each worker, we use the vLLM engine, a state-of-the-art serving library that employs techniques like PagedAttention to achieve high-throughput, low-latency inference.

Inter-Process Communication: The main `asyncio` loop communicates with the `ModelWorker` processes using a robust queue-based system (`multiprocessing.Queue`). A

request, tagged with a unique ID, is placed on a request queue. The main loop then awaits an `asyncio.Future` associated with that ID. The worker process retrieves the request, performs the inference, and places the result on a response queue. A dedicated listener task in the main loop listens for responses and resolves the corresponding `Future`, seamlessly bridging the asynchronous and multi-process components.

4 System Evaluation

We conduct a multi-faceted evaluation to assess the quality of our MedTutor system. Given our focus on the radiology domain, the evaluation is done by three radiologists who scored the outputs on a 5-point Likert scale (1=Poor, 5=Excellent). Annotation guidelines and the annotation interface design are detailed in Appendices F and G. We evaluate both the intermediate "upstream" components of our pipeline and the final "downstream" generated content. Furthermore, we investigate the feasibility of using an LLM-as-a-Judge as a proxy for human evaluation of the AI generated educational content by analyzing its agreement with our human experts.

4.1 Upstream Component Quality

First, we evaluate the quality of the upstream components that feed into the final generator: search query extraction and retrieved paper relevance. This evaluation was conducted on a set of 50 clinical cases. As shown in Table 1, human experts found the search queries extracted by the system to be highly appropriate (Human Avg. score of 3.73). However, they were more critical of the relevance of the retrieved academic papers, giving an average score of 2.88. This suggests that while the system correctly identifies the main topics, the unfiltered, live-retrieved literature can contain articles that are not perfectly aligned with the specific clinical context of the report. In contrast, the LLM judges rated the paper relevance significantly higher (LLM Avg. 4.20), indicating a divergence in the assessment of contextual relevance between human experts and automated metrics.

4.2 Downstream Generation Quality

The primary evaluation focused on the quality of the final, user-facing outputs: textbook summaries, MCQs, and educational material. Three radiologists annotated the outputs from two generator models, Llama-3.3-70B-Instruct and MedGemma-27B. The detailed results are presented in Table 2.

Evaluator	Query Appropriateness	Paper Relevance
Human Evaluators	**3.73**	**2.88**
MedGemma-27B	3.73	4.34
GPT-4.1-mini	4.15	4.52
Gemini-2.5-Flash	4.27	4.58
Gemini-2.5-Pro	4.03	3.37
LLM Avg.	**4.05**	**4.20**

Table 1: Comparison of evaluator scores. The " Evaluators" row represents the combined results from two independent radiologists (n=50 each).

Both models produced high-quality outputs according to our expert evaluators. Llama 3.3-70B-Instruct achieved a respectable average human score of 3.44, demonstrating its capability in synthesizing complex medical information into educational content. MedGemma-27B, a model more specialized for the medical domain, performed slightly better, with an average human score of 3.65. The experts particularly noted the higher quality of the MCQs generated by MedGemma-27B (3.53) compared to those from Llama-3.3-70B-Instruct (3.11). This suggests that the domain-specific nature of MedGemma-27B provides a distinct advantage in generating educational content, such as plausible distractors for multiple-choice questions.

When comparing human evaluations to the LLM-as-a-Judge scores, we note an interesting trend. The LLM judges also preferred MedGemma-27B over Llama 3.3, aligning with the relative ranking of the human experts. However, the LLMs consistently assigned higher absolute scores than the human radiologists. This suggests that while LLM-as-a-Judge can be a valuable tool for scalable, relative comparisons between models, its scoring calibration differs from that of human experts, indicating a tendency for score inflation. These findings suggest a promising path toward semi-automated evaluation while reinforcing the role of human experts as the gold standard for assessing clinical utility. Full LLM-as-a-Judge results are in Tab A.

4.3 Inter-Annotator Agreement

To ensure the reliability of our human evaluations, we measured the inter-annotator agreement (IAA) between the two board-certified radiologists using Krippendorff's Alpha (Krippendorff, 2011).

Model	Evaluator	Textbook Summary	Educational Material	MCQ Quality	Overall Average
Llama 3.3-70B-Instruct	**Human Evaluators**	**3.43**	**3.78**	**3.11**	**3.44**
	MedGemma-27B	3.64	3.66	3.79	3.70
	GPT-4.1-mini	4.34	4.50	4.19	4.34
	Gemini-2.5-Flash	2.82	3.58	4.08	3.49
	Gemini-2.5-Pro	3.95	4.28	4.14	4.12
	LLM Avg.	**3.69**	**4.01**	**4.05**	**3.91**
MedGemma-27B	**Human Evaluators**	**3.58**	**3.84**	**3.53**	**3.65**
	MedGemma-27B	3.65	4.09	4.22	3.99
	GPT-4.1-mini	4.21	4.79	4.60	4.53
	Gemini-2.5-Flash	3.05	4.61	4.47	4.04
	Gemini-2.5-Pro	3.84	4.18	4.15	4.06
	LLM Avg.	**3.69**	**4.42**	**4.36**	**4.16**

Table 2: Main Generation Task Quality: Direct Comparison of Human Expert and LLM-as-a-Judge Evaluations. The "Human Evaluators" scores represent the combined results from three independent radiologists (n=50 each). All scores are on a 1-5 scale (5=best).

The alpha coefficient is calculated as: $\alpha = 1 - \frac{D_o}{D_e}$ Here, D_o is the observed disagreement, calculated as the average difference between the ratings from each human annotator, A_1 and A_2, across all M evaluated items. Specifically, if $r_{i,1}$ and $r_{i,2}$ are the ratings for item i from A_1 and A_2 respectively, then:

$$D_o = \frac{1}{M} \sum_{i=1}^{M} \delta^2(r_{i,1}, r_{i,2})$$

D_e represents the disagreement expected by chance, calculated based on the individual rating distributions of A_1 and A_2. For the difference function δ^2, we first recoded the 1-to-5 Likert scale ratings into a 3-point interval scale (1-2 $\rightarrow$ 1; 3 $\rightarrow$ 2; 4-5 $\rightarrow$ 3) and then applied a squared difference: $\delta^2(u, v) = (u - v)^2$.

The results, presented in Table 4, show a range of agreement levels. We observed good agreement for the *Textbook Summary* from MedGemma-27B ($\alpha = 0.661$) and fair agreement for *Paper Relevance* ($\alpha = 0.493$).

Overall, our annotators demonstrated moderate to good agreement across most tasks (with the exception of MCQ quality), which is in line with agreement levels reported in prior work on high-quality datasets (Liu et al., 2024; Bavaresco et al., 2025). The lower agreement for MCQ evaluation ($\alpha = 0.048$) suggests that the criteria for this specific task may require more detailed guidelines to improve consistency.

5 Conclusion

In this work, we introduce **MedTutor**, a novel, open-source system designed to augment clinical education by transforming clinical reports into structured, evidence-backed learning modules. Our system addresses the critical challenges of factual accuracy and knowledge freshness in medical AI by employing a sophisticated RAG pipeline. This pipeline features a hybrid retrieval mechanism that synthesizes knowledge from both foundational medical textbooks and real-time academic literature, ensuring the generated educational modules are both reliable and current.

Our rigorous evaluation, conducted by board-certified radiologists, confirmed that MedTutor can produce high-quality, clinically valuable educational content. Furthermore, our large-scale LLM-as-a-Judge analysis revealed a moderate but promising correlation with human expert judgments, suggesting a viable path toward scalable automated evaluation while underscoring the continued importance of expert oversight.

By publicly releasing the MedTutor system, its user interface, and the comprehensive evaluation dataset, we make two key contributions. First, we provide a practical tool that can be immediately adapted by other institutions to enhance their own training programs. Second, we offer a valuable benchmark and framework for future research into building trustworthy and effective generative AI systems for the high-stakes medical domain. We believe this work represents a significant step toward fostering more effective and efficient clinician-AI

collaboration in medical education.

6 Limitations

While MedTutor demonstrates a promising approach to augmenting medical education, we acknowledge several limitations that offer avenues for future work.

First, our evaluation is primarily focused on the domain of radiology. Although the system's architecture is designed to be generalizable, its effectiveness and the nuances of its application in other medical specialties with different reporting styles and knowledge structures, such as pathology or cardiology, have not yet been explored. Future studies should assess the adaptability and performance of MedTutor across a broader range of clinical domains.

Second, the human evaluation, while rigorous and conducted by domain experts, was performed on a dataset of 50 clinical cases. A larger-scale study involving a greater number of cases and a more diverse cohort of radiologists would be beneficial to further validate our findings and provide more robust statistical power to the conclusions drawn.

Finally, our analysis of inter-annotator agreement and the LLM-as-a-Judge evaluations highlights challenges in consistently generating high-quality subjective content. The lower agreement scores for MCQs, for instance, suggest that these outputs require further refinement. This indicates that more advanced prompting techniques, fine-tuning of the generator models, or more sophisticated evaluation guidelines may be necessary to improve the reliability and educational value of these more complex, creative tasks.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C1352).

References

Alaa A. Abd-alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, P. Healy, Syed Latifi, S. Aziz, R. Damseh, Sadam Alabed Alrazak, and Javaid Sheikh. 2023. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Medical Education*, 9.

Michael Anderson and S. Anderson. 2019. How should ai be developed, validated, and implemented in patient care? *AMA journal of ethics*, 21 2:E125–130.

Natsuko Atsukawa, Hiroyuki Tatekawa, Tatsushi Oura, Shunichi Matsushita, Daisuke Horiuchi, H. Takita, Yasuhito Mitsuyama, Ayako Omori, T. Shimono, Yukio Miki, and D. Ueda. 2024. Evaluation of radiology residents' reporting skills using large language models: An observational study. *arXiv preprint arXiv:2404.56789*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *Preprint*, arXiv:2406.18403.

J. Bednarczyk, M. Pauls, Jason A. Fridfinnson, and E. Weldon. 2014. Characteristics of evidence-based medicine training in royal college of physicians and surgeons of canada emergency medicine residencies - a national survey of program directors. *BMC Medical Education*, 14:57 – 57.

Dawn MT Bravata, Stephen J Huot, Hadley S Abernathy, K. Skeff, and D. Bravata. 2003. The development and implementation of a curriculum to improve clinicians' self-directed learning skills: a pilot project. *BMC Medical Education*, 3:7 – 7.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *Preprint*, arXiv:2405.19538.

Rong Chen, Siyun Zhang, Yiyi Zheng, Qiuhua Yu, and Chu-huai Wang. 2025. Enhancing treatment decision-making for low back pain: A novel framework integrating large language models with retrieval-augmented generation technology. *arXiv preprint arXiv:2501.34567*.

Wolfgang. Dahnert and Inc. Ovid Technologies. 2017. *Radiology review manual*, 8th ed. edition. Wolters Kluwer, Philadelphia.

D. Daniel, Sue E. Poynter, C. Landrigan, C. Czeisler, J. Burns, and T. Wolbrink. 2020. Pediatric resident engagement with an online critical care curriculum during the intensive care rotation*. *Pediatric Critical Care Medicine*, 21:986 – 991.

Yining Huang, Keke Tang, and Meilian Chen. 2024. A comprehensive survey on evaluating large language model applications in the medical industry. *arXiv preprint arXiv:2401.12345*.

A. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng. 2024. Mimic-cxr database (version 2.1.0).

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2).

Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, and 1 others. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. Technical report, Annenberg School for Communication, University of Pennsylvania.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. *ArXiv*, abs/2311.03731.

Yixin Liu, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *Preprint*, arXiv:2311.09184.

Keer Lu, Zheng Liang, Da Pan, Shusen Zhang, Xin Wu, Weipeng Chen, Zenan Zhou, Guosheng Dong, Bin Cui, and Wentao Zhang. 2025. Med-r^2: Crafting trustworthy llm physicians via retrieval and reasoning of evidence-based medicine. *arXiv preprint arXiv:2505.89012*.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-05-20.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *Preprint*, arXiv:2303.13375.

M. Rogers, Michelle Zeidan, Zac Flinders, A. Presson, and R. Burks. 2019. Educational resource utilization by current orthopaedic surgical residents: A nationwide survey. *JAAOS Global Research & Reviews*, 3.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, and 13 others. 2023. Towards generalist biomedical ai. *Preprint*, arXiv:2307.14334.

Xiao Wang, Mengjue Tan, Qiao Jin, Guangzhi Xiong, Yu Hu, Aidong Zhang, Zhiyong Lu, and Minjia Zhang. 2025. Medcite: Can language models generate verifiable text for medicine? *arXiv preprint arXiv:2503.67890*.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023. Augmenting black-box llms with medical textbooks for biomedical question answering. *arXiv preprint arXiv:2308.12345*.

Adrienne A. Williams and Shana O. Ntiri. 2018. An online, self-directed curriculum of core research concepts and skills. *MedEdPORTAL : the Journal of Teaching and Learning Resources*, 14.

Qianqian Xie, E. Schenck, He S. Yang, Yong Chen, Yifan Peng, and Fei Wang. 2023. Faithful ai in medicine: A systematic review with large language models and beyond. *medRxiv*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. *Preprint*, arXiv:2402.13178.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024b. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2403.45678*.

Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2024. Retrieval-augmented generation for generative artificial intelligence in medicine. *Preprint*, arXiv:2406.12449.

Xiaoman Zhang, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2025a. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. *Preprint*, arXiv:2505.00228.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Weibing Zheng, Laurah Turner, Jess Kropczynski, Murat Ozer, Tri Nguyen, and S. Halse. 2025. Llm-as-a-fuzzy-judge: Fine-tuning large language models as a clinical evaluation judge with fuzzy logic. *arXiv preprint arXiv:2504.78901*.

Yunqi Zhu, Wen Tang, Ying Sun, and Xuebing Yang. 2024. The potential of llms in medical education: Generating questions and answers for qualification exams. *arXiv preprint arXiv:2402.23456*.

Appendix Contents

A LLM-as-a-Judge Evaluation

Model	Judge	Textbook Summary	Educational Material	MCQ	Average
Llama 3.1-8B-Instruct	MedGemma-27B	3.64 (±0.95)	3.49 (±1.10)	3.69 (±1.05)	3.61
	GPT-4.1-mini	4.06 (±0.85)	4.18 (±0.90)	3.86 (±0.92)	4.03
	Gemini-2.5-Pro	3.59 (±1.15)	3.55 (±1.20)	3.92 (±1.18)	3.69
	Gemini-2.5-Flash	3.64 (±1.30)	3.49 (±1.25)	3.88 (±1.28)	3.67
	Avg. (Judges)	**3.73**	**3.68**	**3.84**	**3.75**
Qwen3-8B	MedGemma-27B	3.39 (±1.01)	4.01 (±0.95)	3.78 (±0.88)	3.73
	GPT-4.1-mini	3.42 (±0.90)	4.49 (±0.75)	4.22 (±0.81)	4.04
	Gemini-2.5-Pro	3.45 (±1.10)	4.11 (±0.99)	3.81 (±0.95)	3.79
	Gemini-2.5-Flash	3.39 (±1.25)	4.01 (±1.15)	3.75 (±1.05)	3.72
	Avg. (Judges)	**3.41**	**4.16**	**3.89**	**3.82**
Llama-4-Scout-17B-16E-Instruct	MedGemma-27B	3.68 (±0.88)	4.08 (±0.85)	3.85 (±0.80)	3.87
	GPT-4.1-mini	4.30 (±0.70)	4.28 (±0.65)	4.18 (±0.72)	4.25
	Gemini-2.5-Pro	3.71 (±0.95)	4.15 (±0.90)	4.01 (±0.88)	3.96
	Gemini-2.5-Flash	3.68 (±1.10)	4.08 (±1.05)	3.95 (±1.00)	3.90
	Avg. (Judges)	**3.84**	**4.15**	**4.00**	**4.00**
Qwen3-32B	MedGemma-27B	3.55 (±0.75)	4.64 (±0.60)	3.99 (±0.65)	4.06
	GPT-4.1-mini	3.99 (±0.65)	4.19 (±0.50)	4.48 (±0.55)	4.22
	Gemini-2.5-Pro	3.61 (±0.80)	4.70 (±0.70)	4.25 (±0.75)	4.19
	Gemini-2.5-Flash	3.55 (±1.20)	4.64 (±1.10)	4.18 (±1.00)	4.12
	Avg. (Judges)	**3.68**	**4.54**	**4.23**	**4.15**
Llama-3.3-70B-Instruct	MedGemma-27B	3.64 (±0.68)	3.66 (±0.61)	3.79 (±0.55)	3.70
	GPT-4.1-mini	4.34 (±0.72)	4.50 (±0.55)	4.19 (±0.60)	4.34
	Gemini-2.5-Pro	3.95 (±1.23)	4.28 (±0.38)	4.14 (±0.55)	4.12
	Gemini-2.5-Flash	2.82 (±1.45)	3.58 (±1.44)	4.08 (±1.46)	3.49
	Avg. (Judges)	**3.69**	**4.01**	**4.05**	**3.91**
MedGemma-27B	MedGemma-27B	3.65 (±0.88)	4.09 (±0.61)	4.22 (±0.60)	3.99
	GPT-4.1-mini	4.21 (±0.81)	4.79 (±0.48)	4.60 (±0.51)	4.53
	Gemini-2.5-Pro	3.84 (±1.18)	4.18 (±0.45)	4.15 (±0.72)	4.06
	Gemini-2.5-Flash	3.05 (±1.58)	4.61 (±0.90)	4.47 (±1.18)	4.04
	Avg. (Judges)	**3.69**	**4.42**	**4.36**	**4.16**

Table 3: Aggregated LLM-as-a-Judge evaluation results across all datasets, comparing different judges. The **Avg. (Judges)** row indicates the mean of scores across the judges. All scores are on a 1-5 scale (5=best). Llama-4-Scout-17B-16E-Instruct(Meta, 2025) was inferenced in FP8.

B Detailed Inter-Annotator Agreement

Model Context	Evaluation Metric	Krippendorff's Alpha (α)	Pairwise Kappa (κ)	% Exact Agreement	% Within $\pm$1 Point	Correlation (r)
Upstream	Keyword Appropriateness	0.335	0.627	41%	80%	0.709
	Paper Relevance	0.474	0.675	59%	95%	0.778
Llama3-70B	Textbook Summary	0.347	0.555	48%	84%	0.587
	Educational Material	-0.228	0.382	50%	94%	0.325
	MCQ	-0.159	0.222	29%	81%	0.375
MedGemma-2B	Textbook Summary	0.627	0.812	66%	96%	0.721
	Educational Material	0.354	0.673	72%	100%	0.589
	MCQ	0.114	0.629	46%	90%	0.596

Table 4: Detailed Inter-Annotator Agreement (IAA) between three radiologists across different evaluation tasks. Krippendorff's Alpha (α) and Avg. Pairwise Kappa (κ) measure reliability, while agreement percentages and Pearson correlation (r) provide further insight into rater consistency.

Overall, the agreement scores suggest that MedGemma-27B's outputs were evaluated more consistently by the radiologists than those from Llama3.3-70B. As shown in Table 4, MedGemma-27B's Textbook Summary achieved the highest reliability, with a Krippendorff's Alpha of 0.627, approaching the threshold for acceptable agreement, and a substantial average pairwise Kappa of 0.812. The upstream task of Paper Relevance also demonstrated moderate to substantial agreement across most measures.

Conversely, the outputs from Llama3.3-70B, particularly for more subjective tasks like Educational Material and MCQ evaluation, yielded negative Alpha values, indicating systematic disagreement among the raters (Figure 3). The evaluation of MCQs proved challenging for both models, though agreement was notably higher for MedGemma-27B (Figure 4). These findings highlight that while structured summarization tasks can achieve high inter-rater reliability, evaluating more complex, subjective-generative tasks may require more detailed guidelines to ensure rater consistency.

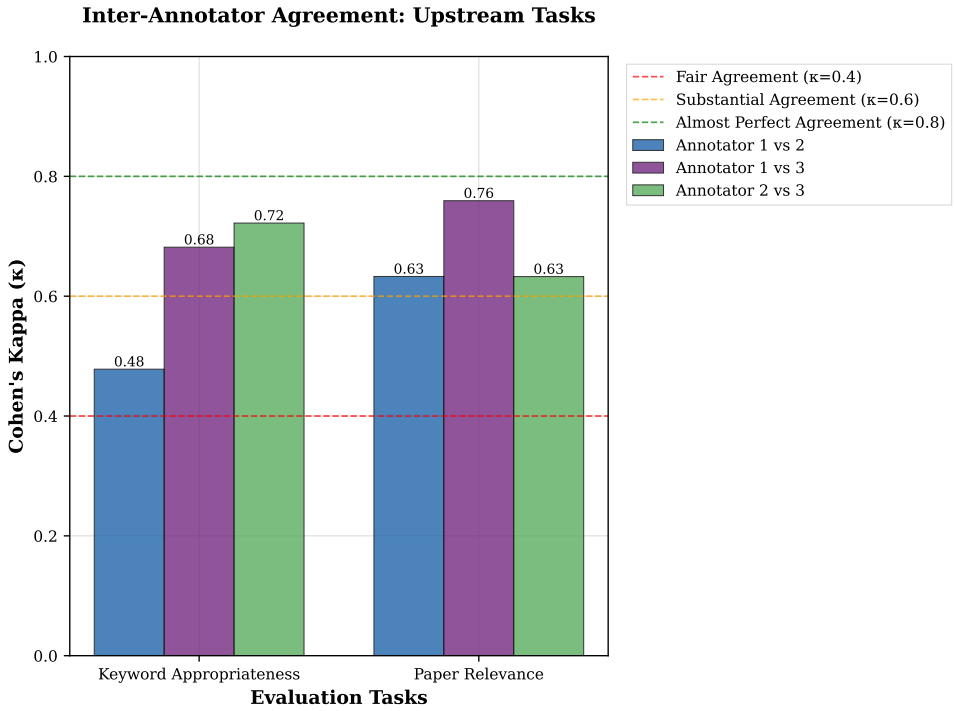


Figure 2: Pairwise Cohen's Kappa (κ) scores for Upstream Tasks. This figure shows the agreement between three pairs of annotators for keyword appropriateness and paper relevance.

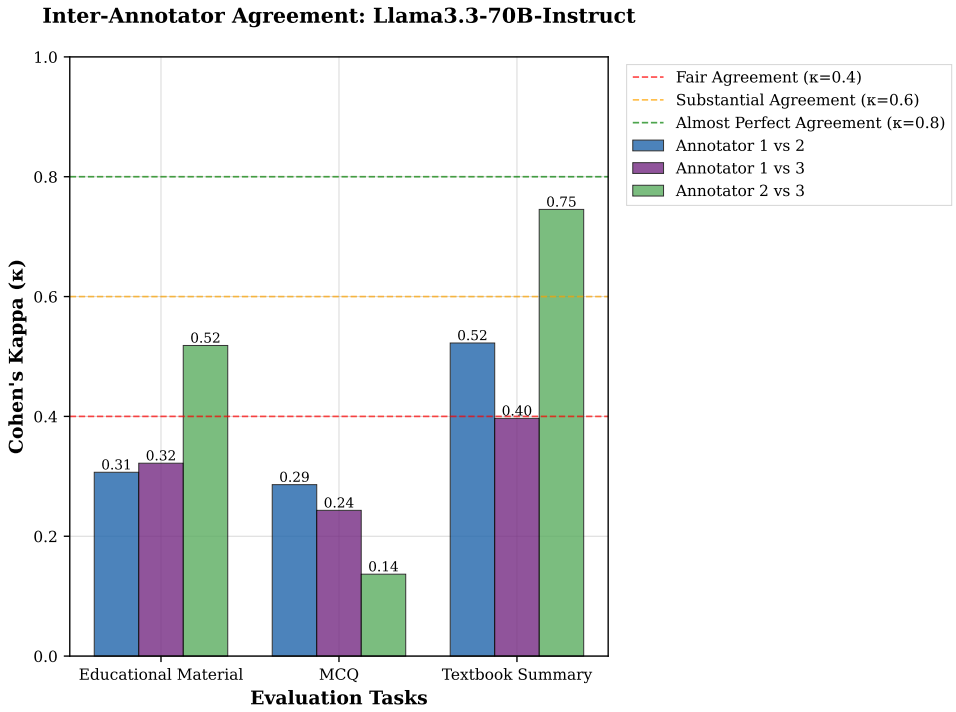


Figure 3: Pairwise Cohen's Kappa (κ) scores for Llama3.3-70B-Instruct Generated Content.

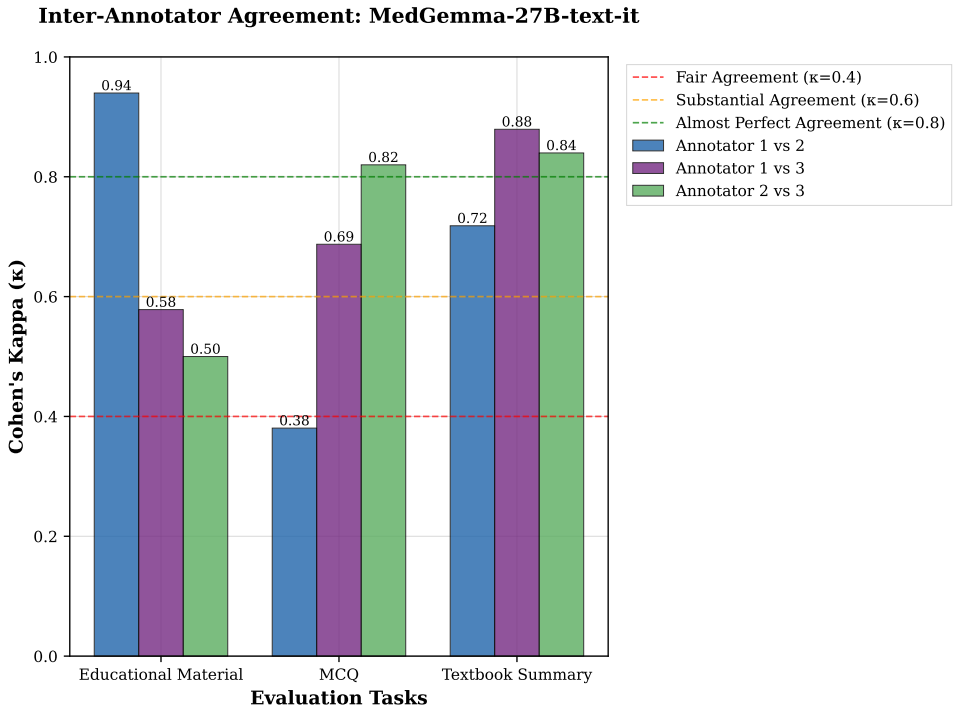


Figure 4: Pairwise Cohen's Kappa (κ) scores for MedGemma-27B-text-it Generated Content.

C MedTutor Dataset Samples and Public Dataset Information

This appendix provides the best inter-annotator agreement examples of the highest and lowest-scoring cases from the 50 cases evaluated by three expert radiologists. These cases were sampled (10 from each of the 5 datasets; Yale Internal, MIMIC-IV-note, MIMIC-CXR, CheXpert-Plus, ReXGradient-160K) and generated by two different models (`Llama-3.3-70B-Instruct` and `MedGemma-27B-text-it`). Each example includes the original clinical report and its corresponding generated report from MedTutor.

Also, we publicly release a large-scale dataset(total 144K) generated by our system. This includes reports from **CheXpert-Plus**, **MIMIC-IV-note**, and **MIMIC-CXR** (2,000 reports each), processed by six different generator models and 4 evaluator models. Due to licensing and de-identification challenges, the Yale-internal and ReXGradient datasets are not included in the public release.

C.1 Highest-scoring case generated by `Llama-3.3-70B-Instruct`.

C.2 Lowest-scoring case generated by `Llama-3.3-70B-Instruct`.

C.3 Highest-scoring case generated by `MedGemma-27B-text-it`.

C.4 Lowest-scoring case generated by `MedGemma-27B-text-it`.

C.1 Highest-Scoring Case with Llama-3.3-70B-Instruct

Case Information

Dataset: `MIMIC-IV-note`
Generator Model: `Llama-3.3-70B-Instruct`
Case ID: `19287224-RR-6`

Original Radiology Report

INDICATION: NO_PO contrast; History: () with abd pain NO_PO contrast// abd pain r/o appendicitiis
TECHNIQUE: Single phase split bolus contrast: MDCT axial images were acquired through the abdomen and pelvis following intravenous contrast administration with split bolus technique.
Oral contrast was administered.
Coronal and sagittal reformations were performed and reviewed on PACS.
DOSE: Acquisition sequence:
1) Stationary Acquisition 7.5 s, 0.5 cm; CTDIvol = 35.2 mGy (Body) DLP = 17.6 mGy-cm.
2) Spiral Acquisition 7.3 s, 55.8 cm; CTDIvol = 9.8 mGy (Body) DLP = 548.4 mGy-cm.
Total DLP (Body) = 566 mGy-cm.
COMPARISON: None.
FINDINGS:
LOWER CHEST: Visualized lung fields are within normal limits. There is no evidence of pleural or pericardial effusion.
ABDOMEN:
HEPATOBILIARY: The liver demonstrates homogenous attenuation throughout. There is no evidence of focal lesions. There is no evidence of intrahepatic or extrahepatic biliary dilatation. The gallbladder is within normal limits.
PANCREAS: The pancreas has normal attenuation throughout, without evidence of focal lesions or pancreatic ductal dilatation. There is no peripancreatic stranding.
SPLEEN: The spleen shows normal size and attenuation throughout, without evidence of focal lesions.
ADRENALS: The right and left adrenal glands are normal in size and shape.
URINARY: The kidneys are of normal and symmetric size with normal nephrogram. There is no evidence of focal renal lesions or hydronephrosis. There is no perinephric abnormality.
GASTROINTESTINAL: The stomach is unremarkable. Small bowel loops demonstrate normal caliber, wall thickness, and enhancement throughout. There are scattered colonic diverticula. The appendix is fluid-filled and dilated with wall hyperemia. A measures 13 mm with periappendiceal stranding and a small appendicular at the base. There is no periappendiceal abscess.
PELVIS: The urinary bladder and distal ureters are unremarkable. There is a small amount of free fluid in the pelvis.
REPRODUCTIVE ORGANS: The visualized reproductive organs are unremarkable.
LYMPH NODES: There is no retroperitoneal or mesenteric lymphadenopathy. There is no pelvic or inguinal lymphadenopathy.
VASCULAR: There is no abdominal aortic aneurysm. No significant atherosclerotic disease is noted.

BONES: There is no evidence of worrisome osseous lesions or acute fracture.
Moderate multilevel degenerative changes noted of the lumbar spine.
SOFT TISSUES: The abdominal and pelvic wall is within normal limits.
IMPRESSION:
1. Acute appendicitis.
2. Colonic diverticulosis without diverticulitis.
NOTIFICATION: Findings wet read to the electronic dashboard and discussed with ()

Generated Educational Materials

Keywords: Acute appendicitis, Colonic diverticulosis

Retrieved Paper 1: Retrospective case-control study to predict a potential underlying appendiceal tumor in an acute appendicitis context based on a CT-scoring system.
OBJECTIVES: To assess CT signs to discriminate an appendiceal tumor versus a non-tumoral appendix in an acute appendicitis context.
METHODS: A 10-year bicentric retrospective case-control study was performed in adults. Patients with a histopathological appendiceal tumor and appendicitis were paired for age and sex with patients with non-tumorous appendicitis (1/3 ratio, respectively). Two senior radiologists blindly analyzed numerous CT findings... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/33454458/ | Source: PubMed

Retrieved Paper 2: Dual energy CT in acute appendicitis: value of low mono-energy.
OBJECTIVES: To assess the potential role of low monoenergetic images in the evaluation of acute appendicitis.
METHODS: A retrospective study of 42 patients with pathology proven acute appendicitis underwent contrast-enhanced-CT conducted on a single-source-DECT before surgery. Attenuation, SNR, and CNR were calculated on both monoenergetic and conventional images and compared to 24 abdominal CT-scans with normal appendix. Representative... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/33992882/ | Source: PubMed

Retrieved Paper 3: Association between colorectal cancer and colonic diverticulosis: case-control study based on computed tomographic colonography.
BACKGROUND: The association between colorectal cancer (CRC) and diverticulosis is uncertain, and there is little data reported in Asia. This study examined the association of both diseases based on the computed tomographic colonography (CTC).
METHODS: This study was designed as a case-control study. Korean subjects undergoing CTC between April 2002 to April 2008 in Samsung Medical Center were eligible. Patients with histologically proven... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/21516446/ | Source: PubMed

Retrieved Paper 4: REVISED CONCEPTS ON DIVERTICULAR DISEASE OF THE COLON. II. SO-CALLED DIVERTICULITIS: DIVERTICULAR SIGMOIDITIS AND PERISIGMOIDITIS; DIVERTICULAR ABSCESS, FISTULA, AND FRANK PERITONITIS.
No Abstract
URL: https://pubmed.ncbi.nlm.nih.gov/14272097/ | Source: PubMed

Textbook Summary 1
Acute appendicitis is inflammation typically caused by luminal obstruction (fecolith, lymphoid hyperplasia) leading to ischemia. Radiologic findings on plain films are infrequent (<50%) but may include an appendicolith, cecal wall thickening, or signs of perforation like extraluminal gas or loss of fat planes. CT is highly accurate, showing a distended appendix (>7mm), wall thickening, periappendicular fat stranding, and potentially an appendicolith or complications like perforation. Nonvisualization of the appendix on CT in the setting of right lower quadrant pain makes appendicitis highly unlikely.

Textbook Summary 2
Colonic diverticulosis involves acquired herniations of the colonic mucosa and submucosa through the muscularis propria, typically appearing as thin-walled, air-filled outpouchings on imaging. Radiographically, these diverticula are often seen as rounded structures, sometimes with a bubbly appearance, and may show residual contrast from prior studies. Ultrasound can show thin-walled outpouchings with echogenic material and potential shadowing, while CT demonstrates rounded outpouchings containing air or contrast, often associated with circumferential colonic wall thickening due to muscular hypertrophy.

Assessment Question 1
According to the provided textbook summary, which CT finding is highly suggestive of acute appendicitis?

- A. Normal appendix diameter
- B. Appendix diameter > 7mm
- C. Absence of periappendicular fat stranding
- D. Nonvisualization of the appendix

Answer: B
Explanation: The textbook summary states that CT findings suggestive of acute appendicitis include a distended appendix (>7mm).

Assessment Question 2
Based on the provided abstracts, which CT finding is a positive predictive factor for an underlying appendiceal neoplasm in the context of acute appendicitis?

- A. Appendicolith
- B. Fat stranding
- C. Mural calcifications
- D. Appendix diameter < 15mm

Answer: C
Explanation: The abstract from the study on appendiceal tumors identifies mural calcifications as a positive predictive factor (OR = 47, p = 0.0001) for an underlying tumor in patients presenting with acute appendicitis.

Assessment Question 3
What is the characteristic radiographic appearance of colonic diverticula as described in the textbook summary?

- A. Thick-walled outpouchings filled with fluid
- B. Thin-walled outpouchings, often rounded, containing air or contrast
- C. Irregular thickening of the colonic wall
- D. Absence of air within the colon

Answer: B
Explanation: The textbook summary describes colonic diverticula as acquired herniations appearing as thin-walled, air-filled outpouchings on imaging, often rounded and potentially containing contrast.

Assessment Question 4
According to the provided abstract analyzing the association between colorectal cancer and diverticulosis, what was the main conclusion regarding this association?

- A. Colorectal cancer is strongly associated with diverticulosis.
- B. Diverticulosis is more common in patients with right-sided colon cancer.
- C. Colorectal cancer is not associated with diverticulosis.
- D. Computed tomographic colonography is not useful for assessing diverticulosis.

Answer: C
Explanation: The abstract from the case-control study using CT colonography concluded that colorectal cancer is not associated with diverticulosis, as the distribution of diverticulosis was similar between the cancer and control groups.

Final Educational Material
Radiology Feedback
 # Acute Appendicitis
• Imaging Findings: CT is the preferred modality. Key findings include an appendix diameter >7 mm, wall thickening, periappendiceal fat stranding, and potentially an appendicolith.
• Differential Diagnosis: Nonvisualization of the appendix on CT in the setting of right lower quadrant pain makes appendicitis highly unlikely.
• Advanced Techniques: Low monoenergetic CT images can enhance visualization of inflamed bowel wall and mucosal enhancement, potentially improving diagnostic accuracy.
• Clinical Correlation: While appendicitis is often caused by obstruction, CT can help identify potential underlying neoplasms, especially in atypical presentations. Features suggesting neoplasm include mural calcifications, an appendix mass, focal asymmetric wall abnormality, or diameter $\geq$ 15 mm.
 # Colonic Diverticulosis
• Imaging Findings: Diverticula appear as thin-walled, air-filled or contrast-filled outpouchings from the colon, typically rounded. CT demonstrates these outpouchings, often associated with circumferential colonic wall thickening due to muscular hypertrophy.
• Clinical Correlation: Diverticulosis itself is the presence of these outpouchings and is generally asymptomatic. It is not directly associated with an increased risk of colorectal cancer.

C.2 Lowest-Scoring Case with Llama-3.3-70B-Instruct

Case Information

Dataset: RexGradient-160K
Generator Model: Llama-3.3-70B-Instruct
Case ID: pGRDNRZB2HZ56F0ZL_aGRDNZSUP68GUCQN2_s1.2.826.0.1.3680043.8.498.16841148712021794680399240737673267273

Original Radiology Report

Indication: Fall yesterday. Fever this morning.
Findings: 4228 hours. The heart size and mediastinal contours are stable. There is stable mild subsegmental atelectasis at both lung bases. No confluent airspace opacity, pleural effusion or pneumothorax. Fracture of the proximal right humerus again noted.
Impression: Stable mild bibasilar atelectasis. No acute cardiopulmonary process.

Generated Educational Materials

Keywords: atelectasis, fracture

Retrieved Paper 1: Debunking a mythology: Atelectasis is not a cause of postoperative fever.
Most physicians appreciate that practicing medicine is a commitment to continuous learning. However, "learning" can be mistakenly understood as simply the acquisition of facts and new knowledge. But learning also necessitates the constant re-examination and challenging of one's existing body of knowledge, as misinformation persists when one's beliefs are not challenged or questioned in the light of new information. One example is the pervasive... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/39566396/ | Source: PubMed

Retrieved Paper 2: Use of artificial intelligence in triaging of chest radiographs to reduce radiologists' workload.
OBJECTIVES: To evaluate whether deep learning-based detection algorithms (DLD)-based triaging can reduce outpatient chest radiograph interpretation workload while maintaining noninferior sensitivity.
METHODS: This retrospective study included patients who underwent initial chest radiography at the outpatient clinic between June 1 and June 30, 2017. Readers interpreted radiographs with/without a commercially available DLD that detects nine... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/37615766/ | Source: PubMed

Retrieved Paper 3: Assessment of proximal tibial fractures with 3D FRACTURE (fast field echo resembling a CT using restricted echo-spacing) MRI-intra-individual comparison with CT.
OBJECTIVES: To evaluate the feasibility and diagnostic performance of a 3D FRACTURE (fast field echo resembling a CT using restricted echo-spacing) MRI sequence for the detection and classification of proximal tibial fractures compared with CT.
METHODS: We retrospectively included 126 patients (85 male; 39.6±14.5 years) from two centers following acute knee injury. Patients underwent knee MRI at 3T including FRACTURE-MRI. Additional CT was... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/40126605/ | Source: PubMed

Retrieved Paper 4: How I Do It: Evaluating Cardiac Implantable Devices and Noncardiac Mimics on Chest Radiographs.
Cardiac implantable electronic devices (CIEDs), including pacemakers and defibrillators, are increasingly used to manage various cardiac conditions. This article reviews the radiographic appearance, typical components, and placement of CIEDs, including newer technologies like leadless pacemakers and MRI-conditional devices. The article also highlights the imaging findings of common complications such as lead dislodgement, fracture, and... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/40358448/ | Source: PubMed

---

Textbook Summary 1
Atelectasis, or lung collapse, presents radiologically with increased lung density, vessel crowding, and potential fissure/mediastinal displacement. Obstructive atelectasis involves air resorption distal to a blockage (tumor, mucus plug, foreign body), while nonobstructive atelectasis retains some air. Other forms include passive (pleural effusion/pneumothorax), adhesive (surfactant deficiency), cicatrizing (fibrosis), and discoid/rounded atelectasis (often related to pleural inflammation or obstruction). Specific patterns, like the "Luftsichel" sign, can indicate left upper lobe collapse.

Textbook Summary 2
Enteropathy-associated T-cell lymphoma, a type of non-Hodgkin lymphoma, often presents with bowel wall thickening, ulceration, or strictures, particularly in the proximal small bowel. Radiologic findings may include circumferential wall thickening, mesenteric fat infiltration, and nonbulky lymphadenopathy, with a high frequency of FDG uptake on PET scans. Complications like bowel perforation are common, especially in Type II lymphoma, and differentiating it from large B-cell lymphoma or refractory celiac disease is crucial.

Assessment Question 1
According to the provided abstract debunking a mythology, what is the common misconception regarding atelectasis?

- A. Atelectasis is a common cause of postoperative consolidation.
- B. Atelectasis is a common cause of postoperative fever.
- C. Atelectasis is difficult to differentiate from pneumonia on imaging.
- D. Atelectasis is primarily caused by pleural effusion.

Answer: B
Explanation: The abstract explicitly states that the pervasive belief that postoperative atelectasis causes fever is a myth, despite ample evidence to the contrary.

Assessment Question 2
Which of the following is NOT a type of atelectasis mentioned in the textbook summary?

- A. Obstructive atelectasis
- B. Passive atelectasis
- C. Adhesive atelectasis
- D. Vascular atelectasis

Answer: D
Explanation: The textbook summary lists obstructive, nonobstructive, passive, adhesive, cicatrizing, and discoid/rounded atelectasis, but not vascular atelectasis.

Assessment Question 3
According to the abstract comparing FRACTURE-MRI and CT for proximal tibial fractures, what was the sensitivity of FRACTURE-MRI for detecting fractures?

- A. 78.2
- B. 85.7
- C. 91.5
- D. 97.1

Answer: C
Explanation: The abstract states that fractures were detected by FRACTURE-MRI with a sensitivity of 91.5

Assessment Question 4
The abstract on evaluating cardiac implantable devices mentions which potential complication related to device leads?

- A. Dislodgement
- B. Calcification
- C. Consolidation
- D. Cardiomegaly

Answer: A
Explanation: The abstract explicitly lists lead dislodgement, fracture, and perforation as common complications of cardiac implantable electronic devices.

Final Educational Material
Okay, here is the synthesized feedback focusing on atelectasis and fracture, based on the provided information:
 # Atelectasis
• Clinical Context: Atelectasis is commonly seen postoperatively and in the context of fever, but it is not a direct cause of fever itself. This is a persistent misconception.
• Imaging Appearance: Radiographically, atelectasis appears as increased lung density, often with crowding of pulmonary vessels. It can be subsegmental, as seen in this case.
• Differentiation: It's important to differentiate atelectasis from consolidation (infection), which typically presents with air bronchograms and may have a more lobar distribution.
• AI Application: AI-based triaging tools can help identify atelectasis on chest radiographs, potentially reducing radiologist workload while maintaining diagnostic sensitivity.
 # Fracture
• Imaging Modalities: While CT is the standard for fracture evaluation, specialized MRI sequences like FRACTURE-MRI are being investigated as alternatives, potentially reducing radiation exposure and costs, especially when MRI is already indicated.
• MRI Performance: Studies show FRACTURE-MRI has high sensitivity and specificity for detecting and classifying fractures, comparable to CT in certain locations like the proximal tibia.
• Complications: Imaging is crucial for identifying complications of implanted devices, such as lead fracture in pacemakers/defibrillators, which can be visualized on chest radiographs.

C.3 Highest-Scoring Case with MedGemma-27B-text-it

Case Information

Dataset: `CheXpert-Plus`
Generator Model: `MedGemma-27B-text-it`
Case ID: `23803`

Original Radiology Report

Findings: Two views of the chest demonstrate reticular opacities bilaterally, right greater than left, with a basilar predominance, likely secondary to underlying fibrotic lung disease. There is no evidence of pulmonary edema or pleural effusion. There is cardiomegaly. Left anterior chest wall cardiac pacer appears unchanged in position, with two ventricular leads.
Impression:
1. FINDINGS CONSISTENT WITH FIBROTIC LUNG DISEASE AS DESCRIBED ABOVE.
2. NO EVIDENCE OF PULMONARY EDEMA.
3. STABLE CARDIOMEGALY WITH STABLE ARRANGEMENT OF PACER LEADS.

Generated Educational Materials

Keywords: Fibrotic Lung Disease

Retrieved Paper 1: Collaborative radiologic and histopathologic assessment of fibrotic lung disease.
The idiopathic interstitial pneumonias (IIPs) are a seemingly disconnected collection of diseases usually associated with the presence of pulmonary fibrosis. Categorization of the IIPs continues to be problematic despite recent attempts to refine the diagnostic criteria and suggests that rather than separate diseases, these pneumonias represent a spectrum of injury and abnormal repair of the alveolar wall. Although the initiating injury or... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/20501710/ | Source: PubMed

Retrieved Paper 2: From the radiologic pathology archives: organization and fibrosis as a response to lung injury in diffuse alveolar damage, organizing pneumonia, and acute fibrinous and organizing pneumonia.
Organization, characterized by fibroblast proliferation, is a common and nearly universal response to lung injury whether it is focal or diffuse. Despite the vast range of injurious agents, the lung's response to injury is quite limited, with a similar pattern of reaction seen radiologically and histologically regardless of the underlying cause. Although there is a tendency to divide organization into distinct entities, the underlying injury to... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/24224590/ | Source: PubMed

Textbook Summary 1
In fibrotic lung disease, key radiologic findings include a reticulonodular pattern, ground-glass opacities, and honeycombing, with a basilar and subpleural distribution. Diagnostic considerations for fibrotic lung disease, such as usual interstitial pneumonia (UIP), involve identifying a combination of clinical and imaging features, including progressive dyspnea, clubbing, and bibasilar crackles, along with characteristic HRCT findings like traction bronchiectasis and macrocystic honeycombing. HRCT plays a crucial role in differentiating UIP from other idiopathic interstitial pneumonias and guiding biopsy sites for definitive diagnosis.

Assessment Question 1
What is a characteristic radiologic finding in fibrotic lung disease?

- A. Pulmonary edema
- B. Reticulonodular pattern
- C. Cardiomegaly
- D. Pleural effusion

Answer: B
Explanation: Fibrotic lung disease is characterized by specific radiologic findings, including a reticulonodular pattern, which is a key feature of this condition.

Assessment Question 2
What is the typical distribution of radiologic findings in fibrotic lung disease?

- A. Apical and central
- B. Basilar and subpleural
- C. Upper lobe and peripheral
- D. Diffuse and homogeneous

Answer: B
Explanation: Fibrotic lung disease typically shows a basilar and subpleural distribution of radiologic findings, which is an important characteristic for diagnosis.

Final Educational Material
Fibrotic Lung Disease
Fibrotic lung disease is characterized by key radiologic findings including a reticulonodular pattern, ground-glass opacities, and honeycombing, typically with a basilar and subpleural distribution. The diagnosis often involves identifying a combination of clinical features such as progressive dyspnea, clubbing, and bibasilar crackles, along with characteristic HRCT findings like traction bronchiectasis and macrocystic honeycombing. High-resolution computed tomography (HRCT) plays a crucial role in differentiating usual interstitial pneumonia (UIP) from other idiopathic interstitial pneumonias and in guiding biopsy sites for a definitive diagnosis. A collaborative diagnostic approach combining radiologic and histopathologic assessments can lead to a more reliable identification of the predominant pathways leading to pulmonary fibrosis, which has implications for therapy and future research directions.

C.4 Lowest-Scoring Case with MedGemma-27B-text-it

Case Information

Dataset: RexGradient-160K
Generator Model: MedGemma-27B-text-it
Case ID: pGRDNRZB2HZ56F0ZL_aGRDNZSUP68GUCQN2_s1.2.826.0.1.3680043.8.498.16841148712021794680399240737673267273

Original Radiology Report

Indication: Fall yesterday. Fever this morning.
Findings: 4228 hours. The heart size and mediastinal contours are stable. There is stable mild subsegmental atelectasis at both lung bases. No confluent airspace opacity, pleural effusion or pneumothorax. Fracture of the proximal right humerus again noted.
Impression: Stable mild bibasilar atelectasis. No acute cardiopulmonary process.

Generated Educational Materials

Keywords: atelectasis, fracture

Retrieved Paper 1: Debunking a mythology: Atelectasis is not a cause of postoperative fever.
Most physicians appreciate that practicing medicine is a commitment to continuous learning. However, "learning" can be mistakenly understood as simply the acquisition of facts and new knowledge. But learning also necessitates the constant re-examination and challenging of one's existing body of knowledge, as misinformation persists when one's beliefs are not challenged or questioned in the light of new information. One example is the pervasive... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/39566396/ | Source: PubMed

Retrieved Paper 2: Use of artificial intelligence in triaging of chest radiographs to reduce radiologists' workload.
OBJECTIVES: To evaluate whether deep learning-based detection algorithms (DLD)-based triaging can reduce outpatient chest radiograph interpretation workload while maintaining noninferior sensitivity.
METHODS: This retrospective study included patients who underwent initial chest radiography at the outpatient clinic between June 1 and June 30, 2017. Readers interpreted radiographs with/without a commercially available DLD that detects nine... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/37615766/ | Source: PubMed

Retrieved Paper 3: Assessment of proximal tibial fractures with 3D FRACTURE (fast field echo resembling a CT using restricted echo-spacing) MRI-intra-individual comparison with CT.
OBJECTIVES: To evaluate the feasibility and diagnostic performance of a 3D FRACTURE (fast field echo resembling a CT using restricted echo-spacing) MRI sequence for the detection and classification of proximal tibial fractures compared with CT.
METHODS: We retrospectively included 126 patients (85 male; 39.6±14.5 years) from two centers following acute knee injury. Patients underwent knee MRI at 3T including FRACTURE-MRI. Additional CT was... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/40126605/ | Source: PubMed

Retrieved Paper 4: How I Do It: Evaluating Cardiac Implantable Devices and Noncardiac Mimics on Chest Radiographs.
Cardiac implantable electronic devices (CIEDs), including pacemakers and defibrillators, are increasingly used to manage various cardiac conditions. This article reviews the radiographic appearance, typical components, and placement of CIEDs, including newer technologies like leadless pacemakers and MRI-conditional devices. The article also highlights the imaging findings of common complications such as lead dislodgement, fracture, and... (Omitted)
URL: https://pubmed.ncbi.nlm.nih.gov/40358448/ | Source: PubMed

---

Textbook Summary 1
Atelectasis is characterized by radiologic findings of increased lung density, crowding of pulmonary vessels, and displacement of fissures, hilum, and mediastinum, often with elevation of the hemidiaphragm. On MRI, obstructive atelectasis shows high signal intensity on T2WI, while nonobstructive atelectasis shows low signal intensity. The diagnosis of atelectasis considers various causes, including obstructive, nonobstructive, passive, adhesive, and cicatrizing types, each with distinct pathophysiology and imaging features.

Textbook Summary 2
Fracture is not a primary focus in the provided textbook pages, which mainly discuss lymphoma and malignant melanoma, but radiologic findings related to fractures are not mentioned. However, in the context of lymphoma, particularly enteropathy-associated T-cell lymphoma, complications such as bowel perforation can occur, which may be associated with a fracture of the bowel wall. Diagnostic considerations for lymphoma include imaging findings like circumferential wall thickening on CT and FDG uptake on PET.

Assessment Question 1
What is a characteristic radiologic finding of atelectasis?

- A. Decreased lung density
- B. Increased lung density
- C. Dilatation of pulmonary vessels
- D. Elevation of the diaphragm without crowding of pulmonary vessels

Answer: B
Explanation: Atelectasis is characterized by radiologic findings of increased lung density, crowding of pulmonary vessels, and displacement of fissures, hilum, and mediastinum.

Assessment Question 2
Which of the following statements about postoperative atelectasis is supported by the provided context?

- A. Atelectasis is a common cause of postoperative fever.
- B. Atelectasis is rarely seen in postoperative patients.
- C. The belief that atelectasis causes postoperative fever is a pervasive myth.
- D. Atelectasis always leads to significant respiratory complications.

Answer: C
Explanation: The context mentions that the belief that postoperative atelectasis causes fever is a myth, despite ample evidence to the contrary, as discussed in the abstract titled "Debunking a mythology: Atelectasis is not a cause of postoperative fever."

Assessment Question 3
What is the sensitivity of FRACTURE-MRI in detecting fractures, as mentioned in the provided context?

- A. 70
- B. 80
- C. 91.5
- D. 95

Answer: C
Explanation: The context states that fractures were detected by FRACTURE-MRI with a sensitivity of 91.5

Assessment Question 4
What is a potential advantage of using FRACTURE-MRI for diagnosing proximal tibial fractures, as mentioned in the context?

- A. Higher radiation exposure
- B. Reduced accuracy in fracture classification
- C. Reduced costs and radiation exposure
- D. Increased need for CT scans

Answer: C
Explanation: The context suggests that FRACTURE-MRI might be a helpful alternative to computed tomography in an acute trauma setting by reducing costs and radiation exposure in patients requiring a preoperative MRI anyway.

Final Educational Material
Atelectasis
Atelectasis is characterized by increased lung density, crowding of pulmonary vessels, and displacement of fissures, hilum, and mediastinum, often with elevation of the hemidiaphragm. It's essential to differentiate atelectasis from consolidation, as their imaging characteristics and clinical implications differ. Atelectasis can be caused by various factors, including obstructive, nonobstructive, passive, adhesive, and cicatrizing types. Notably, atelectasis is not a cause of postoperative fever, contrary to a common myth. The use of artificial intelligence in triaging chest radiographs can help reduce the workload of radiologists while maintaining sensitivity for detecting atelectasis and other radiologic findings.
Fracture
Fracture diagnosis and classification can be accurately performed using 3D FRACTURE (fast field echo resembling a CT using restricted echo-spacing) MRI sequences, which offer a feasible alternative to CT scans in acute trauma settings. This can help reduce radiation exposure and costs, especially in patients requiring preoperative MRI. When evaluating fractures, particularly in the context of cardiac implantable devices, it's crucial to identify potential complications such as lead dislodgement, fracture, and perforation. Accurate identification of these devices and their positioning on chest radiographs is vital for early detection of complications and directing appropriate patient care.

D Default System Prompts for MedTutor

D.1 Keyword Generation Prompt

Keyword Generation Prompt

System Prompt:

```
You are an expert medical language model. Given the full radiology report
and the extracted Impression section, extract all specific disease names,
diagnostic labels, and named pathological entities mentioned or implied in
either section. Focus only on established or suspected diagnoses, such as
named conditions.

Only include diagnoses that are positively identified or suspected in the
report. Do not include any conditions that are explicitly ruled out,
negated, or stated as absent.

Do not include general phrases, symptoms, or clinical findings that are not
formal diagnoses.

Output your answer as a valid JSON object with the following format:

{ "keywords": ["diagnosis 1", "diagnosis 2", "diagnosis 3"] }

If no diagnoses are present, return:

{
"keywords": []
}
```

User Instruction Template:

```
Final_report: {full_report_text}
Impression: {impression_text}
```

D.2 Textbook Summary Prompt

Textbook Summary Prompt

System Prompt:

```
You are a concise and accurate radiology assistant, skilled in summarizing
medical texts.
```

User Instruction Template:

```
Please summarize the following textbook pages focusing on the keyword '{keyword}'.
The summary should highlight key radiologic findings and diagnostic
considerations. Be concise, using 2-3 sentences and your own words.
Output only the summary text itself, with no additional conversational
text or headers.

Textbook Pages Content:
{pages_block_text}
```

D.3 MCQ Generation Prompt

Multiple Choice Q&A Generation Prompt

System Prompt:

```
You are a specialized AI assistant for creating multiple-choice questions (MCQs)
for radiology education. You must focus *exclusively* on the provided
**Primary Diagnostic Keywords**.
```

User Instruction Template:

```
### Primary Diagnostic Keywords to Focus On:
- {keywords_list_str}

### Full Context (for reference)
{mcq_input_context}

### Your Task
Based *only* on the provided context, generate 2 multiple-choice questions
**for each Primary Diagnostic Keyword listed above**. Do not generate questions
for any other terms or topics mentioned in the context. Each question must
test understanding of the information related to the primary keywords.

Follow this format exactly:

### Multiple Choice Questions

#### {{Diagnosis Keyword 1}}

Q1. {{Question stem}}
A. {{Option A}}
B. {{Option B}}
C. {{Option C}}
D. {{Option D}}
Answer: {{Correct Option Letter}}
Explanation: {{Brief explanation based on the provided context.}}
```

D.4 Educational Material Generation Prompt

Educational Material Prompt

System Prompt:

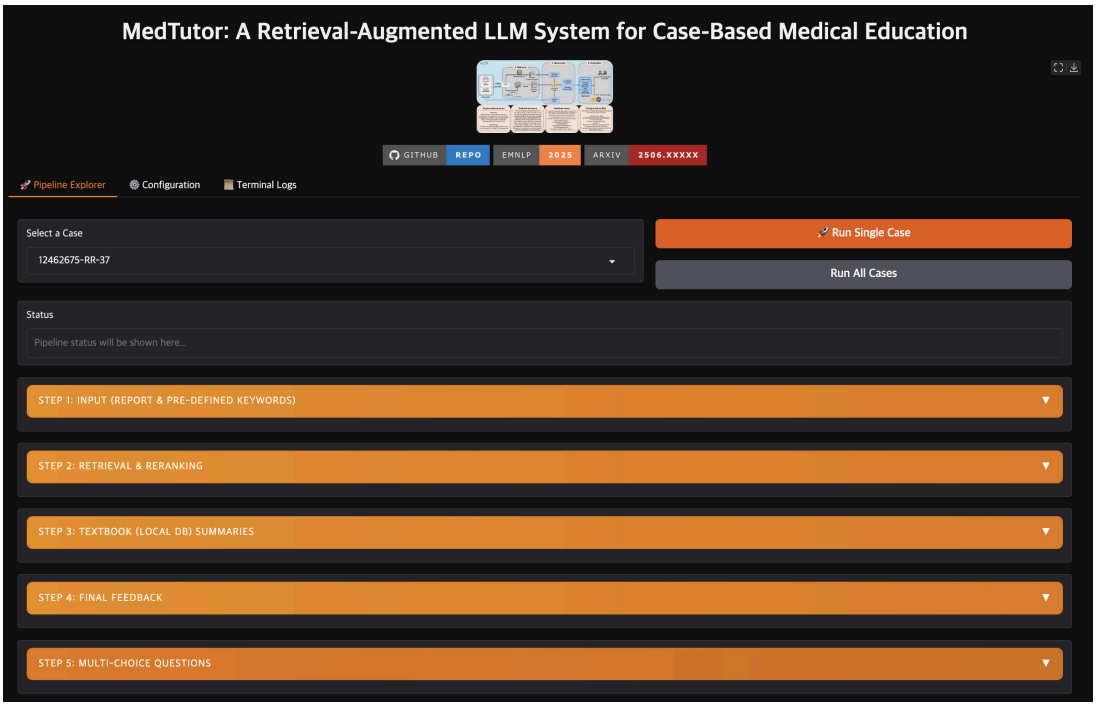
```
You are an expert radiology AI assistant. Your task is to synthesize the
provided information into concise, educational feedback focused *only* on the
primary diagnostic keywords provided. Do not explain or elaborate on other
terms from the original report unless they are directly relevant to the
primary keywords.
```

User Instruction Template:

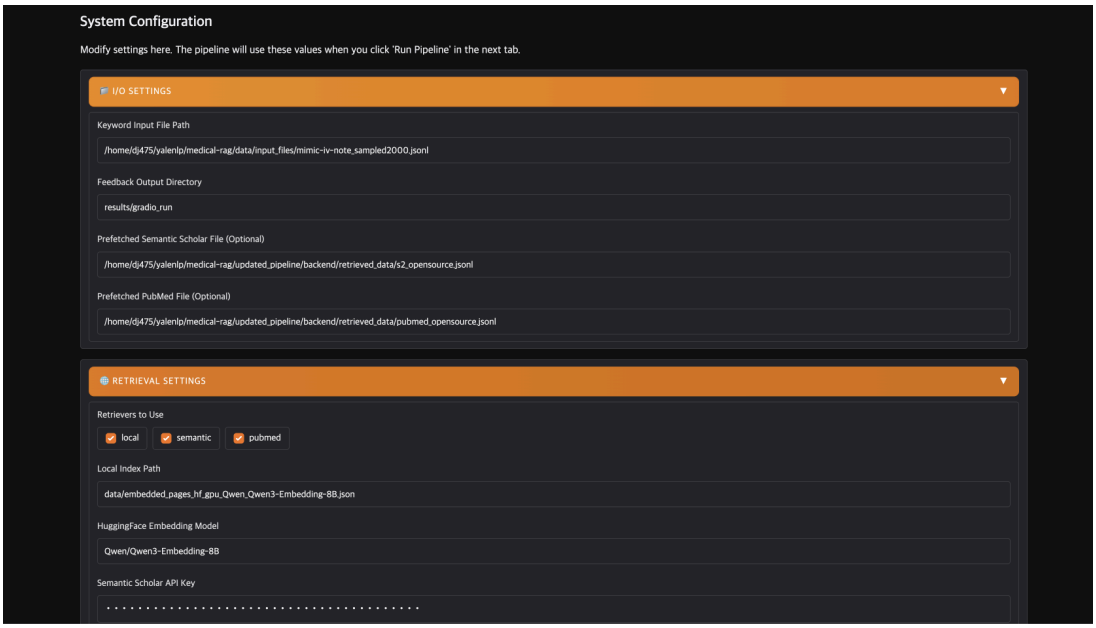
```
### Primary Diagnostic Keywords
- {keywords_list_str}

### Original Reviewer Report (for context only)
```

```
{original_reviewer_report}

### Supporting Educational Material
{user_block_for_final_stages}

### Your Task
Based on all the information above, provide a concise, synthesized feedback.
Structure your response with a section for each **Primary Diagnostic Keyword**.
Focus only on clinical teaching points and imaging pearls related to these
primary keywords.
```

E MedTutor System UI

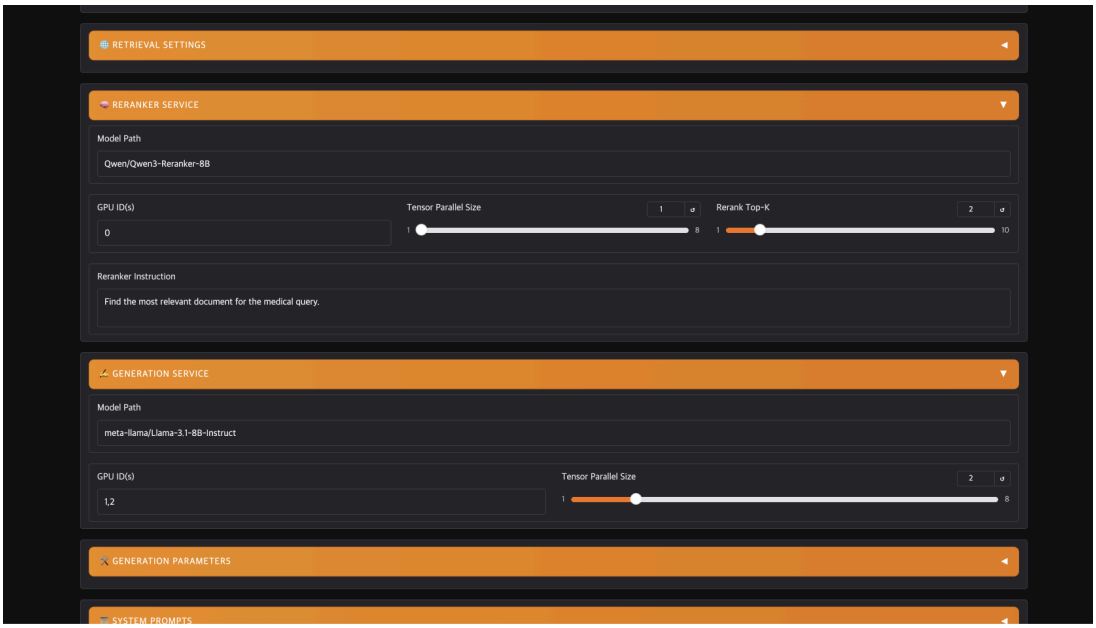


(a) Main user interface of MedTutor.

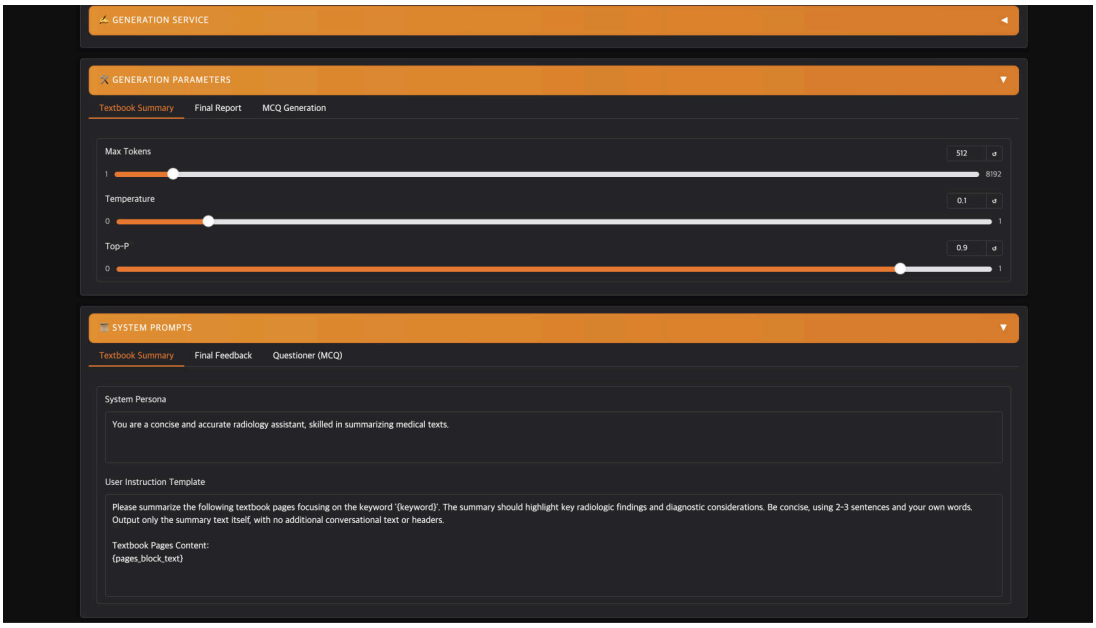


(b) Configuration settings for model selection and system prompts.

Figure 5: The MedTutor UI (Part 1 of 2): Main dashboard and initial configuration settings.



(a) Further configuration for data sources and retrieval.



(b) Finalizing configuration and execution options.

Figure 6: The MedTutor UI (Part 2 of 2): Additional configuration panels for data processing and task execution.

F Human Annotation Guideline

I. Evaluation of Information Quality per Keyword

For each diagnostic keyword identified from the original report, we evaluate the following components:

F.1 Retrieved & Reranked Academic Papers

- **Relevance to Keyword & Original Report:** How directly related is each paper or retrieved snippet to the given keyword and the context of the original radiology report?

F.2 Generated Textbook Summary

- **Accuracy & Factuality:** Is the summary an accurate and factual representation of information related to the keyword (compared to general radiology knowledge or, if available, the source textbook)?
- **Helpfulness & Relevance:** Is the summary helpful and related to the case report provided as input?
- **Coverage of Key Information:** Does the summary include the most critical information (e.g., key imaging findings, diagnostic criteria) related to the keyword?

F.3 Example Multiple Choice Questions

- **Relevance & Correctness:** Are the questions relevant to the keyword? Is the answer provided and rationale correct? Are the answer choices relevant?

II. Evaluation of the "Educational material" Paragraph (per Keyword)

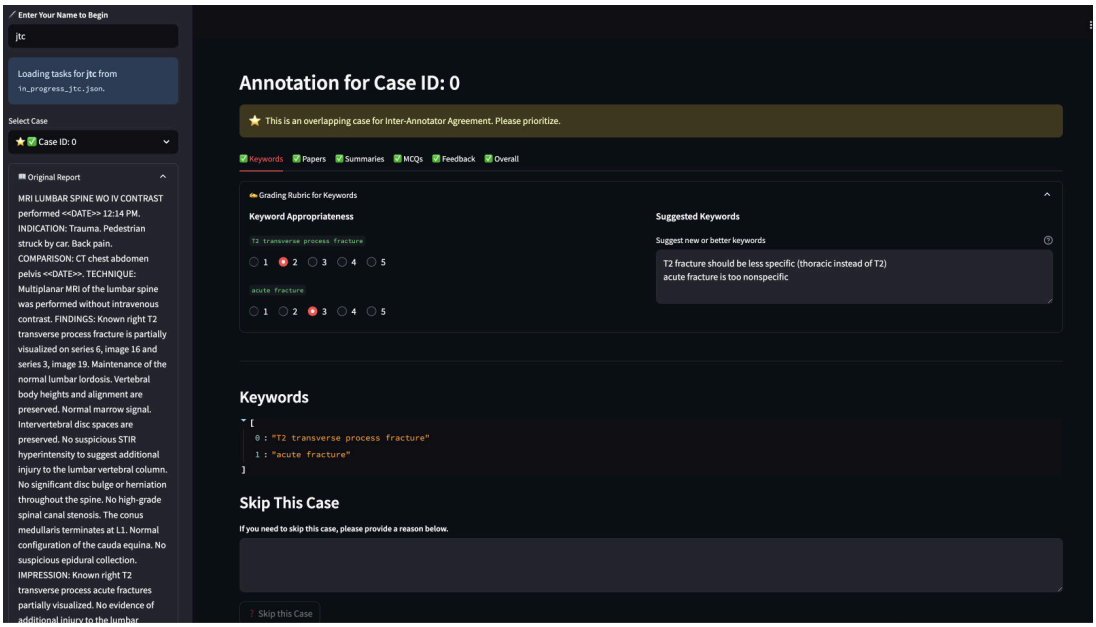
- **Clinical & Educational Utility:** How clinically relevant, accurate, and educationally valuable is this educational material paragraph for a radiology trainee in understanding the keyword within the context of the original report? (This encompasses quality, clinical insight, contextual appropriateness, and trustworthiness.)

III. Evaluation of Overall Educational Material Structure & Quality

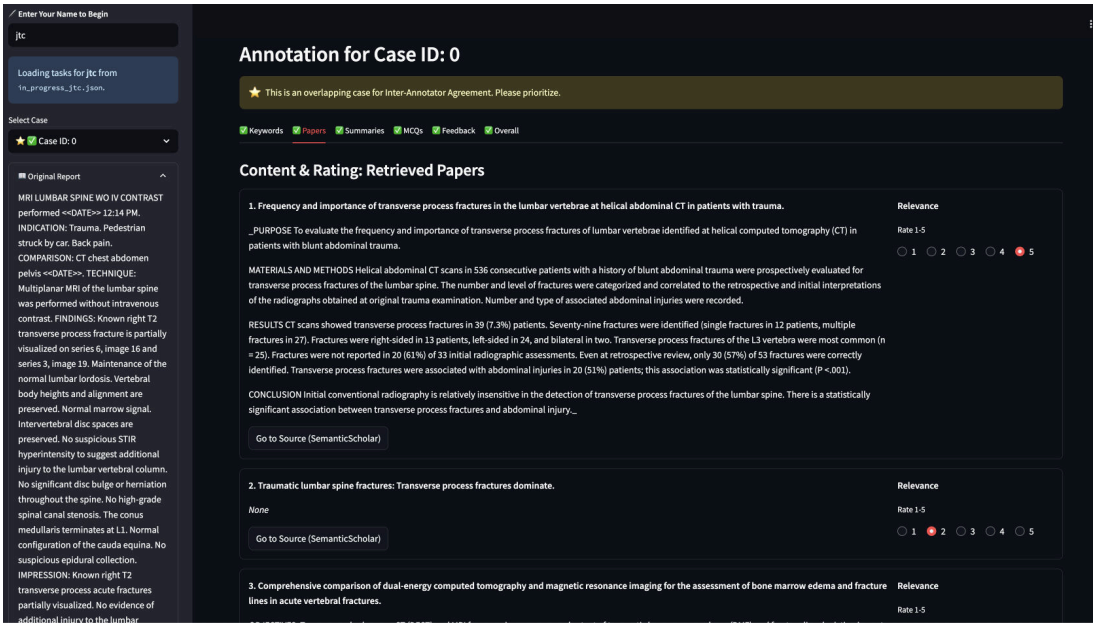
For the entire generated report:

- **Appropriateness of Keywords:** Are the keywords (used to structure the feedback) appropriate and comprehensive for the given original radiology report? Specifically, is the keyword general enough that it can be searched in a textbook or Radiopaedia (e.g., “rib fracture”, not “anterior 4th rib fracture”), and related to a pathology worth learning more about (e.g., “cholangiocarcinoma”, not “mass”)?

G Human Annotator System UI

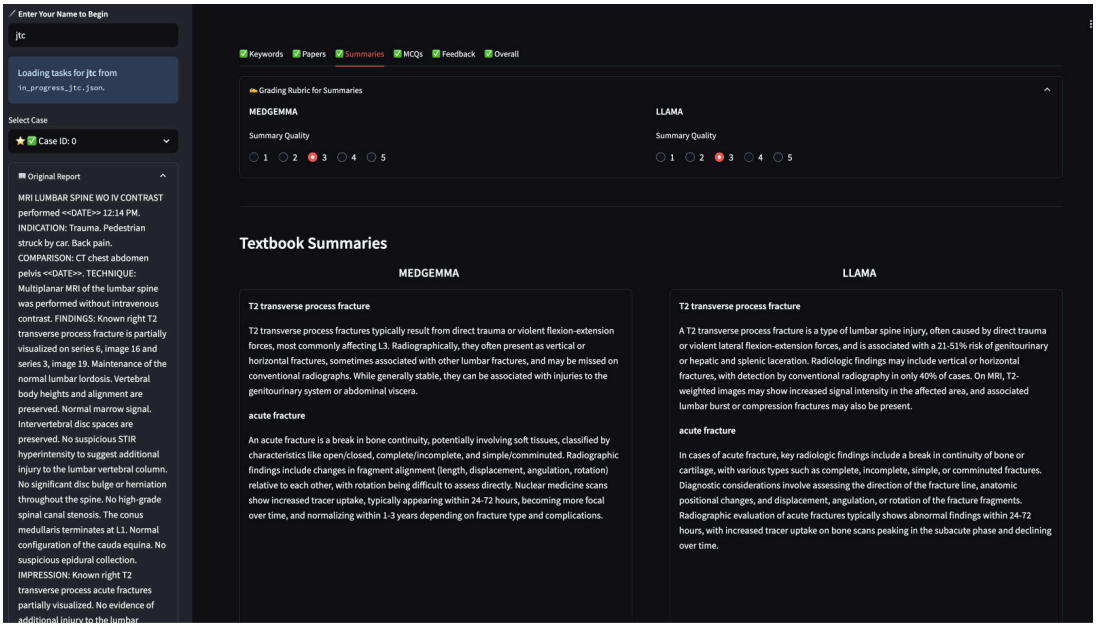


(a) Keyword Evaluation Page

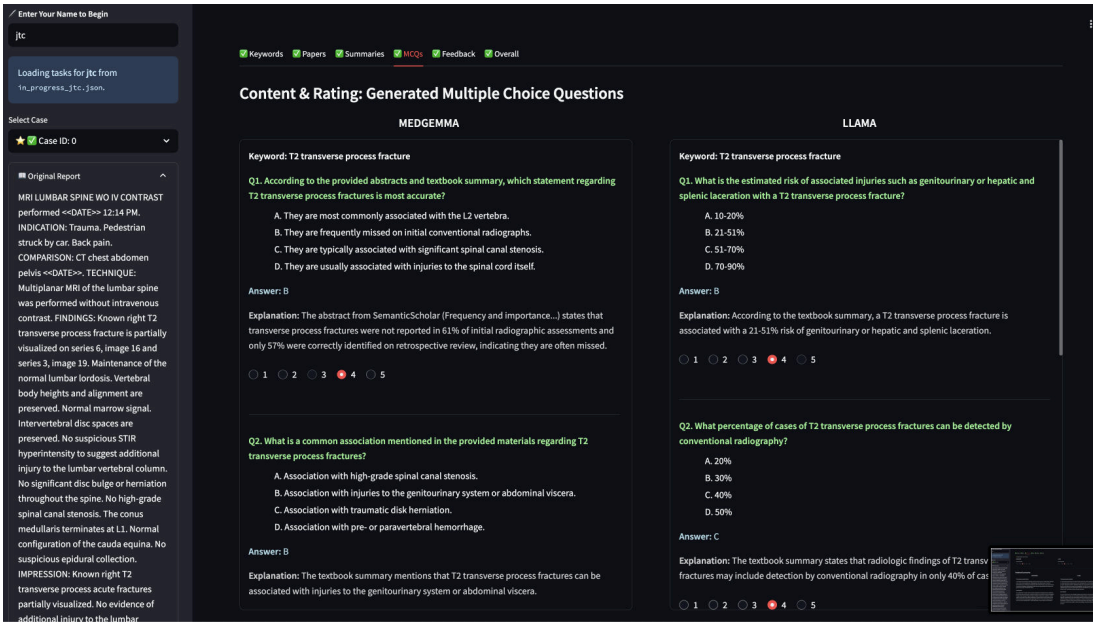


(b) Paper Evaluation Page

Figure 7: Annotation system UI (Part 1 of 3): Interfaces for evaluating keywords and retrieved papers.

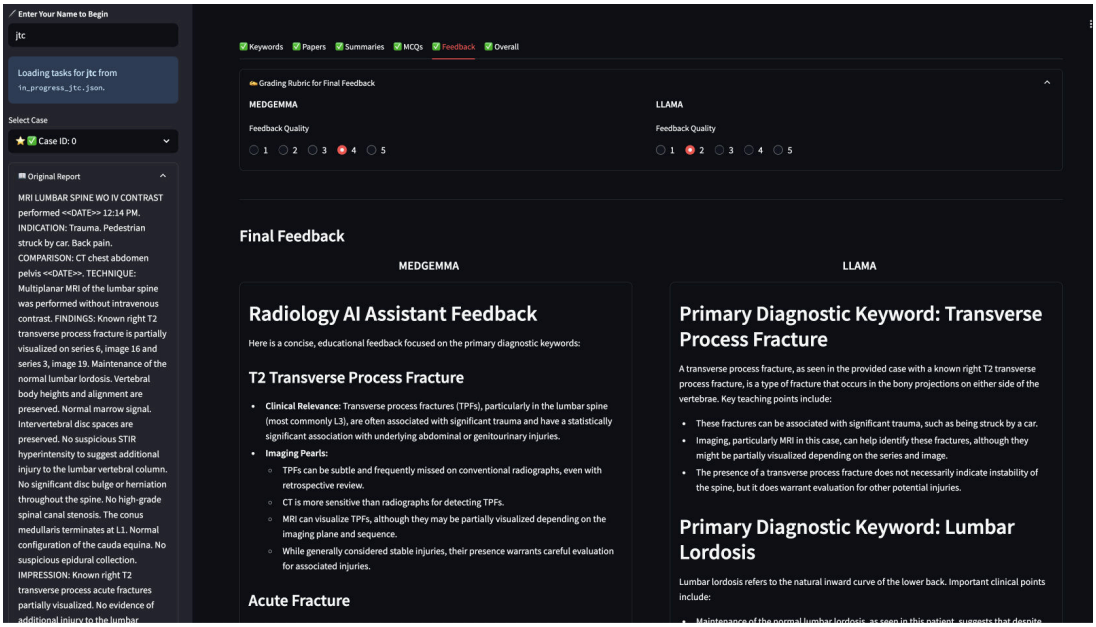


(a) Textbook Summary Evaluation Page

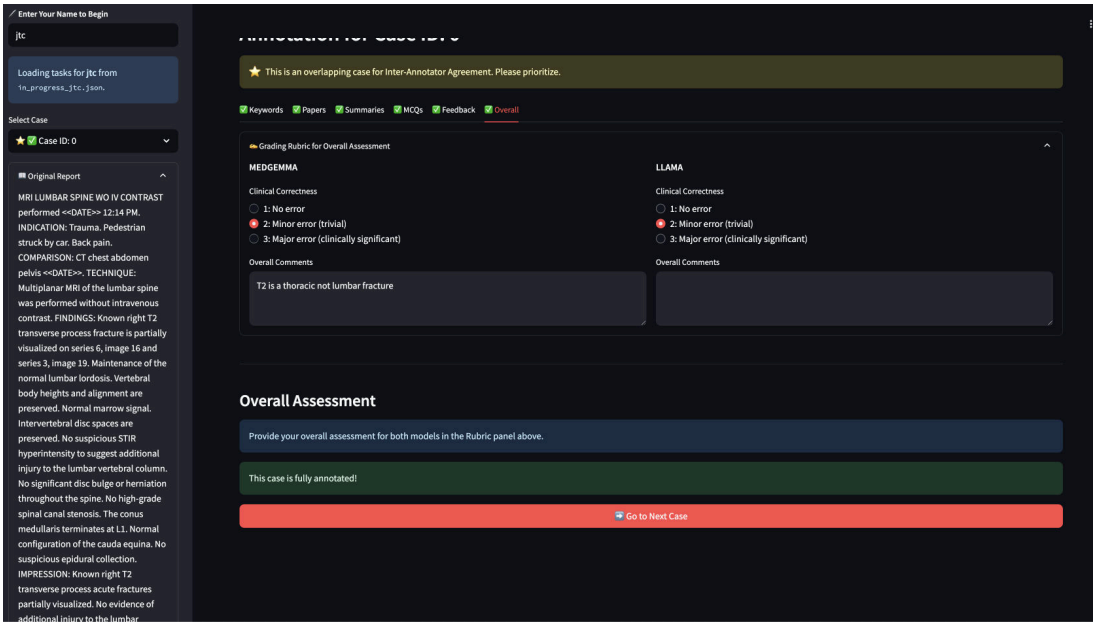


(b) MCQ Evaluation Page

Figure 8: Annotation system UI (Part 2 of 3): Interfaces for evaluating textbook summaries and multiple-choice questions.



(a) Educational Material Evaluation Page



(b) Overall Quality Evaluation Page

Figure 9: Annotation system UI (Part 3 of 3): Interfaces for evaluating the final synthesized educational material and overall quality.