

OPTIMIZING FGR MANAGEMENT THROUGH ARTIFICIAL INTELLIGENCE APPLICATION: A COMPARATIVE ANALYSIS OF RETRIEVAL-AUGMENTED GENERATION (RAG) SYSTEMS VS. STANDARD LARGE LANGUAGE MODELS

AUTHORS —————

Franciszek Ługowski, Julia Babińska, Paweł Jan Stanirowski

AFFILIATION —————

1st Department of Obstetrics and Gynecology, Medical University of Warsaw

INTRODUCTION

Large Language Models (LLMs) offer potential for clinical decision support but are prone to "hallucinations" and the conflation of differing international guidelines. In the management of Fetal Growth Restriction (FGR), discrepancies in Doppler-based staging and delivery timing are of particular importance.

OBJECTIVE

This study aims to develop a Retrieval-Augmented Generation (RAG) system grounded in specific fetal medicine guidelines and to evaluate its accuracy against a standard non-augmented LLM in the management of FGR.

METHODOLOGY

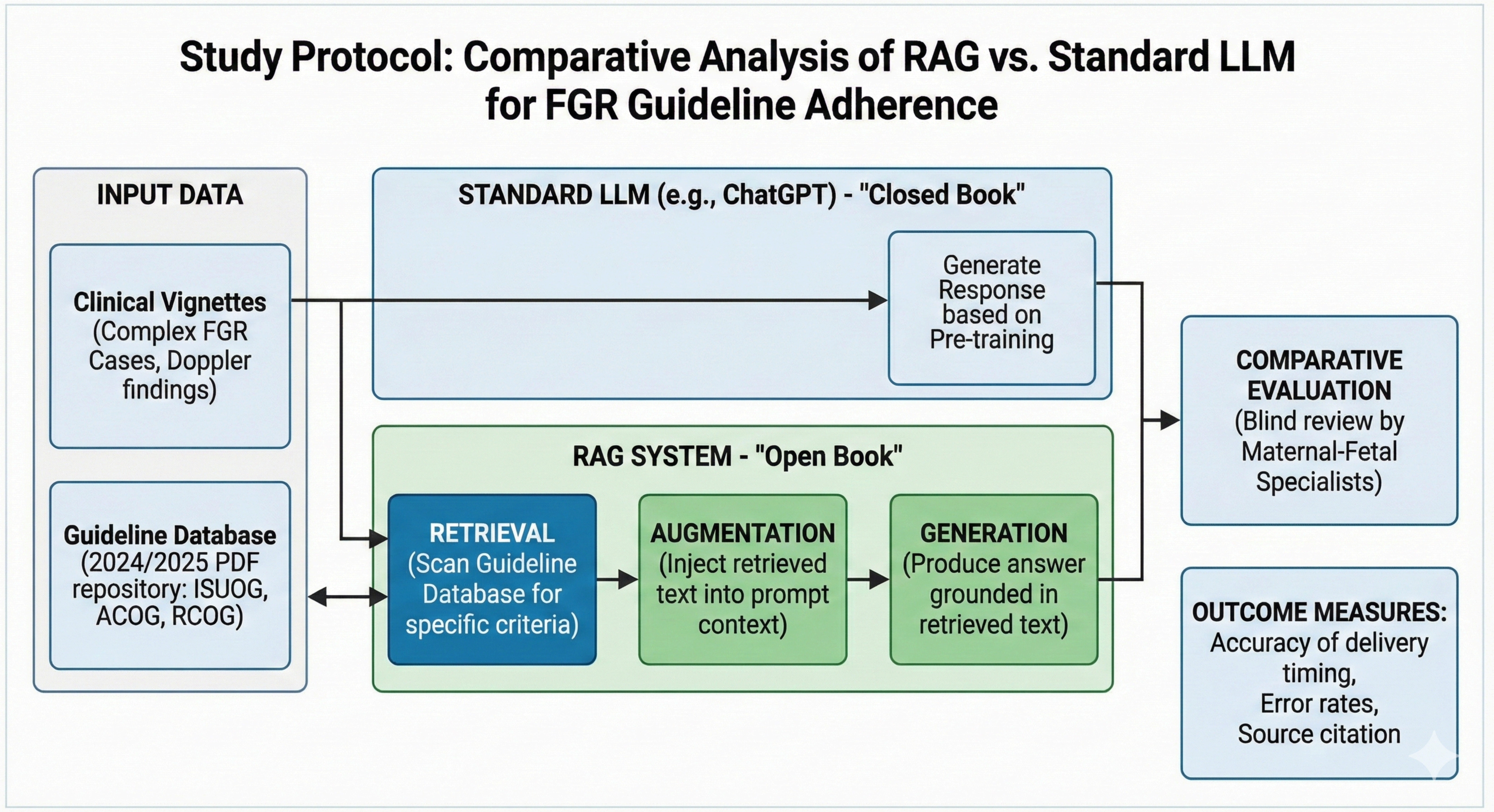
We engineered a domain-specific RAG system ("The Open Book" model) specifically indexing the latest 2024/2025 practice guidelines from ISUOG, RCOG, and ACOG. We curated a dataset of 50 complex clinical vignettes featuring varying FGR stages and Doppler findings. Each vignette was queried against two models: (1) a standard commercial LLM (GPT-5.1) relying solely on pre-training, and (2) the RAG-enhanced model. Expert professionals in maternal-fetal medicine conducted a blinded assessment of the responses. The primary outcome was the accuracy of the recommended delivery timing. Secondary outcomes included proper guideline citation and correct interpretation of Doppler staging.

RESULTS

The standard LLM demonstrated a baseline accuracy of 62% for delivery timing, frequently recommending interventions based on outdated criteria or mixing society guidelines. In contrast, the RAG-enhanced system achieved an accuracy of 96% (p<0.001). Critical errors (defined as delivery recommendations deviating by >1 week from guidelines) occurred in 18% of standard LLM responses compared to 0% in the RAG cohort. Furthermore, the RAG system successfully provided source verification (page/paragraph citations) for every management decision.

CONCLUSION

Standard LLMs pose significant risks in FGR management due to the potential for stochastic hallucinations of clinical protocols. RAG architecture significantly outperforms standard models by strictly grounding responses in validated literature, suggesting its implementation for the secure integration of Artificial Intelligence within FGR management protocols.



Fundusze Europejskie
dla Rozwoju Społecznego



Rzeczpospolita
Polska

Dofinansowane przez
Unię Europejską

