Advance Publication

Circulation Reports
doi:10.1253/circrep.CR-25-0094

RESEARCH LETTER

Performance Evaluation of Large Language Models With Retrieval-Augmented Generation in Cardiology Specialist Examinations in Japan

Hiromasa Hayama, MD, PhD; Tu Hao Tran, BMedSc (Hons), MBBS; Jin Kirigaya, MD, PhD; Yosuke Katayama, MD, PhD; Tomoko Negishi, MD; Koya Ozawa, MD, PhD; Kazuaki Negishi, MD, PhD, MSc

Background: Large language models (LLMs) have shown potential in medical education, but their application to cardiology specialist examinations remains underexplored. We compared the performances of a retrieval-augmented generation LLM (RAG-LLM) 'CardioCanon' against general-purpose LLMs.

Methods and Results: A total of 96 publicly available text-based open-source multiple-choice questions from the Japanese Cardiology Specialist Examination (1997–2022) were used. CardioCanon showed similar option-level accuracy to ChatGPT-4o and Gemini 2.0 Flash (81.0%, 76.0%, and 77.2%, respectively), but higher case-based accuracy than ChatGPT (57.3% vs. 29.2%, P<0.001).

Conclusions: RAG techniques can enhance AI-assisted examination performance by improving case-level reasoning and decision-making.

Key Words: Cardiology examination; Large language model; Medical education; Retrieval augmented generation

Large language models (LLMs) such as ChatGPT have shown growing potential in medical education, particularly for high-stakes standardized assessments such Medical Licensing Examinations.[1,2] However, general-purpose LLMs remain limited by misinformation, hallucinations, and insufficient domain-specific contextual knowledge. The application of LLMs to medical specialty board examinations, such as cardiology, remains underexplored.

To address this, we developed CardioCanon, a retrieval-augmented generation (RAG) architecture that integrates a pre-trained LLM (ChatGPT-4o) with a cardiology-specific corpus.[3,4] This study evaluated its performance on Japanese Cardiology Specialist Examination questions compared with that of general-purpose LLMs.

Methods

We compiled a dataset of 96 publicly available multiple-choice questions from the Japanese Cardiology Specialist Examination (1997–2022). In CardioCanon, 664 documents (publicly available clinical guidelines and clinical trial data up to year 2024) were embedded using the text-embedding-ada-002 model (OpenAI) to create 1,536 high-dimensional vector representations.[3] These embeddings were stored and managed using cloud-based vector storage, Pinecone, which enabled similarity-based retrieval of the guideline content relevant to the input query. During inference, CardioCanon retrieved semantically similar and relevant passages from the vector store using a conversational retrieval QA chain that integrated 3 key components: (1) the Pinecone Retriever for document retrieval, (2) ChatGPT-4o model as the inference engine with temperature set to 0.4 for output stability, and (3) Conversation Summary Memory to preserve dialogue history for multi-turn interactions. The architecture follows a modular design, allowing real-time retrieval-augmented reasoning without fine-tuning the underlying LLM.

Received May 18, 2025; accepted May 18, 2025; J-STAGE Advance Publication released online July 2, 2025 Time for primary review: 1 day

School of Clinical Medicine, University of New South Wales, Sydney, NSW (H.H., J.K., Y.K., T.N., K.O., K.N.); The Ingham Institute for Applied Medical Research, Sydney, NSW (H.H., J.K., Y.K., T.N., K.O., K.N.); Imaging and Phenotyping Laboratory, Charles Perkins Centre, University of Sydney, Sydney, NSW (T.H.T.); Department of Cardiology, Liverpool Hospital, Liverpool, NSW (K.N.); and Victor Chang Cardiac Research Institute, Darlinghurst, NSW (K.N.), Australia

Mailing address: Kazuaki Negishi, MD, PhD, MSc, Ingham Institute for Applied Medical Research, 1 Campbell St, Liverpool NSW 2170, Sydney, Australia. email: k.negishi@unsw.edu.au

ISSN-2434-0790

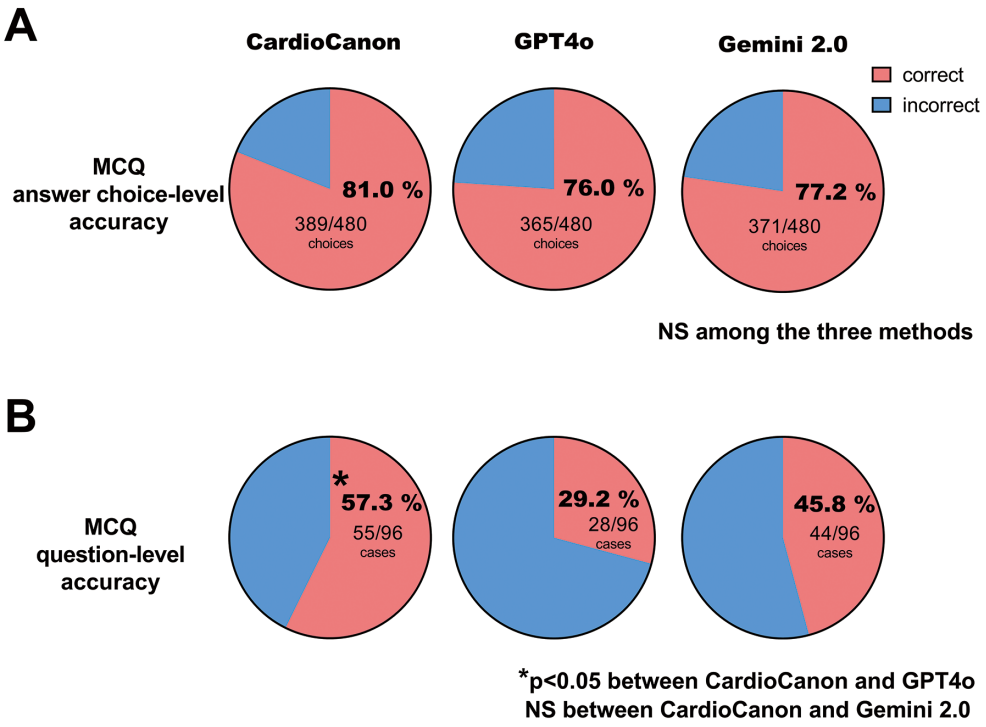

Figure. Comparison of model performance: (**A**) option-level accuracy (proportion of correct choices across all MCQ options); (**B**) case-based accuracy (proportion of fully correct responses per question). MCQ, multiple-choice question; NS, not significant.

CardioCanon was compared to general-purpose LLMs ChatGPT-4o (Open AI) and Gemini-2.0-Flash (Google DeepMind), evaluated using 2 metrics: (1) option-level accuracy, defined as the proportion of correct answer choices across all options, and (2) case-based accuracy, defined as the proportion of fully correct responses per question. These metrics were chosen to distinguish between partial and complete task resolution and assess the model's ability to produce clinically coherent responses. A chi-square test was used for statistical comparison. Two-sided $P<0.05$ was accepted as indicating statistical significance.

Results

CardioCanon achieved an option-level accuracy of 81.0% (389/480 choices), which was comparable to that of ChatGPT-4o (76.0%, 365/480 choices) (P=0.07) and Gemini-2.0-Flash (77.2%, 371/480) (P=0.18) (**Figure A**). However, CardioCanon showed a significantly higher case-based accuracy: 57.3% (55/96), compared with 29.2% (28/96 cases) (P<0.001) for ChatGPT-4o and 45.8% (44/96) (P=0.15) for Gemini-2.0-Flash (**Figure B**).

Discussion

This study demonstrated that RAG enhanced the reasoning performance of LLMs in cardiology specialty certification examinations. CardioCanon showed higher accuracy in question-level accuracy than ChatGPT-4o, although similar performances were observed in choice-level accuracy. Importantly, performance improvement was achieved without fine-tuning the underlying LLM. Instead, it resulted from dynamic linkage to external expert knowledge, which is especially valuable in high-stakes testing environments where reference consistency and reliability are essential. The observed improvement in case-based accuracy suggests that the model could go beyond pattern recall and generate clinically coherent responses grounded in structured guideline information. Similar findings were reported for US orthopedic board questions.[5]

These findings indicate that RAG-based architecture may serve as an effective tool in medical education and specialist assessment, particularly where reliable literature-based responses are critical. The multimodal abilities of ChatGPT-4o will need to be formally assessed for performance in the accurate analysis of medical images and graphics when it is more capable, likely in future versions.

Study Limitation

Our analysis was constrained to the text-based capabilities of LLMs because image-based questions were excluded from analysis, as the multimodal capabilities of current LLMs in accurate medical image interpretation remain in their infancy.

Conclusions

CardioCanon, a RAG-LLM model, demonstrated a significantly higher case-based accuracy than general-purpose LLMs in the context of Japanese cardiology board examination questions. These results highlight the utility of RAG in enhancing AI-driven medical education and assessment.

With the advent of multimodal models, further research is warranted.

References

1. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health* 2023; **2:** e0000198.
2. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of generative pretrained transformer on the National Medical Licensing Examination in Japan. *PLoS Digital Health* 2024; **3:** e000043.
3. OpenAi. Vector embeddings: OpenAI API documentation 2024, https://platform.openai.com/docs/guides/embeddings (accessed May 17, 2025).
4. Tran T, Joseph V, Smith L, Hopkins A, Lo S, Hennessy A, et al. CardioCanon: A customised chatbot for cardiology inquiry with retrieval augmented generation to reduce hallucinations and improve performance of large language models. *Heart Lung Circ* 2024; **33:** S379–S380.
5. Eskenazi J, Krishnan V, Konarzewski M, Constantinescu D, Lobaton G, Dodds SD. Evaluating retrieval augmented generation and ChatGPT's accuracy on orthopaedic examination assessment questions. *Ann Joint* 2025; **10:** 12, https://aoj.amegroups.org/article/view/9194 (accessed May 17, 2025).