Development, Evaluation, and Assessment of Large Language Models (DEAL) Checklist: Reporting Methods and Results for LLM-Based Radiology Research

Satvik Tripathi, Researcher, Department of Radiology , University of Pennsylvania
Dana Alkhulaifat, MD; Florence Doo, MD; Pranav Rajpurkar, PhD; Rafe Mcbeth, PhD; Dania Daye, MD, PhD; Tessa Cook, MD, PhD, FSIIM

Background/Problem Being Solved

Large language models (LLMs) are increasingly employed in radiology for automated reporting, workflow enhancement, and decision support tasks. Despite their transformative potential, radiology research involving LLMs often suffers from inconsistent methodology reporting, limiting reproducibility, generalizability, and clinical integration.

Intervention(s)

We developed the Development, Evaluation, and Assessment of Large Language Models (DEAL) Checklist to address these challenges. This standardized reporting framework is designed specifically for LLM applications and builds on existing guidelines, such as the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) and EQUATOR network. It provides a comprehensive structure for documenting methodologies, evaluation metrics, and ethical considerations in radiology-focused LLM research.

Barriers/Challenges

Radiology presents unique challenges in the development and application of LLMs. These include dataset biases arising from differences in radiology reports across various imaging modalities, subspecialties, and institutions, as well as the need for models to be compatible with combined textual and imaging information. Additionally, the inherent stochastic variability in LLM outputs complicates the reliability and clinical applicability of results. The absence of standardized protocols for fine-tuning LLMs and optimizing prompt engineering furthers the disparities in model performance, complicating their deployment across diverse clinical settings.

Outcome

The DEAL Checklist offers two reporting pathways: DEAL-A, for studies focusing on the development and fine-tuning of LLMs tailored to radiology tasks, and DEAL-B, for studies utilizing proprietary or pre-trained models. Both pathways emphasize the reporting of model specifications, dataset preparation, evaluation metrics, and ethical considerations, ensuring transparency and reproducibility.

Conclusion/Statement of Impact/Lessons Learned

The DEAL Checklist is a crucial step in advancing rigorous and reproducible LLM-based radiology research. Improving standardization facilitates the reliable integration of LLMs into clinical practice, ultimately enhancing diagnostic accuracy, workflow efficiency, and patient care outcomes.

Figure(s)

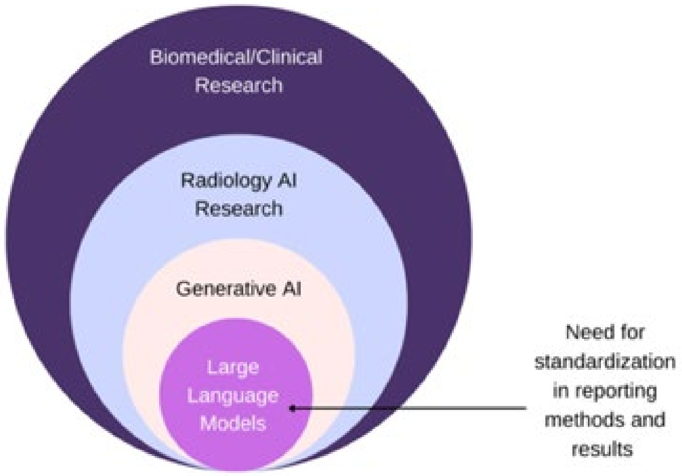
Biomedical/Clinical Research
Radiology AI Research
Generative AI
Large Language Models
Need for standardization in reporting methods and results

Figure 1. Venn diagram representing the hierarchy of research areas. Biomedical and clinical research encompasses radiology AI research, which includes generative AI and specifically large language models (LLMs). There exists a need of standardization in reporting methods and results, particularly for LLM research.

Keywords

Applications; Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity; Emerging Technologies; Imaging Research