HealthManagement
Promoting Management and Leadership

# Enhancing Radiology Reporting with Large Language Models

Radiology reports are essential in clinical practice, providing key diagnostic information from imaging studies. These reports typically comprise two parts: detailed imaging findings and an impression section that delivers a diagnostic summary, including differential diagnoses and recommendations. The interpretation and generation of these impressions require significant expertise, and mistakes or inconsistencies can occur. To address these challenges, recent advancements in artificial intelligence, particularly large language models (LLMs), are being explored as tools to generate radiologic impressions automatically. A recent article in Radiology examines the development, evaluation, and potential implications of constructing an LLM specifically for radiology report generation.

### Challenges in Radiologic Impression Generation

Due to the specialised and nuanced nature of radiological data, generating radiologic impressions is a complex task. Radiologists need to integrate multiple elements—such as patient history, clinical symptoms, and imaging findings—to deliver a comprehensive and accurate diagnostic impression. Variability among radiologists in interpreting these findings, coupled with cognitive biases, can result in differing conclusions, even with similar imaging data.

One of the central challenges in automating this process is ensuring that the generated impressions are clinically accurate and linguistically appropriate. Early attempts to use general-purpose LLMs, such as GPT-2 and GPT-3, showed that while the language models could match radiologists in grammar and readability, their diagnostic accuracy was lower. This highlighted the need for domain-specific training to create models that could handle the unique demands of radiology. Consequently, a focused effort was made to construct LLMs that are tailored specifically for radiological data, fine-tuned to recognise the complexities and subtleties involved in radiology reporting.

### Developing a Radiology-Specific LLM

In response to these challenges, researchers developed a specialised LLM using a vast dataset of medical texts, radiology reports, and other relevant clinical information. The model, known as WiNGPT-7B, was trained on 20 GB of both medical and general-purpose text, and fine-tuned with 1.5 GB of data that included 800 radiology reports paired with instructions. This extensive pretraining and fine-tuning process was designed to equip the model with the ability to generate impressions from various imaging modalities, such as CT, MRI, and radiography, across multiple anatomical sites.

The training process involved several critical steps. First, data from various sources were converted into textual format and rigorously cleaned to remove noise, such as irrelevant links or advertisements. The model's backbone, Llama 2 7B, was used as a foundation, providing robust natural language understanding capabilities. Instruction learning techniques, such as Self-Instruct and Evol-Instruct, were employed to refine further the model's ability to follow complex prompts and generate contextually appropriate outputs.

After training, the model was evaluated using a large dataset of real-world radiology reports. The results demonstrated that the LLM could generate impressions that were both linguistically and clinically consistent with the final impressions written by radiologists. However, the study also highlighted certain limitations, particularly in specific diagnostic accuracy, which underscores the need for continuous refinement of such models.

### Evaluating the Performance of LLM-Generated Impressions

The evaluation of the LLM's performance was a key component of this research. The model's generated impressions were compared against radiologist-written impressions using metrics such as recall, precision, and F1 scores. In a large test set of nearly 4,000 patients, the LLM

achieved a median recall of 0.775, a precision of 0.84, and an F1 score of 0.772. These results indicate a high level of agreement between the LLM and human experts, particularly in terms of comprehensiveness and factual consistency.

In addition to these quantitative metrics, an expert panel conducted a qualitative assessment of the LLM's performance, evaluating impressions based on scientific terminology, coherence, specific diagnosis, differential diagnosis, management recommendations, correctness, comprehensiveness, and lack of bias. The LLM consistently scored well in most areas, especially in coherence and scientific terminology. However, its performance was somewhat less favourable in providing specific diagnoses, reflecting the inherent difficulty of replicating human diagnostic reasoning without access to a broader clinical context.

Despite these limitations, the overall evaluation showed that the LLM-generated impressions were clinically useful and aligned closely with the final impressions crafted by radiologists. The expert panel's feedback confirmed that the LLM could be a valuable tool for supporting radiologists in generating high-quality, standardised reports, especially in high-volume settings where time constraints may lead to variability in reporting quality.

**Conclusion**

The development of LLMs tailored for radiology marks a significant advancement in medical AI. They offer the potential to streamline the radiology reporting process and enhance the consistency and accuracy of diagnostic impressions. While challenges remain, particularly in improving the model's ability to generate specific diagnoses, current research demonstrates the feasibility of using LLMs in clinical practice.

By integrating LLMs into the radiology workflow, hospitals and clinics could potentially reduce the cognitive load on radiologists, allowing them to focus on more complex cases and patient interactions. Furthermore, as these models continue to evolve, their ability to assist in other areas of radiology, such as image analysis and clinical decision-making, may also expand. Future research should focus on refining these models, addressing their current limitations, and exploring their broader applications in clinical medicine.

**Source: RSNA Radiology**

**Image Credit: iStock**

Published on : Mon, 23 Sep 2024