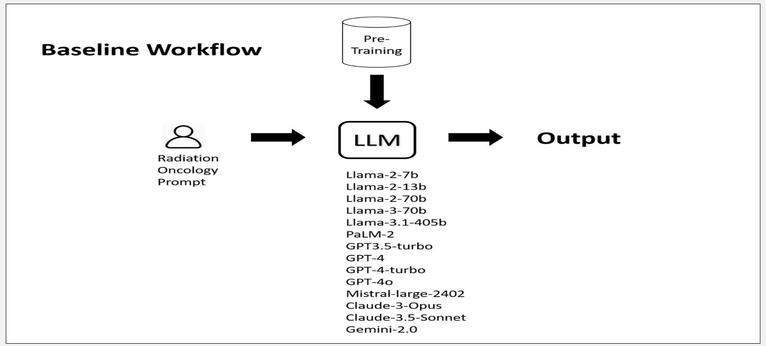
Quantifying Domain-Specific Knowledge of Large Language Models: Analysis Using the American College of Radiology's Radiation Oncology In-Training Examination

Nikhil G. Thaker, Navid Redjal, Adam Dicker, Arturo Loaiza-Bonilla, Vivek Subbiah, Aaron Law, Trevor Royce, Vikash Deendyal, Ajay Choudhri, Gautam H. Thaker

Capital Health Hopewell Medical Center, Pennington, NJ

Introduction

Large Language Models (LLMs) show promise in assisting knowledge-intensive fields such as oncology, where up-to-date information and multidisciplinary expertise are critical. Traditional LLMs risk hallucinations and reliance on static, possibly outdated data. Retrieval-Augmented Generation (RAG) has emerged as a strategy to address these issues by incorporating real-time, domain-specific information from external knowledge repositories.



Methods

We developed an end-to-end RAG pipeline ("Iridium Model") that embeds user prompts via the OpenAI text-embedding-3-small model and matches them against a specialized radiation oncology database (textbooks, guidelines, Q&A archives). The top five semantically relevant chunks are retrieved from a Chroma vector database and merged with the user prompt to create an augmented query. This augmented query is fed into 14 pre-trained LLMs (i.e. GPT 3.5/4) We benchmarked performance using the 2021 American College of Radiology (ACR) in-training exam in radiation oncology.

Results

Larger, more recent LLMs had higher baseline scores, with some exceeding scores of post-graduate year (PGY) 5 residents. Introducing RAG improved performance for all top models, with gains of 2–7 percentage points. The highest-performing model (RAG-GPT-4o) achieved 85.6% accuracy, surpassing graduating resident averages by nearly 15%. Domain-specific analysis in clinical, statistics, biology, and physics showed uniform or near-uniform improvements under RAG, with exceptions in certain models. Majority voting further boosted aggregate accuracy when individual model correctness exceeded 0.5.

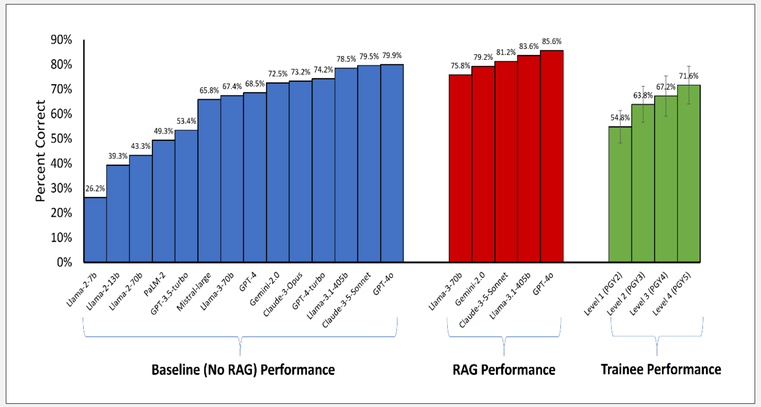


Figure 1 .Comparison of model performances between baseline (no RAG) zero-shot workflow (blue columns), RAG workflow (red columns), and trainee performance (Level 1 through Level 4, corresponding to PGY2-5) (green columns).

capitalhealth
Minds Advancing Medicine

Conclusion

Our findings demonstrate that a RAG framework significantly improves LLM accuracy in radiation oncology, mitigating hallucinations and enhancing domain specificity. By integrating real-time, externally sourced knowledge, RAG can support clinical education, decision-making, and patient education in oncology. Future work will expand the domain-specific databases, refine retrieval methods, and explore multi-modal data integration to further optimize LLM performance in complex medical fields.

	Baseline	RAG	Improvement
Gemini-2.0-Flash-Exp	73%	79%	7%
Llama-3.1-405b-instruct	79%	84%	5%
Claude-3.5-Sonnet	80%	81%	2%
GPT-4o	80%	86%	6%

Table 1. Improvement in the top performing LLMs. Largest gains were seen for Gemini-2.0 and GPT-4o, with GPT-4o having the best benchmark performance of all tested models. RAG, Retrieval Augmented Generation

Affiliations

- Capital Health, Pennington, New Jersey
- Bayta Systems, Newtown, Pennsylvania
- Thomas Jefferson, Philadelphia, PA
- St Luke's University Health Network
- Sarah Cannon Research Institute, Nashville, TN
- Atrium Health Wake Forest Baptist, Winston Salem, NC
- Lockheed Martin, Pennsylvania