A Layered Approach to Reducing Hallucinations in LLMs for Clinical Data

MSR4

Ashwin Rai,[1] Devika Bhandary,[2] Victoria Ikoro,[3] Andre Ng[3]

[1]Thermo Fisher Scientific, Waltham, MA, USA; [2]Thermo Fisher Scientific, London, UK; [3]Thermo Fisher Scientific, Ontario, Canada

Introduction

- Large language models (LLMs) are being piloted to automate literature review, summarize electronic health records (EHRs), and extract clinical insights from structured and unstructured data.[1,2]
- Their promise is speed and scale, but reliability is critical when outputs inform care or policy.
- Hallucinations, defined as fabricated or misleading content, remain a key risk without careful prompt design and operational safeguards.[3]
- These risks are heightened in healthcare, where inaccurate or nontransparent outputs may compromise trust and decision-making.[2,3]

Problem statement

- Naive prompting (single "stuff" chain): With a generic "answer based on context" template and no retrieval guardrails, the LLM produced speculative outputs and failed to tie claims to evidence, such as vague diabetes summaries and unsupported hemoglobin A1C (HbA1c) generalizations when specific data were required.
- This behavior illustrates hallucination risk and misuse of context in clinical questions and answers (QA), underscoring the need for structured controls.[3]

Proposed remedy (layered RAG and agent)

- **Layer 1 – Prompting & Validation:** Domain-specific, context-only prompts, which enforce factuality and signal when evidence is absent.[4]
- **Layer 2 – Orchestration (LangChain):** Modular chains with prompt templates, retrieval-augmented generation (RAG),[1-3,5] and lightweight agentic control[3] to separate instructions from logic, route to appropriate models, and ensure only retrieved evidence is used.
  - This layered design enhances reproducibility, transparency, and scalability, while significantly reducing hallucinations and improving clinical relevance.

Objectives

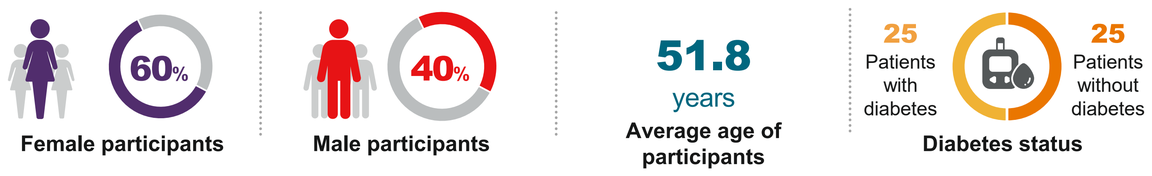
- **The main objective of this study is to reduce hallucinations in LLMs when applied to structured (e.g., tabular EHR data) and unstructured (e.g., clinical text) sources.**
- **To design a two-layer framework where (1) refined prompts enforce factual, context-specific outputs and (2) LangChain operationalizes RAG with agentic control.**
- **To demonstrate, through case examples, how prompt validation and retrieval-aware orchestration improve factual grounding and transparency in clinical and research workflows.**
- **To provide a scalable, interpretable, and compliant approach for deploying LLMs in healthcare analytics and decision support.**

Methods

Data sources

- **Unstructured clinical input:** Article from Tap.Health on "Obtaining a Sample History from a Patient with Diabetes," providing narrative expert insights for testing context-grounded summarization.
- **Structured data set:** Synthetic EHR-like, comma-separated values including patient ID, age, gender, body mass index (BMI), diabetes status, fasting glucose, HbA1c, and cholesterol.
  - Designed to evaluate LLM performance on quantitative tasks (e.g., comparing average HbA1c across cohorts, identifying high-risk demographic segments).

Figure 1. Descriptive statistics of the EHR-diabetic data set



Proposed methodology framework

Layer 1 — Clinical prompt alignment

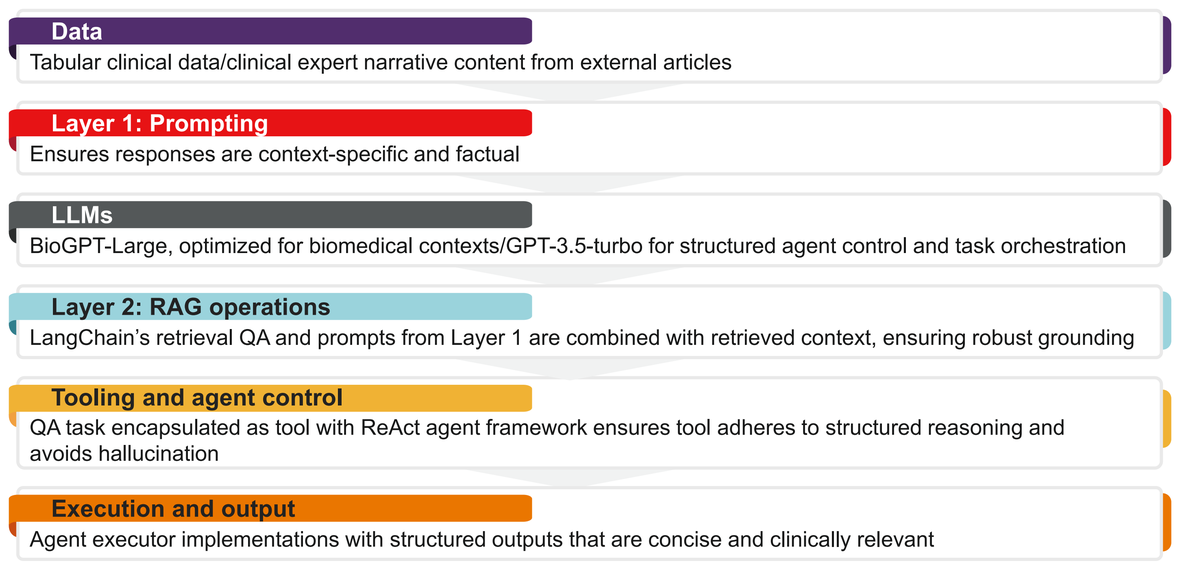
- Prompts strictly enforce factual, context-based answers ("Use ONLY retrieved context..."; "Insufficient evidence" if context absent). Templates (via ChatPromptTemplate) separate instructions from logic, enabling flexible updates.
  - **Domain-specific model:** BioGPT-Large (clinical grounding)
  - **Control model:** GPT-3.5-turbo-0125 (orchestration)

Layer 2 — LangChain RAG and agent control

- **RAG chain:** Combines retriever[6] and Layer 1 prompts into a RetrievalQA chain
- **Tooling:** QA chain wrapped as LangChain Tool (ehr_qa), explicitly handles missing context scenarios
- **Agent Controller:** ReAct agent[7] prioritizes retrieved context, structured tool usage, and controlled actions
- **Execution:** AgentExecutor[8] ensures robust execution and traceability
- **Workflow:** Agent retrieves context, invokes QA prompts, synthesizes summaries using BioGPT-Large, and strictly provides answers grounded in retrieved data

Methods (cont.)

Figure 2. Steps for the proposed methodology framework



Abbreviations: LLM = large language model; QA = questions and answers; RAG = retrieval-augmented generation

Results

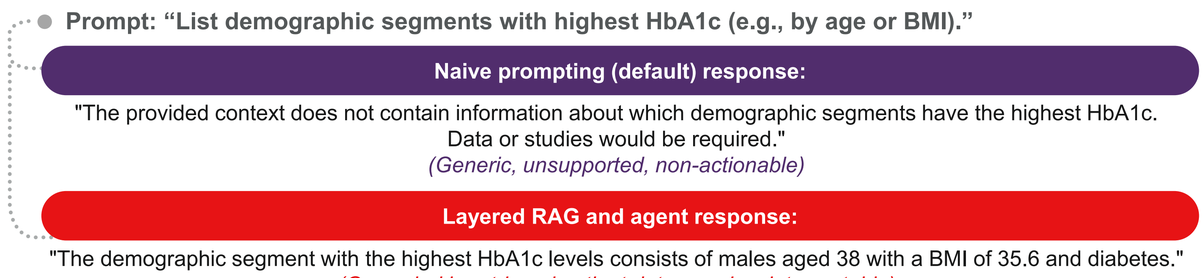
- The framework was tested across different input sources, including structured EHR-like data and unstructured clinical text.
- As shown in **Table 1**, naive prompting produced generalized or speculative outputs, whereas the layered RAG and agent approach generated precise, evidence-grounded responses.

Table 1. Illustration of prompt and response showing naive prompting (generic) versus layered RAG and agent (grounded)

Aspect	Naive Prompting (Baseline)	Layered RAG and Agent Framework
Transparency and accuracy	Generic, speculative, low accuracy (e.g., "Diabetics often have high HbA1c")	Grounded, specific, high accuracy (e.g., "HbA1c = 8.1% (diabetic) vs. 5.6% (non-diabetic)")
Evidence use	No link to retrieved data; outputs not tied to context	Strictly based on retrieved context; answers reference-extracted rows
Clinical reliability	Risk of misinformation; low trustworthiness; unsupported summaries	Traceable, reliable, aligned with domain requirements; auditable outputs
Reproducibility	Variable outputs; little consistency; different answers to same prompt	Consistent workflows through modular orchestration; structured chain ensures repeatability
Context handling	Weak: often ignores or misuses data; skips available demographics	Strong: explicitly tied to structured and unstructured inputs; finds highest HbA1c segment
Output interpretability	Little clarity on reasoning or evidence path; no way to verify sources	Transparent chain of reasoning; source-linked answers; retrieval and computation shown
Practical utility	Limited for healthcare decision support; not usable for risk stratification	Suitable for analytics, evidence synthesis, and policy use; supports clinical alignment

Abbreviations: HbA1c =hemoglobin A1c; RAG = retrieval-augmented generation

Figure 3. Illustration of prompt and response showing naive prompting (generic) versus layered RAG and agent (grounded)



Abbreviations: BMI = body mass index; HbA1c =hemoglobin A1c; RAG = retrieval-augmented generation

Conclusions

- **A layered RAG approach substantially reduced hallucinations and improves factual accuracy compared with naive prompting.**
- **LangChain's modular orchestration framework enables scalability, flexibility, and transparency, supporting reliable deployment in diverse healthcare contexts.**
- **Together, these methods produce transparent, reproducible, and clinically aligned pipelines that strengthen trust in LLM-assisted healthcare analytics and decision support.**
- **This approach provides a scalable pathway for adoption, balancing performance with compliance, and paving the way for safer, evidence-based use of LLMs in real-world practice.**

References

1. Lewis PS, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. 2020; **2.** Asai A, et al. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. 2024; **3.** Yao S, et al. ReAct: Synergizing reasoning and acting in language models. 2023; **4.** LangChain Docs. Prompt Templates (ChatPromptTemplate & PromptTemplate concepts). https://python.langchain.com/api_reference/core/prompts/langchain; **5.** LangChain Docs. Build a Retrieval-Augmented Generation (RAG) app (tutorial). https://js.langchain.com/docs/tutorials/rag/; **6.** Douze M, et al. arXiv preprint arXiv:240108281. 2024; **7.** LangChain Docs. create_react_agent (agent creation & notes on production use). https://python.langchain.com/api_reference/langchain/agents/langchain.agents; **8.** LangChain Docs. AgentExecutor (tool-using agent runtime). https://python.langchain.com/api_reference/langchain/agents/langchain.agents.agent.AgentExecutor

Disclosures

AR, **DB**, **AN** and **VI** are employees of PPD™ Evidera™ Real-World Data & Scientific Solutions, Thermo Fisher Scientific. This poster was funded by Thermo Fisher Scientific.

Acknowledgements

Editorial and graphic design support was provided by Caroline Cole and Karissa Calara of Thermo Fisher Scientific.