Transforming Healthcare Education: Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation

Yasmina Al Ghadban[1,2], Huiqi Yvonne Lu[1,3] , Uday Adavi[1], Ankita Sharma[1,2], Sridevi Gara[1], Neelanjana Das[1], Bhaskar Kumar[1], Renu John[1], Praveen Devarsetty[1] , Jane E Hirst[1,2,4]

1. *The George Institute for Global Health*
2. *Nuffield Department of Women's & Reproductive Health, University of Oxford, UK*
3. *Computational Health Lab, Department of Engineering Sciences, University of Oxford, UK*
4. *Imperial College London, UK*

Background

- There is a need to enhance medical education for frontline health workers, particularly in resource-constrained environments
- **SMART*health* Pregnancy** is a digitally supported tool for frontline health workers (ASHAs) to identify, screen and refer high risk pregnant women
- ASHAs and pregnant women asked us for support of their continuous learning and for a way to provide real-time answers to questions women want to know about their pregnancy

Aim

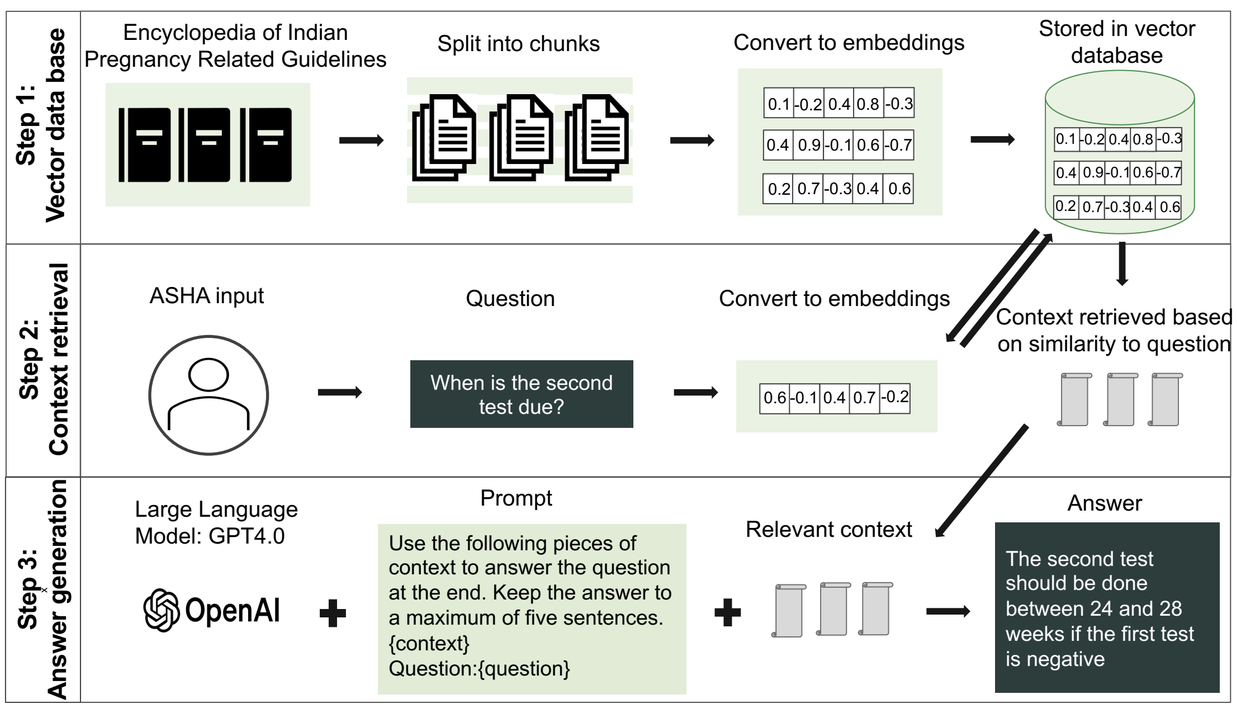
To develop, technically and clinically validate, an LLM suitable for community health workers in rural India to improve healthcare education and support guideline-based pregnancy care

Methods

Retrieval-augmented generation (RAG) pipeline

RAG enables LLMs to access information from non-parametric storage (**Figure 1**). It was chosen for traceability to source material, scalability across vast knowledge bases, and seamless adaptability to evolving clinical guidelines.

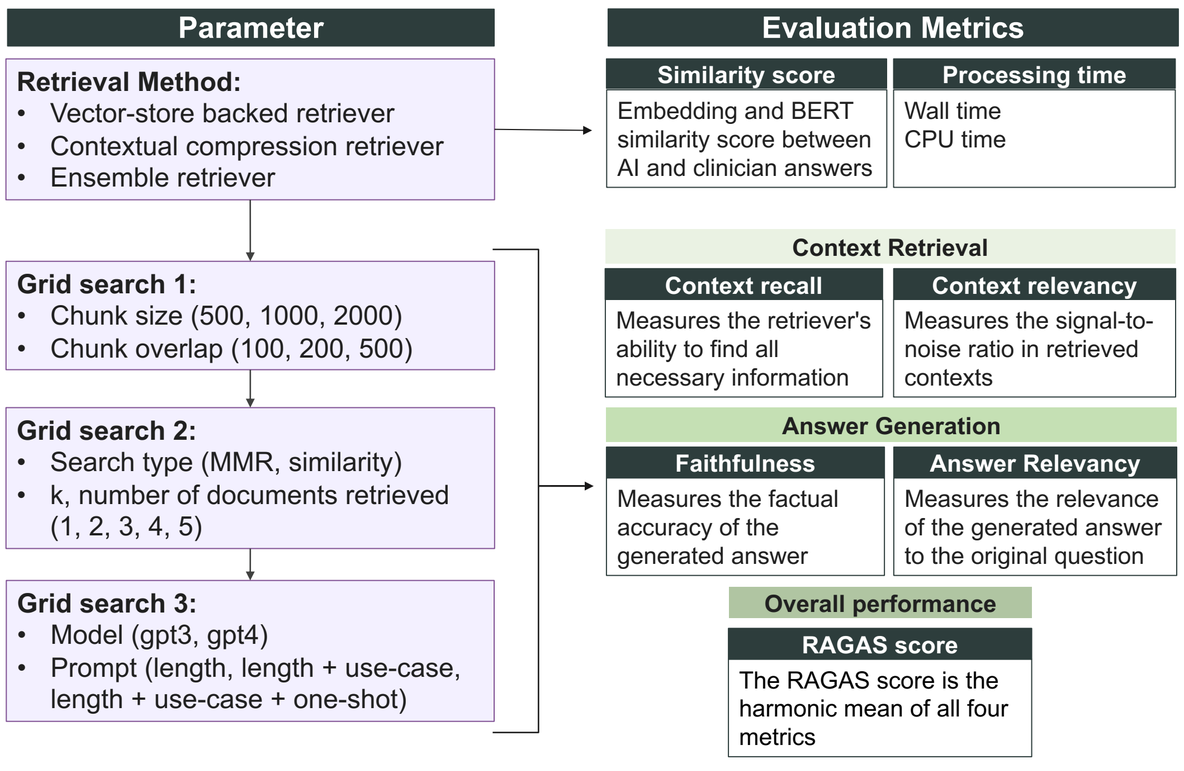
Figure 1. Flow diagram of the RAG process in SMART*health* GPT



Parameter optimisation

To select the model parameters, we developed an evaluation pipeline using similarity scores, processing time and the RAGAS framework, which allows evaluation of both generation and retrieval steps alone, and an assessment of overall performance (**Figure 2**).

Figure 2. Parameter optimisation and evaluation pipeline



Clinician Evaluation

Twelve community medicine clinicians and two obstetricians, across two states, rated AI-generated answers based on accuracy, completeness, appropriateness, and presence of bias on a 3-point Likert scale. The 180 questions included in this round of clinical validation were developed with ASHAs directly through focus groups and community engagement.

Results

Step 1: Development of the encyclopaedia

The final repository included 20 pregnancy guidelines, with a focus on anaemia, gestational diabetes, and hypertension in pregnancy.

Step 2: Context Retrieval

The answer quality, measured with similarity scores, was similar between the simplest vector-store based retriever, the compression retriever and the ensemble retriever, while the processing time increased. The chunk size, chunk overlap, search type and number of retrieved documents (k) parameters minimally impacted model performance. As a result, model parameters were primarily chosen based on processing time and technical considerations.

Step 3: Answer generation

The model choice and prompt template also minimally impacted RAGAS metrics.

Table 1. Selected parameters for SMART*health* GPT v1

Parameter	Value	Rationale
Retriever	Vector-store backed retriever	Processing time
Chunk size	1000 characters	Processing time
Chunk overlap	200 characters	Processing time
Search type	MMR	Diversity in retrieval minimises repetition
k	2 chunks	Balance in completeness and duplication
Model	gpt-4	Lower likelihood of hallucinations
Prompt	One-shot prompt template	More predictable responses

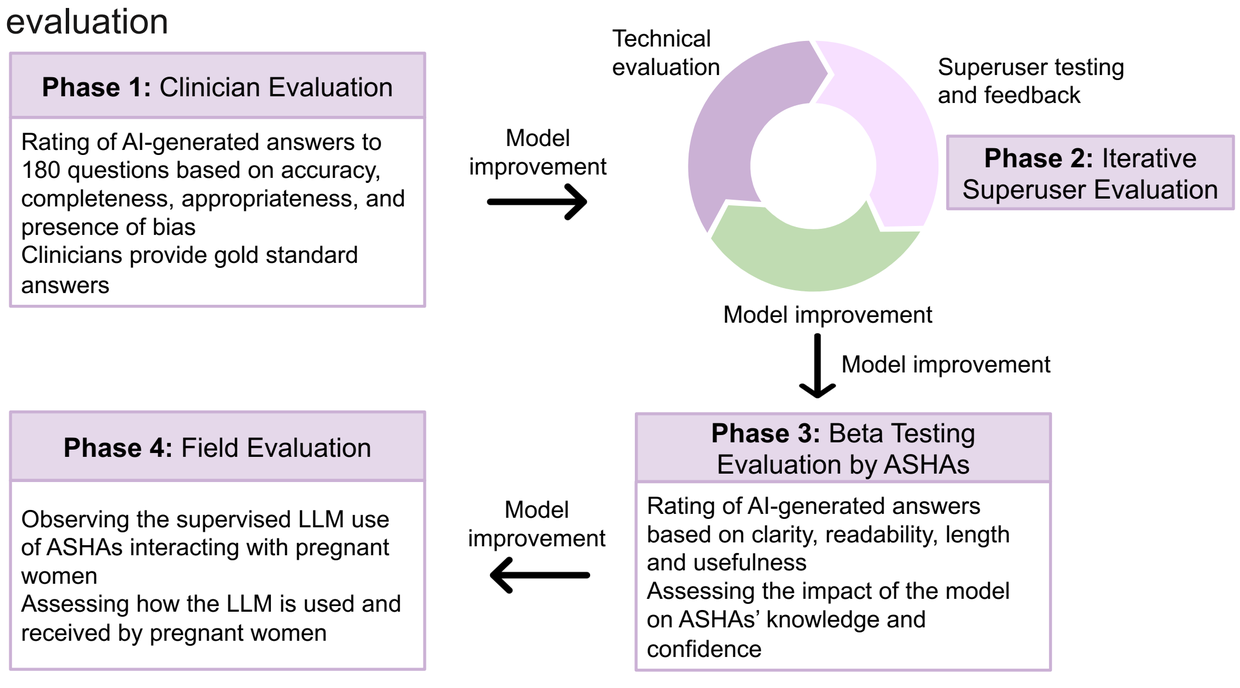
Clinician evaluation

Each question was rated by at least two clinicians. For 141 (79%) questions, all clinicians agreed that the AI-generated answer was completely or partially accurate; and that the AI-generated answer was either unbiased or actively promoted equity. However, all clinicians rated the completeness of the AI-generated answer as adequate or comprehensive for only 49 (35%) questions.

Discussion

The clinical validation has allowed to gain significant insight into the performance of the model, and importantly, to identify the cases where the model is failing (the answers with low ratings). In addition to rating the model's answer to the 180 questions, clinicians also provided an ideal response that is clear, uncontroversial, and appropriate for ASHAs. The standardised answers will constitute the "gold-standard" responses and the repository of 180 question-answer pairs can be used to fine-tune the model and further improve its prompts. The LLM will be improved iteratively based on the feedback from each phase of validation (**Figure 3**).

Figure 3. Iterative LLM improvement through the four phases of clinical evaluation



Conclusion

SMART*health* GPT showcases the promising role of RAG and LLMs in medical education and provides insights for future applications of generative AI in diverse educational settings.