Exploring The Use Of Customized Large Language Models For Potential In Vascular Surgery: A Pilot Study Involving The VESAP Board Exam

Peter Vien, Elie Donath MD MPH[1], Joshua Le, Aryan Naik, Sean Liebscher MD[2], Ben Carriveau, Daniel Bertges[2]

[1]Dept of Medicine, University of Vermont, Burlington VT

[2]Dept of Vascular Surgery, University of Vermont, Burlington VT

elie.donath@uvmhealth.org

UVMHealth.org/MedCenter

OBJECTIVE

- To develop large language models (LLM's) customized for vascular surgery and compared to proprietary models in answering the VESAP-5 questionnaire

BACKGROUND

- Large language models (LLMs) have demonstrated capabilities that rival human performance across multiple disciplines of medical study.
- While commercially available LLM's are extremely powerful, recent advances have materialized allowing for customization enabling them to be tailored for highly specific tasks or domains
- We explore a vascular surgery-focused LLM customization strategy revolving around the use of retrieval augmented generation (RAG).

AIMS

- We use the VESAP-5 standardized board exam to test the ability of a customized LLM compared to a "zero-shot" (or baseline) LLM.
- Little is known about LLM customization strategies in healthcare - they may pave the way for more sophisticated and compact models that could allow for broader implementation in medical subspecialties

PROMPT PROVIDED TO LLM

Use the following pieces of information to answer the user's question.

If you don't know the answer, just say that you don't know, don't try to make up an answer.

Context: {context}

Question: {question}|

Answer the question and provide additional helpful information, based on the pieces of information, if applicable.

Be elaborate in the rationale you provide for your answer.

Always start your answer off with "The answer is " containing the multiple choice response (which has to be either A,B,C,D,E).

Responses should be properly formatted to be easily read.

METHODS

- MCQ's from the Society for Vascular Surgery's Self-Assessment Program (VESAP) consisted of 680 questions and covered thirteen subsections of vascular surgery
- Three base (i.e. Zero-Shot) LLMs were examined
  - Anthropic's closed-source Claude 3.0 (March 14, 2024)
  - OpenAI's closed-source ChatGPT-4o (May 14, 2023)
  - Meta's open-source LLAMA-3 70B model (April 18, 2024). LLAMA-3 was further interrogated with a RAG-based architecture.
- Three separate sources of vascular surgery text were selected to build an embeddings framework which the RAG architecture could access.
  - Vascular Surgery Exam Prep (Audible Bleeding)
  - Transcripts from 200 videos of varying lengths from the YouTube channel created by Houston Methodist DeBakey Heart and Vascular Center
  - Rutherford's Vascular Surgery Textbook
- Primary outcome was test performance of the different LLM's. Also ROUGE/BERT scoring to compare how well the LLM-provided rationale compared to the underlying "gold-standard" rationale provided by VESAP5.

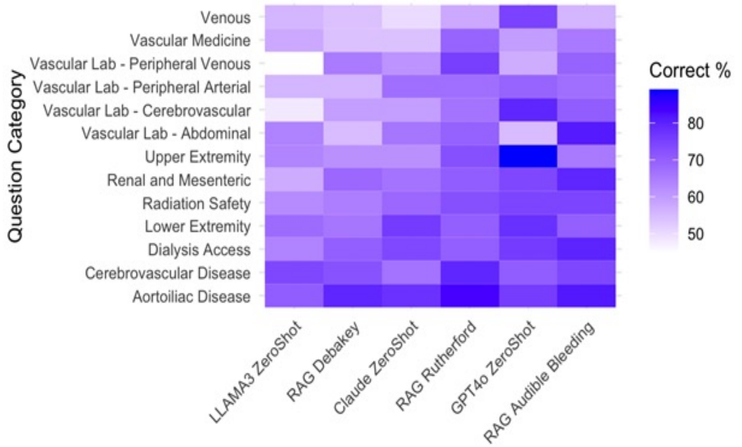
RESULTS

Overall Model Performance

We compared three base LLM's to three version of LLAMA-3 customized by additional vascular surgery knowledge

LLM Version	Accuracy	Percentage
LLAMA3 - base	412/680	60.6%
Claude	444/680	65.3%
ChatGPT	489/680	71.9%
LLAMA3- Audible Bleeding	444/680	71.9%
LLAMA3 - Rutherford	484/680	71.2%
LLAMA3 - Debakey	432/680	63.5%

RESULTS



ROUGE/BERT Scoring

- We compared the "gold standard" VESAP-provided Discussion to an LLM-provided "rationale" for its response
- The zero-shot Claude model aligned the closest in terms of meaning to the "gold standard" text
- The RAG-based models demonstrated a high degree of variability in scoring – in some cases the were able to very closely adhere to the VESAP-provided discussion but in many cases they began to hallucinate

CONCLUSIONS

- LLMs generate accurate responses to vascular surgery questions with semantically appropriate explanations.
- Customized models using approaches like RAG show equal promise, suggesting the potential for LLMs to be trained on specialist information.

REFERENCES

1. OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/
2. Grattafiori, A, Dubey A, et al. The LLAMA3 Herd of Models. arXiv:2407.21783
3. Anthropic. (2023). Claude (Oct 8 version) [Large language model]. https://www.anthropic.com/