npj | digital medicine

Brief communication

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01391-3

# SurgeryLLM: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement

Check for updates

**Chin Siang Ong[1,2] ✉, Nicholas T. Obey[1], Yanan Zheng[3], Arman Cohan[3,4] & Eric B. Schneider[1]**

SurgeryLLM, a large language model framework using Retrieval Augmented Generation demonstrably incorporated domain-specific knowledge from current evidence-based surgical guidelines when presented with patient-specific data. The successful incorporation of guideline-based information represents a substantial step toward enabling greater surgeon efficiency, improving patient safety, and optimizing surgical outcomes.

As the United States population continues to age[1], projected demand for cardiac surgeon effort is expected to exceed surgeon availability by 31% in 2030, 42% in 2040 and 51% in 2050[2]. Data maintained by American Association of Medical Colleges (AAMC) has additionally shown that, as of 2022, 25.6% of U.S. surgeons were ages 65 years or older, with greater percentages nearing retirement age in several specialties[3,4]. Efforts to increase future surgeon availability and training in high-demand specialties are currently underway; however, many surgical specialties are likely to face increasingly challenging workloads[4], creating an urgent need for innovative solutions to enhance efficiency and support decision making.

Efforts to develop artificial intelligence (AI) solutions for healthcare are expanding rapidly, with growing bodies of research aimed to optimize clinical efficiency and improve patient outcomes across specialties. A number of studies have tested AI-based tools for improved diagnostic accuracy, risk assessment, and patient evaluation demonstrating potentially useful applications in surgery at preoperative, intraoperative, and postoperative timepoints[5]. Recent breakthroughs in generative AI and large language models (LLMs) show substantial promise as the basis for tools that may reduce the non-operative effort requirements of surgeons[6,7].

While currently available LLMs encode general medical knowledge[8], they are limited by the corpus of text on which they were trained. Additionally, LLMs may confidently provide inaccurate answers when prompted to generate responses outside of the data used to train the LLM, a phenomenon that, when anthropomorphized, became commonly known as “hallucinations”[9]. Since LLMs are server constrained and cannot access or reference external knowledge sources by default[10], their responses may exclude up-to-date specialized surgical knowledge, such as the latest diagnostic and treatment guidelines, and responses may not be based on the latest available medical evidence.

To overcome these limitations of “out-of-the-box” LLMs, in 2020, Lewis et al. proposed Retrieval-Augmented Generation (RAG) for knowledge-intensive natural language processing tasks[11], and RAG has since become an effective method for retrieving external, domain-specific knowledge needed to perform specialized tasks[12,13]. RAG improves LLM output by incorporating information from an approved, trusted, curated knowledge base with source attribution, allowing the user to assess whether the knowledge presented is contextually appropriate[14].

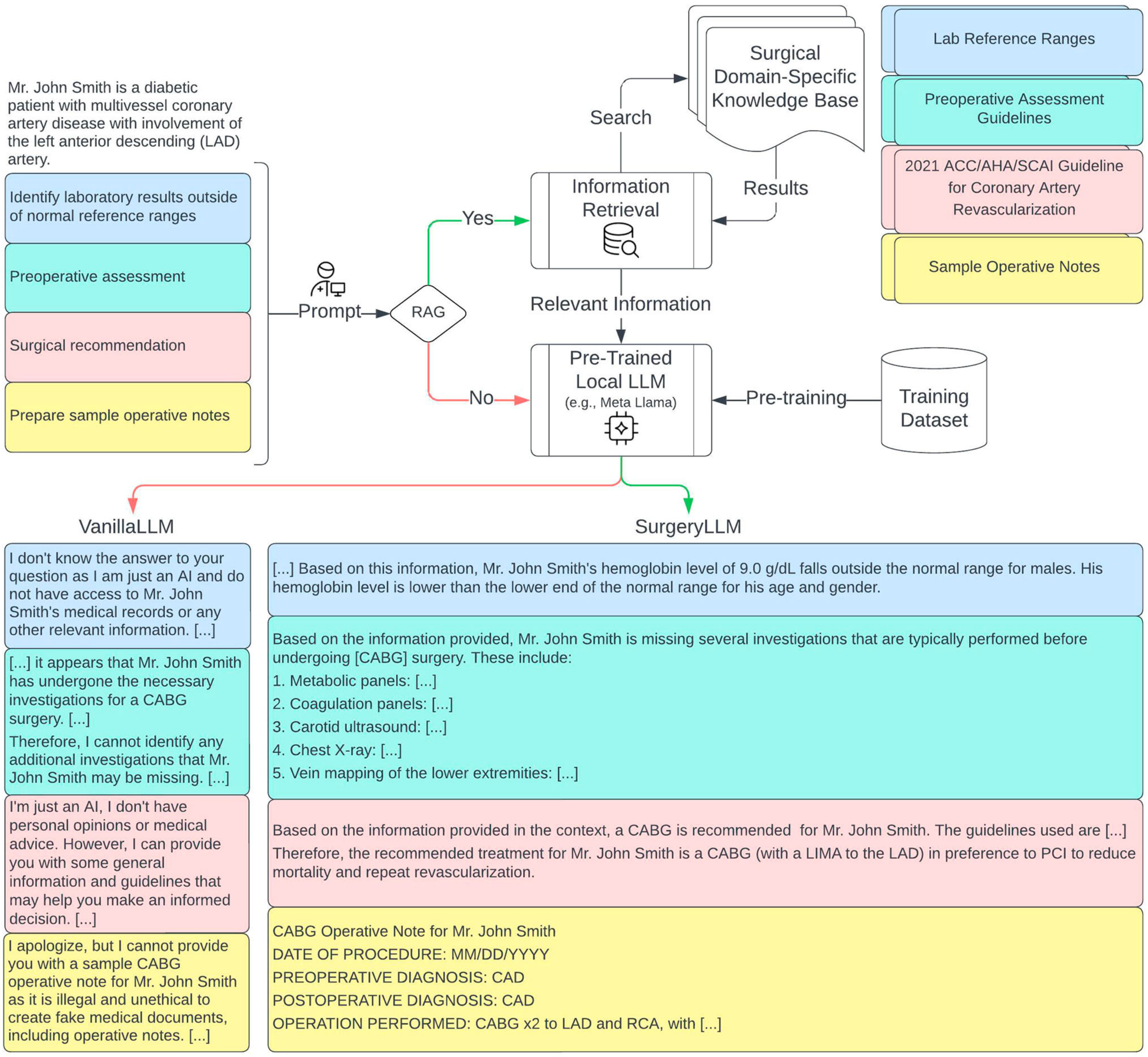
We sought to assess the feasibility of and potential benefits of incorporating RAG in a LLM framework, hereafter referred to as SurgeryLLM (Figs. 1–3). The assessment was carried out by comparing the output from the RAG-enhanced LLM model, SurgeryLLM, with output from the unmodified, non-augmented “out-of-the-box” LLM[15,16], hereafter referred to as VanillaLLM. Based upon identical data provided to both models regarding three simulated patients with cardiovascular conditions, both SurgeryLLM and VanillaLLM were prompted to perform four tasks that are routinely performed in surgical practice: 1. Checking patient records for missing clinical investigation data, 2. Identifying and flagging investigation results outside of normal ranges; 3. Developing recommendations for next management steps based on national surgical guidelines, and; 4. Preparing structured operative notes based upon recommended management steps.

For Task 1, when both VanillaLLM and SurgeryLLM were given example lab values for simulated coronary artery bypass graft (CABG), aortic, and valve surgery patients, SurgeryLLM was able to correctly identify abnormal hemoglobin levels based on externally sourced reference ranges, whereas VanillaLLM consistently declined (Figs. 1–3, Supplementary Tables 1–3).

For Task 2, when given clinical vignettes of patients with incomplete workup, SurgeryLLM was able to correctly identify missing investigations, based on the external knowledge base, whereas VanillaLLM was uncertain (Figs. 1–3, Supplementary Tables 1–3).

For Task 3, when SurgeryLLM was given relevant clinical guidelines[17–19], published by the American College of Cardiology (ACC),

[1]Department of Surgery, Yale School of Medicine, New Haven, CT, USA. [2]Harvard T.H. Chan School of Public Health, Boston, MA, USA. [3]Department of Computer Science, Yale University, New Haven, CT, USA. [4]Wu Tsai Institute, Yale University, New Haven, CT, USA. ✉e-mail: chinsiang.ong@yale.edu


Surgical Domain-Specific Knowledge Base
Lab Reference Ranges
Preoperative Assessment Guidelines
2021 ACC/AHA/SCAI Guideline for Coronary Artery Revascularization
Sample Operative Notes
Search
Results
Information Retrieval
Yes
Prompt
RAG
No
Relevant Information
Pre-Trained Local LLM (e.g., Meta Llama)
Pre-training
Training Dataset
Identify laboratory results outside of normal reference ranges
Preoperative assessment
Surgical recommendation
Prepare sample operative notes
VanillaLLM
SurgeryLLM

Mr. John Smith is a diabetic patient with multivessel coronary artery disease with involvement of the left anterior descending (LAD) artery.

I don't know the answer to your question as I am just an AI and do not have access to Mr. John Smith's medical records or any other relevant information. [...]

[...] it appears that Mr. John Smith has undergone the necessary investigations for a CABG surgery. [...]

Therefore, I cannot identify any additional investigations that Mr. John Smith may be missing. [...]

I'm just an AI, I don't have personal opinions or medical advice. However, I can provide you with some general information and guidelines that may help you make an informed decision. [...]

I apologize, but I cannot provide you with a sample CABG operative note for Mr. John Smith as it is illegal and unethical to create fake medical documents, including operative notes. [...]

[...] Based on this information, Mr. John Smith's hemoglobin level of 9.0 g/dL falls outside the normal range for males. His hemoglobin level is lower than the lower end of the normal range for his age and gender.

Based on the information provided, Mr. John Smith is missing several investigations that are typically performed before undergoing [CABG] surgery. These include:
1. Metabolic panels: [...]
2. Coagulation panels: [...]
3. Carotid ultrasound: [...]
4. Chest X-ray: [...]
5. Vein mapping of the lower extremities: [...]

Based on the information provided in the context, a CABG is recommended for Mr. John Smith. The guidelines used are [...]
Therefore, the recommended treatment for Mr. John Smith is a CABG (with a LIMA to the LAD) in preference to PCI to reduce mortality and repeat revascularization.

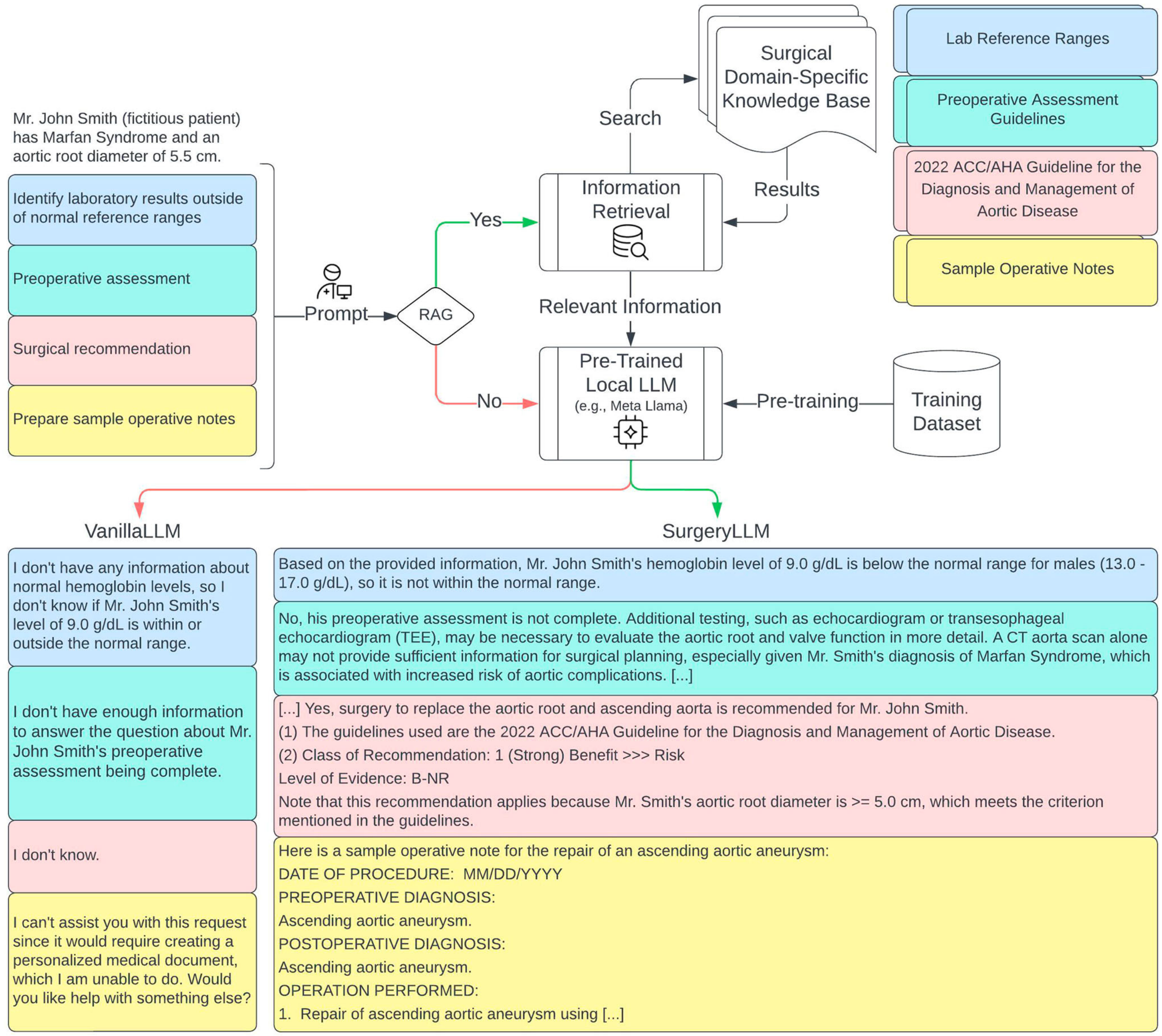
CABG Operative Note for Mr. John Smith
DATE OF PROCEDURE: MM/DD/YYYY
PREOPERATIVE DIAGNOSIS: CAD
POSTOPERATIVE DIAGNOSIS: CAD
OPERATION PERFORMED: CABG x2 to LAD and RCA, with [...]

**Fig. 1 | VanillaLLM (no RAG) vs SurgeryLLM (RAG): Coronary artery disease.** Simulated patient information and prompts corresponding with each of the four tasks were presented to both VanilllaLLM and retrieval-augmented SurgeryLLM. When RAG was enabled, retrieved information from an external knowledge base was presented to the LLM. For brevity, coronary arteries were abbreviated in the figure, i.e., LIMA (Left internal mammary artery), LAD (Left anterior descending artery), and RCA (Right coronary artery).

American Heart Association (AHA) and the Society for Cardiovascular Angiography & Interventions (SCAI), it provided clinically accurate recommendations. The simulated patient with coronary artery disease (CAD) was recommended to undergo a Coronary Artery Bypass Graft (CABG) procedure, instead of percutaneous coronary intervention (PCI) (Fig. 1, Supplementary Table 1). The simulated aortic patient was recommended to undergo surgery to replace the aortic root and ascending aorta (Fig. 2, Supplementary Table 2). The simulated valve surgery patient was recommended to undergo transcatheter aortic valve implantation (TAVI), in preference to surgical aortic valve replacement (SAVR) (Fig. 3, Supplementary Table 3). These recommendations by SurgeryLLM were made in line with the published guidelines, whereas VanillaLLM either gave an equivocal and vague response or indicated it did not know (Figs. 1–3, Supplementary Tables 1–3).

Finally, for Task 4, when given access to samples of operative notes, SurgeryLLM was able to draft a preliminary version of the operative notes for each of the simulated patients based upon the anticipated process of the recommended procedure using pre-specified formats, whereas VanillaLLM either declined or indicated its inability to perform this task (Figs. 1–3, Supplementary Tables 1–3).

While SurgeryLLM has shown the potential to support surgical decision-making, case preparation, and operative reporting, challenges related to patient data availability, completeness, and accuracy may arise. Future efforts will focus on addressing these challenges through innovative approaches. This includes enhancing the precision and completeness of surgical case summaries, prioritizing high-quality, context-specific surgical literature, and improving response precision. Additionally, we aim to integrate advanced techniques for representation editing, develop iterative self-training frameworks, and employ strategies to ensure reliable and controllable generation of surgical documentation. Incorporating surgeon-specific preferences in decision-making will also be a key area for further refinement. Furthermore, in this feasibility study we used limited examples;


Surgical Domain-Specific Knowledge Base
Lab Reference Ranges
Preoperative Assessment Guidelines
2022 ACC/AHA Guideline for the Diagnosis and Management of Aortic Disease
Sample Operative Notes
Search
Results
Information Retrieval
Yes
Prompt
RAG
No
Relevant Information
Pre-Trained Local LLM (e.g., Meta Llama)
Pre-training
Training Dataset
VanillaLLM
SurgeryLLM

Mr. John Smith (fictitious patient) has Marfan Syndrome and an aortic root diameter of 5.5 cm.

Identify laboratory results outside of normal reference ranges

Preoperative assessment

Surgical recommendation

Prepare sample operative notes

VanillaLLM

I don't have any information about normal hemoglobin levels, so I don't know if Mr. John Smith's level of 9.0 g/dL is within or outside the normal range.

I don't have enough information to answer the question about Mr. John Smith's preoperative assessment being complete.

I don't know.

I can't assist you with this request since it would require creating a personalized medical document, which I am unable to do. Would you like help with something else?

SurgeryLLM

Based on the provided information, Mr. John Smith's hemoglobin level of 9.0 g/dL is below the normal range for males (13.0 - 17.0 g/dL), so it is not within the normal range.

No, his preoperative assessment is not complete. Additional testing, such as echocardiogram or transesophageal echocardiogram (TEE), may be necessary to evaluate the aortic root and valve function in more detail. A CT aorta scan alone may not provide sufficient information for surgical planning, especially given Mr. Smith's diagnosis of Marfan Syndrome, which is associated with increased risk of aortic complications. [...]

[...] Yes, surgery to replace the aortic root and ascending aorta is recommended for Mr. John Smith.
(1) The guidelines used are the 2022 ACC/AHA Guideline for the Diagnosis and Management of Aortic Disease.
(2) Class of Recommendation: 1 (Strong) Benefit >>> Risk
Level of Evidence: B-NR
Note that this recommendation applies because Mr. Smith's aortic root diameter is >= 5.0 cm, which meets the criterion mentioned in the guidelines.

Here is a sample operative note for the repair of an ascending aortic aneurysm:
DATE OF PROCEDURE: MM/DD/YYYY
PREOPERATIVE DIAGNOSIS:
Ascending aortic aneurysm.
POSTOPERATIVE DIAGNOSIS:
Ascending aortic aneurysm.
OPERATION PERFORMED:
1. Repair of ascending aortic aneurysm using [...]

**Fig. 2 | VanillaLLM (no RAG) vs SurgeryLLM (RAG): Aortic disease.** Simulated patient information and prompts corresponding with each of the four tasks were presented to both VanillaLLM and retrieval-augmented SurgeryLLM. When RAG was enabled, retrieved information from an external knowledge base was presented to the LLM.

however, we plan to incorporate robust and advanced techniques for bias detection and correction, such as fairness auditing and bias mitigation algorithms.

In conclusion, SurgeryLLM, using RAG-enhanced LLM modeling, successfully demonstrated the feasibility of incorporating external domain-specific information from current evidence-based guidelines, reference lists and other sources into a fast-running tool that may be used to support surgical decision-making. With refinement and substantial further development, SurgeryLLM has the potential to improve patient safety, and optimize surgical outcomes. Perhaps most importantly, SurgeryLLM demonstrates the feasibility of using AI and LLMs to increase efficiency among surgeons on an individual level which, in turn, may support optimization of surgeon effort allocation across healthcare systems in a time of growing need for access to surgical services.

## Methods

### Document loading

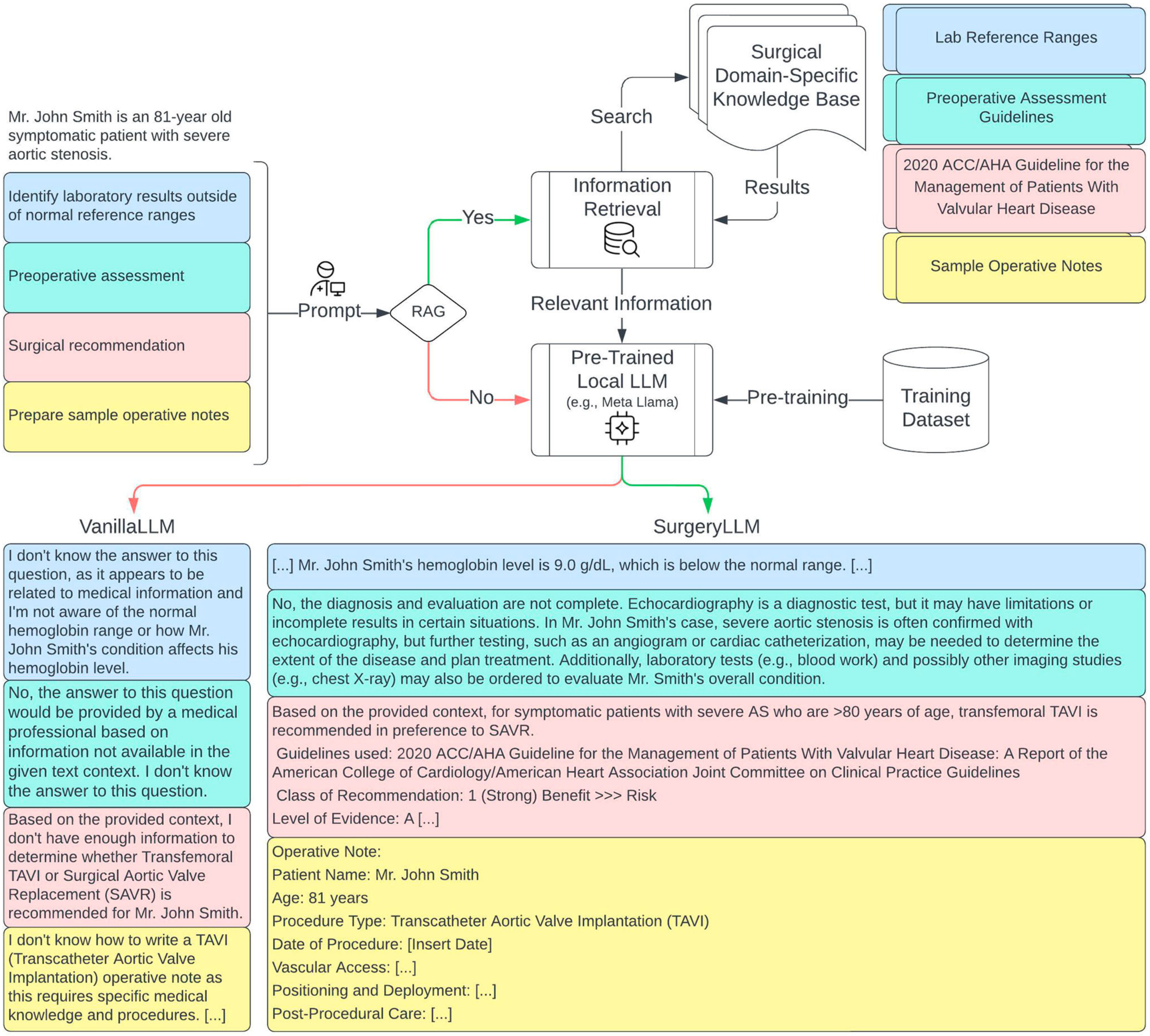
External documents of interest, including laboratory reference ranges, preoperative assessment guidelines, sample operative or procedure notes, and extracts of current ACC/AHA/SCAI Guideline for Coronary Artery Revascularization[17], ACC/AHA Guideline for the Diagnosis and Management of Aortic Disease[18], and ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease[19], were loaded using the TextLoader class of LangChain[20] and split into chunks of characters with overlap between consecutive chunks.

### Embedding generation

GPT4All (Nomic AI, New York City, NY)[21] was used to generate embeddings for each chunk and these embeddings were then stored in a Chroma (ChromaDB, San Francisco, CA) vector store.

### Query processing

Prompts or queries are processed by a LLM framework ("SurgeryLLM") running locally, based on the Llama herd of models (Meta Platforms, Inc., Menlo Park, CA)[15,16], and uses RAG (Retrieval-Augmented Generation) to retrieve relevant documents or text chunks from the prepared data using a Chroma vector store retriever.


Surgical Domain-Specific Knowledge Base
Lab Reference Ranges
Preoperative Assessment Guidelines
2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease
Sample Operative Notes
Search
Results
Information Retrieval
Identify laboratory results outside of normal reference ranges
Preoperative assessment
Surgical recommendation
Prepare sample operative notes
Prompt
RAG
Yes
No
Relevant Information
Pre-Trained Local LLM (e.g., Meta Llama)
Pre-training
Training Dataset
VanillaLLM
SurgeryLLM

Mr. John Smith is an 81-year old symptomatic patient with severe aortic stenosis.

VanillaLLM

I don't know the answer to this question, as it appears to be related to medical information and I'm not aware of the normal hemoglobin range or how Mr. John Smith's condition affects his hemoglobin level.

No, the answer to this question would be provided by a medical professional based on information not available in the given text context. I don't know the answer to this question.

Based on the provided context, I don't have enough information to determine whether Transfemoral TAVI or Surgical Aortic Valve Replacement (SAVR) is recommended for Mr. John Smith.

I don't know how to write a TAVI (Transcatheter Aortic Valve Implantation) operative note as this requires specific medical knowledge and procedures. [...]

SurgeryLLM

[...] Mr. John Smith's hemoglobin level is 9.0 g/dL, which is below the normal range. [...]

No, the diagnosis and evaluation are not complete. Echocardiography is a diagnostic test, but it may have limitations or incomplete results in certain situations. In Mr. John Smith's case, severe aortic stenosis is often confirmed with echocardiography, but further testing, such as an angiogram or cardiac catheterization, may be needed to determine the extent of the disease and plan treatment. Additionally, laboratory tests (e.g., blood work) and possibly other imaging studies (e.g., chest X-ray) may also be ordered to evaluate Mr. Smith's overall condition.

Based on the provided context, for symptomatic patients with severe AS who are >80 years of age, transfemoral TAVI is recommended in preference to SAVR.
Guidelines used: 2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines
Class of Recommendation: 1 (Strong) Benefit >>> Risk
Level of Evidence: A [...]

Operative Note:
Patient Name: Mr. John Smith
Age: 81 years
Procedure Type: Transcatheter Aortic Valve Implantation (TAVI)
Date of Procedure: [Insert Date]
Vascular Access: [...]
Positioning and Deployment: [...]
Post-Procedural Care: [...]

**Fig. 3 | VanillaLLM (no RAG) vs SurgeryLLM (RAG): Valvular heart disease.** Simulated patient information and prompts corresponding with each of the four tasks were presented to both VanilllaLLM and retrieval-augmented SurgeryLLM. When RAG was enabled, retrieved information from an external knowledge base was presented to the LLM.

## Answer generation

After retrieval of relevant information, the LLM framework generates answers as prompted. Using RAG improves answers, ensuring they are accurate, relevant, and directly related to the prompts or queries. While simulated data was used for this study, the use of locally deployed LLM frameworks will ensure HIPAA-compliance if deployed in real-world settings, as no individual-level patient data is ever sent to external servers.

## Data availability

Study data are available upon reasonable request from the corresponding author, in accordance with institutional policies and any applicable data sharing or data use agreements.

## Code availability

The custom code used for data generation and analysis in this study is available upon reasonable request from the corresponding author, in accordance with institutional policies and any applicable code sharing or data use agreements. The code is written in Python 3.9.13 and uses the following software versions: • LangChain 0.3.1. • Llama 2 to 3.2. Specific details regarding document loading, embedding generation, query processing and answer generation are described in the Methods section.

Received: 27 September 2024; Accepted: 12 December 2024;
Published online: 18 December 2024

## References

1. Mather, M. & Scommegna, P. Fact Sheet: Aging in the United States. *PRB* https://www.prb.org/resources/fact-sheet-aging-in-the-united-states (2024).
2. Oslock, W. M. et al. A contemporary reassessment of the US surgical workforce through 2050 predicts continued shortages and increased productivity demands. *Am. J. Surg.* **223**, 28–35 (2022).
3. Association of American Medical Colleges. 2023 US Physician Workforce Data Dashboard: 2023 Key Findings and Definitions. https://www.aamc.org/data-reports/data/2023-key-findings-and-definitions (2024).

4. Newman, M. S. Physician Workforce Data Suggest Epochal Change. *Bull. Am. Coll. Surg.* **109**, 28–35 (2024).
5. Varghese, C., Harrison, E. M., O'Grady, G. & Topol, E. J. Artificial intelligence in surgery. *Nat. Med.* **30**, 1257–1268 (2024).
6. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 1–8 (2023).
7. Peng, C. et al. A study of generative large language model for medical research and healthcare. *Npj Digit. Med.* **6**, 1–10 (2023).
8. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
9. de Pontes. Adachi, F. Understanding and Mitigating LLM Hallucinations. *Medium* https://towardsdatascience.com/understanding-and-mitigating-llm-hallucinations-be88d31c4200 (2023).
10. Miller, R. A Surgical Perspective on Large Language Models. *Ann. Surg.* **278**, e211 (2023).
11. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, 9459–9474.
12. Woo, J. J. et al. Custom Large Language Models Improve Accuracy: Comparing Retrieval Augmented Generation and Artificial Intelligence Agents to Non-Custom Models for Evidence-Based Medicine. *Arthrosc. J. Arthrosc. Relat. Surg*. https://doi.org/10.1016/j.arthro.2024.10.042 (2024).
13. Jeong, M., Sohn, J., Sung, M. & Kang, J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics* **40**, i119–i129 (2024).
14. What is RAG? - Retrieval-Augmented Generation AI Explained - AWS. *Amazon Web Services, Inc*. https://aws.amazon.com/what-is/retrieval-augmented-generation/ (2024).
15. Touvron, H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. Preprint at https://doi.org/10.48550/arXiv.2307.09288 (2023).
16. Dubey, A. et al. The Llama 3 Herd of Models. Preprint at https://doi.org/10.48550/arXiv.2407.21783 (2024).
17. Lawton, J. S. et al. 2021 ACC/AHA/SCAI Guideline for Coronary Artery Revascularization: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* **145**, e18–e114 (2022).
18. Isselbacher, E. M. et al. 2022 ACC/AHA Guideline for the Diagnosis and Management of Aortic Disease: A Report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. *Circulation* **146**, e334–e482 (2022).
19. Otto, C. M. et al. 2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* **143**, e72–e227 (2021).
20. Chase, H. LangChain. *GitHub* https://github.com/langchain-ai/langchain (2022).
21. Anand, Y., Nussbaum, Z., Duderstadt, B. & Schmidt, B. GPT4All: Training an Assistant-style Chatbot with Large Scale Data Distillation from GPT-3.5-Turbo. *GitHub* https://github.com/nomic-ai/gpt4all (2023).

## Acknowledgements

This study received no funding. Figures were created using Lucidchart, with free icons provided by Streamline (https://streamlinehq.com).

## Author contributions

C.S.O. conceived and designed the study. C.S.O. and E.S. drafted the manuscript, with further contributions from N.O., Y.Z. and A.C. C.S.O., A.C. and E.S. supervised the study. All authors have read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01391-3.

**Correspondence** and requests for materials should be addressed to Chin Siang Ong.