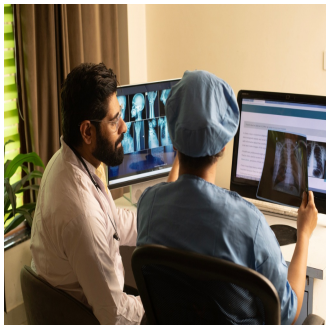
HealthManagement
Promoting Management and Leadership

Fine-Tuned LLMs Improve Radiology Report Accuracy



Radiology plays a critical role in diagnosing thoracic conditions, but its effectiveness depends heavily on the clarity and accuracy of reports. Unfortunately, various sources of error—ranging from speech recognition software limitations to human interpretation biases—pose risks to patient safety and care quality. Traditional solutions, such as structured reporting and deep learning models, offer partial improvements but face limitations in adoption and granularity. In light of this, recent advancements in large language models (LLMs) are now being explored to detect errors in radiology reports, providing a promising path toward automated medical proofreading.

Training LLMs to Detect Radiology Errors

To evaluate how LLMs might enhance error detection, researchers developed a two-part dataset. The first segment consisted of 1,656 synthetic chest radiology reports generated using GPT-4, evenly split between error-free texts and those embedded with four distinct error types: negation, left/right mislabeling, interval changes and transcription mistakes. The second segment included 614 reports derived from the MIMIC-CXR database, matched with synthetic versions containing deliberate errors. This comprehensive dataset enabled the team to rigorously test various LLMs, including GPT-4, BiomedBERT and Llama-3 models, using different prompting and fine-tuning strategies.

Must Read: ***Comparing Open- and Closed-Source LLMs for Radiology Error Detection***

Among the tested models, the fine-tuned Llama-3-70B-Instruct demonstrated superior performance. Under zero-shot prompting, it achieved a macro F1 score of 0.780, with particularly high accuracy in detecting transcription (F1 = 0.828) and left/right errors (F1 = 0.772). These results indicated that, when properly fine-tuned, generative LLMs could effectively identify subtle and diverse errors in radiology text. This performance was further validated when 200 reports were reviewed by radiologists, confirming that the model-detected errors aligned with expert evaluations in most cases.

Impact of Model Configuration and Prompt Design

Model performance varied depending on scale, training and prompting strategy. Fine-tuning emerged as a critical factor, significantly improving results across all error types. When comparing model sizes, the larger Llama-3-70B-Instruct outperformed its 8B counterpart across most tasks, demonstrating greater consistency in detecting complex errors such as interval changes. Conversely, the smaller model showed isolated strengths, particularly in identifying negation errors, suggesting task-specific trade-offs related to model size.

Prompting strategies also influenced performance. Zero-shot prompting, which requires no prior examples, offered robust generalisation across all error types. However, few-shot prompting—especially four-shot random prompts—proved more effective for tasks like transcription error detection, where contextual understanding benefits from illustrative examples. Interestingly, specified prompts did not consistently outperform randomly selected ones, highlighting that example relevance, not merely specificity, determines effectiveness in few-shot scenarios.

Validating in Real-World Contexts

To assess generalisability, researchers applied the models to 55,339 real-world radiology reports from New York-Presbyterian/Weill Cornell Medical Center. A voting mechanism combining predictions from the fine-tuned Llama-3-70B-Instruct and GPT-4 identified 606 reports likely to contain errors. Subsequent human review of 200 randomly selected reports showed that 99 were confirmed by both radiologists and 163 by at least one, resulting in an overall model-confirmed detection rate of 81.5%. This real-world validation demonstrated that the models could maintain high accuracy outside the confines of a synthetic dataset, with transcription and spatial orientation errors being particularly well identified.

Despite these promising results, challenges remain. The fine-tuning process demands significant computational resources and access to high-quality annotated datasets. Furthermore, while synthetic data offers a privacy-preserving training method, it may not fully capture the nuance and variability of real-world reporting styles and errors. Researchers acknowledged this limitation and took steps to ensure diverse, representative synthetic examples were used, with radiologists verifying error accuracy during model training.

The study confirmed that fine-tuned LLMs can meaningfully enhance the detection of errors in radiology reports, positioning them as valuable tools for automated medical proofreading. With Llama-3-70B-Instruct outperforming both domain-specific and general-purpose models, fine-tuning appears essential for deploying these systems in clinical environments. Furthermore, the findings emphasise the importance of thoughtful prompt design and real-world validation to ensure robust, generalisable performance. While computational and data-related barriers must still be addressed, the integration of LLMs into radiology workflows has the potential to reduce diagnostic errors, improve report quality and ultimately support better patient care.

Source: **Radiology**

Image Credit: **iStock**

Published on : Fri, 30 May 2025