Performance of Large Language Models in Real-World Interventional Cardiology Scenarios: The ILLUMINATE Randomised, Blinded Evaluation Study

Authors: *Attilio Lauretti,[1,9] Iginio Colaiori,[1] Simone Calcagno,[2] Enrico Romagnoli,[3] Fabrizio D'Ascenzo,[4,5] Antonio Di Matteo,[1] Francesco Gemelli,[1] Gaetano Pero,[1] Marco Bernardi,[1,6] Luigi Spadafora,[1,6] Antonio Esposito,[8] Marco Borgi,[1] Giuseppe Biondi-Zoccai,[6,7] Francesco Versaci[1]

1. Division of Cardiology, Santa Maria Goretti Hospital, Latina, Italy
2. Cardiology Unit, Department of Emergency and Admission, San Paolo Hospital, Civitavecchia, Italy
3. Department of Cardiovascular Sciences, Fondazione Policlinico Agostino Gemelli IRCCS, Rome, Italy
4. Division of Cardiology, Cardiovascular and Thoracic Department, Città della Salute e della Scienza, Turin, Italy
5. Division of Cardiology, Department of Medical Sciences, University of Turin, Italy
6. Department of Medical-Surgical Sciences and Biotechnologies, Sapienza University of Rome, Latina, Italy
7. Maria Cecilia Hospital, GVM Care & Research, Cotignola, Italy
8. ICOT Marco Pasquali Institute, Cardiovascular Department, Latina, Italy
9. Department of Clinical and Molecular Medicine, Sapienza University of Rome, Italy

*Correspondence to attilio.lauretti@uniroma1.it

Disclosure: The authors have declared no conflicts of interest.

Keywords: AI, interventional cardiology, large language models (LLM).

Citation: EMJ Int Cardiol. 2025;13[1]:33-35. https://doi.org/10.33590/emjintcardiol/WUXT8325

BACKGROUND

The integration of AI in cardiology has advanced considerably with the emergence of large language models (LLM), which offer new perspectives for clinical and interventional decision support.[1,2] However, few studies to date have assessed their reliability in complex, real-world interventional cardiology cases.[3-5] The ILLUMINATE[6] study is a randomised, blinded evaluation that compares multiple LLMs in high-complexity clinical scenarios reflective of contemporary interventional practice.

METHODS

This study involved 20 anonymised cases (10 coronary artery disease and 10 structural heart disease), each presenting significant diagnostic or therapeutic complexity. Six LLMs were tested: default ChatGPT (ChatGPTd; OpenAI, San Francisco, California, USA), ChatGPT with embedded European Society of Cardiology guidelines (ChatGP-gl), ChatGPT with internet-enabled search (ChatGPTi), Perplexity AI (San Francisco, California, USA), Mistral AI (Paris, France), and Gemini (Google, San Francisco, California, USA). For each case, models were prompted to offer a conclusive clinical recommendation. Their outputs were then randomised, anonymised, and blindly scored by five independent interventional cardiologists based on five predefined criteria: appropriateness, accuracy, relevance, clarity, and clinical utility. Each criterion was rated on a 0–10 scale, with composite scores calculated for comparative analysis using a mixed linear model.

RESULTS

A total of 120 evaluations were conducted. The mean composite score was 7.1 (95% CI: 7.0–7.2), though performance varied significantly across different models ($p<0.001$). ChatGPTi and ChatGPT-gl demonstrated superior performance with scores of 7.8 (95% CI: 7.5–8.0) and 7.7 (95% CI: 7.4–7.9), respectively. Intermediate performance was seen with Mistral AI (7.0), Perplexity AI (7.0), and ChatGPTd (6.9), while Gemini scored the lowest (6.3). No performance differences were found between coronary artery disease and structural heart disease cases ($p=0.900$), suggesting robustness across clinical domains. (Figure 1)

Models equipped with web search or guideline integration consistently outperformed those without, underscoring the value of external data access

for accurate, actionable responses. Nonetheless, no model reached optimal scores, and additional prompting was often required to elicit a definitive recommendation, underlining current limitations in LLM autonomy and clinical reasoning. Inter-rater reliability scoring variability was also observed.

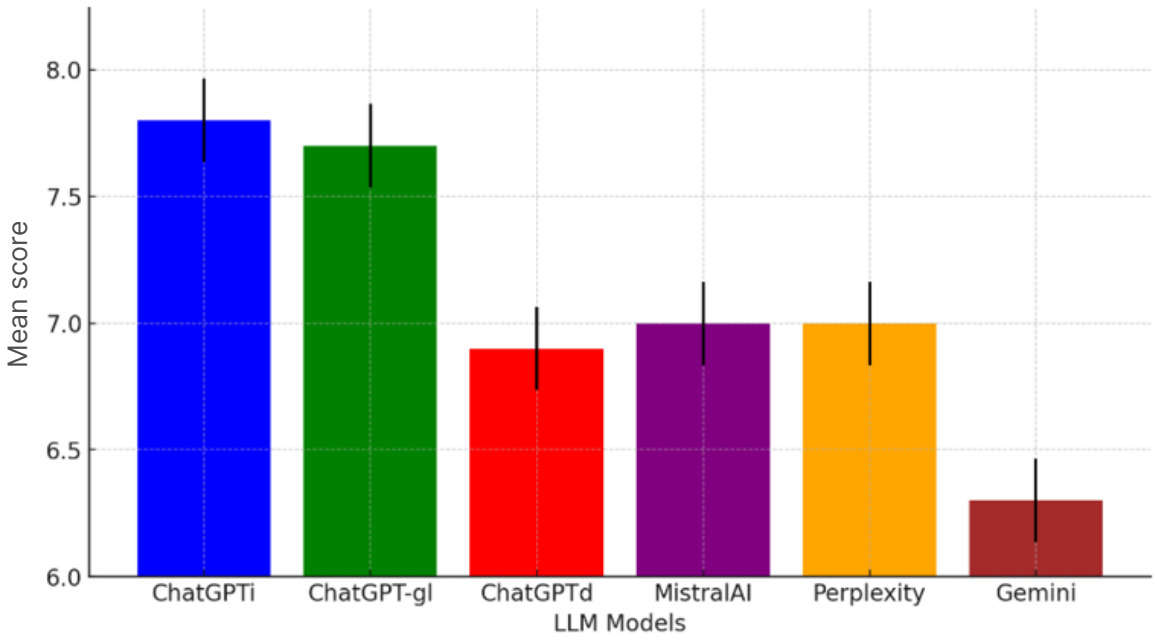
CONCLUSION

The implications of these findings are twofold. First, LLMs may represent a useful adjunct in the management of interventional cardiology cases, particularly when enhanced with guideline-based or real-time data access. Second, these tools remain currently immature for autonomous decision-making and require further development to ensure consistency, contextual awareness, and safety in patient care.

Importantly, the ILLUMINATE study highlights the need for a regulatory oversight and physician involvement in AI deployment. While LLMs show promise as decision-support tools, their integration into clinical workflows must proceed cautiously. Future research should focus on improving interpretability, minimising hallucinations, and enabling dynamic updating with the latest evidence.

In conclusion, the ILLUMINATE study demonstrates that while LLMs can assist in complex interventional cardiology scenarios, their performance is highly variable and contingent on model configuration. The best-performing systems were those equipped with structured access to medical guidelines and web data. These results support the potential of LLMs as a valuable complement, and not as a replacement, to human expertise in high-stakes cardiovascular care.

Figure 1: Graphical summary about the mean performances of large language models with confidence intervals.



ChatGPTd: default ChatGPT; ChatGPT-gl: ChatGPT with embedded European Society of Cardiology guidelines; ChatGPTi: ChatGPT with internet-enabled search; LLM: large language model.

References

1. Alexandrou M et al. Performance of ChatGPT on ACC/SCAI interventional cardiology certification simulation exam. JACC Cardiovasc Interv. 2024;17(10):1292-3.
2. Geneş M et al. Assessment of ChatGPT's compliance with ESC-acute coronary syndrome management guidelines at 30-day intervals. Life (Basel). 2024;14(10):1235.
3. Itelman E et al. AI-assisted clinical decision making in interventional cardiology: the potential of commercially available large language models. JACC Cardiovasc Interv. 2024 Aug 12;17(15):1858-60.
4. Masanneck L et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. J Med Internet Res. 2024;26:e53297.
5. Salihu A et al. Towards AI-assisted cardiology: a reflection on the performance and limitations of using large language models in clinical decision-making. EuroIntervention. 2023;19(10):e798-801.
6. Lauretti et al. Illuminate study: comparing large language models in complex interventional cardiology cases. Abstract A66015AL. EuroPCR 2025, 20-23 May, 2025.