SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Synthesizing Diagnostic Insights from Radiology Reports: A RAG-Based LLM Method for Reducing Hallucinations and Preventing Catastrophic Forgetting

Briana Malik, University of Pittsburgh

Introduction/Background

Disparities in data quality and context availability can introduce biases in Large Language models (LLMs), affecting their accuracy. These issues are particularly pronounced in the field of radiology, where precise interpretation and understanding of reports is critical. Incorporating accurate and contextually relevant information is essential to LLM performance, reducing hallucinations and catastrophic forgetting.

Methods/Intervention

We hypothesize that integrating Retrieval-Augmented Generation (RAG) with contextual search will significantly reduce hallucinations by grounding LLMs with accurate and contextually relevant information. We used 500 radiology reports from a chest X-ray collection, tokenized the text, and generated embeddings using Large Language Model (LLM) tokenizer. These embeddings were stored in LevelDB database for efficient storage and retrieval. A similarity search index was built to facilitate efficient contextual retrieval. Queries related to specific radiological conditions were processed through RAG system. RAG system retrieved the most relevant context, which was then combined with the query and input into LLM (GPT-2) to generate contextually rich responses.

Results/Outcome

The RAG-based method significantly improved LLM's understanding of radiology reports and rare conditions by grounding the model and reducing hallucinations. The number of words in responses decreased from 388 to 223, showing more concise outputs. Unique words decreased from 40 to 109 with RAG, indicating less repetition. The repetition rate fell from 0.897 to 0.511. ROUGE-1 F1 score improved from 0.015 to 0.024, and ROUGE-1 precision increased from 0.007 to 0.013. ROUGE-1 recall remained constant at 0.500. Perplexity increased from 1.605 to 10.369, reflecting more contextually rich responses.

Conclusion

The RAG-based approach enhances the accuracy and relevance of responses and improves understanding of rare conditions. Reduced repetition rates and improved ROUGE scores demonstrate more accurate responses. Higher perplexity with RAG indicates richer, more contextual responses compared to lower perplexity and incoherence without RAG.

Statement of Impact

This study demonstrates the potential of RAG-based LLMs to advance radiology report interpretation and rare disease diagnosis. By providing more accurate, contextually relevant answers, this approach enhances diagnostic quality and patient care, addressing critical gaps in traditional LLM training which is not domain specific, suffers from under/not training if very rare words do not appear in vocab generated during frontier model training.

See figures in this link: https://www.abstractscorecard.com/uploads/Tasks/upload/23112/CRAJXJCV-1889033-1-ANY(1).pdf
Heatmap with and without RAG system, ROGUE Score, Perplexity Score, Unique Words, Precision and Recall, Cosine Similarity.

Keywords

Retrieval Augmented Generation (RAG)**;** Large Language Models (LLM); Deep Learning; Artificial Intelligence; Transformer based Models**;** Hallucination