Co-pilot, Not Autopilot: A Practical Method for Using Large Language Models in Interventional Cardiology

Authors:	*Louis-Marie Desroche,[1] Christopher Cook,[2,3] Thomas Modine[4] 1. Interventional Cardiology Unit, University Hospital of La Réunion, Saint-Denis, France 2. The Essex Cardiothoracic Centre, Basildon, UK 3. Anglia Ruskin University, Cambridge, UK 4. Department of Cardiac Surgery, CHU de Bordeaux, Haut-Lévêque Hospital, France *Correspondence to louis-marie.desroche@chu-reunion.fr
Disclosure:	Cook has served as a consultant for Edwards, Abbott, and Philips; and holds equity in Viz.ai and cerebria.ai. Modine reports receiving grants/contracts and consulting fees from Abbott, Edwards, Medtronic, and Jenscare; advisory board participation for Abbott, Edwards, Medtronic, and Jenscare; and patents planned, issued or pending; leadership/fiduciary roles; and stock/stock options in Versa Vascular. Desroche has declared no conflicts of interest.
Acknowledgements:	This feature is based on the AI Lab session 'AI fundamentals for busy cardiologists', presented at EuroPCR 2026, Paris, France.
Keywords:	AI, clinical decision support, Heart Team, interventional cardiology, large language models (LLM), prompt engineering.
Citation:	EMJ Int Cardiol. 2026;14[1]:22-25. https://doi.org/10.33590/emjintcardiol/0I89LD76

WHEN A LARGE language model (LLM) gives a cardiologist a poor answer, it is not always the model that is the only problem. More often than we care to admit, it is the briefing (or 'prompt') that is poor. Put simply, the way we ask a question shapes the answer that we get, and most of us were never taught how to ask.

This is not a fringe concern; these tools are already in daily use, whether that be for discharge letters, guideline checks, or translation. Indeed, the adoption of LLMs applied to everyday clinical tasks has proliferated faster than anyone has taught us to use them. Furthermore, the evidence base for the accuracy of LLMs does not seem to match the enthusiasm of its adopters. A large systematic review of 519 studies found that only 5% used real patient-care data; 44.5% tested examination-style knowledge and 84.2% addressed question-answering.[1] Some LLMs reach passing or near-passing scores on selected cardiology board-style examinations; however, this reflects 'exam' evidence, not 'bedside' evidence. Accordingly, we have a tool that is already in widespread use, but with limited proof of appropriateness at the bedside, compounded by no shared method for using it optimally. This article offers one practical way to narrow that gap.

WHAT ACTUALLY IS A LLM, AND THE DISTINCTION THAT MATTERS

In simplified terms, a LLM is a very large 'autocomplete'; the same predictive-text mechanism that suggests the next word on your mobile phone, but scaled up tremendously. Trained on much of the public internet, LLMs have learned one task; specifically, given a sequence of words,

to predict probabilistically the next most likely word to follow. Therefore, contrary to many people's belief or understanding, there is no medical 'brain' intrinsic to LLMs. Furthermore, there is no causal reasoning, no awareness of your patient, and unless it is connected to retrieval or external tools, its built-in knowledge is limited to what it was exposed to during its training. Scaled massively, that simple mechanism is impressive, indeed capable of passing cardiology board examinations, but fluency in language does not equate to accuracy of language.

For our specialty, one key distinction is decisive. Two different technologies travel under the umbrella label of 'AI in cardiology'. Task-specific deep learning, the networks behind AI-ECG and AI-echocardiography, is regulatory-cleared and has been validated in large cohorts and randomised trials. For example, the AI-ECG screen for low ejection fraction was tested in a pragmatic randomised trial of 22,641 patients.[2] However, LLMs have not been tested or validated anywhere near as robustly.

Indeed, to the authors' knowledge, no LLM-based cardiology decision-support system has yet been CE-marked or FDA-cleared for autonomous clinical decision-making. Acknowledging this, the EU AI Act classifies medical AI as 'high-risk', which mandates human oversight. As such, the clinician remains entirely responsible for the decision.

WHERE THE EVIDENCE STANDS: CO-PILOT, NOT AUTOPILOT

One of the few controlled signals in cardiology so far comes from a randomised comparison in which the same general cardiologists managed complex cardiomyopathy cases twice, once alone and once assisted by an LLM. Blinded subspecialists preferred the LLM-assisted assessments (46.7% versus 32.7%) and found fewer clinically significant errors (13.1% versus 24.3%).[3] Crucially, the cardiologist stayed in the loop: the model was a co-pilot, and the autopilot was never tested.

> **The way we ask a question shapes the answer that we get, and most of us were never taught how to ask**

But availability is not integration, and a randomised trial proves it. Physicians given access to a leading LLM scored no better than those using conventional resources (76% versus 74%; adjusted difference: 2 percentage points; 95% CI: -4–8), even though the model alone outscored both groups by 16 points.[4] Worse, when an LLM is confidently wrong, clinicians can be dragged down with it, even AI-literate ones.[5] The lesson is uncomfortable: the hardest skill is perhaps to overrule the model when it is confidently wrong.

THE LEVER IS CONTEXT, AND THE METHOD IS A BRIEFING

The conceptual shift of 2026 is from polishing the wording of a prompt to engineering the context around it. The most interventional-native evidence we have makes the point directly: across 20 complex coronary and structural cases scored by five blinded interventional cardiologists (600 evaluations), the same model scored 6.9/10 by default, 7.8 with web search enabled, and 7.7 with European guidelines supplied in the prompt.[6] Same question, richer context, better-rated answers.

> **With one and the same model, how you brief it determines the answer you get, and thus how safely you can act upon it**

A useful mental model is a brilliant Nobel laureate newly arrived at your hospital, who has read every textbook but has never met your patient, does not know your centre, and does not know your local pathways. That is the LLM. A prompt is, therefore, a clinical briefing; the same kind you give a fellow before a complex case. If you would not brief the laureate that way, do not brief the model that way.

The briefing has a three-pillar structure: frame, ground, and verify, each supported by evidence and undone by a concrete failure (Table 1).

Frame is how you formulate the question. Across 300 simulated hypertension cases, the model-and-prompt configuration shifted decision accuracy from 63% to 91%, and some poorly designed configurations even dragged physicians below their own baseline.[7]

Ground is the patient context you supply. Given structured Heart Team data on 150 patients with severe aortic stenosis, ChatGPT (OpenAI, San Francisco, California, USA) matched the actual Heart Team decision 77% of the time, and 90% for transcatheter implantation.[8]

Verify is the format that lets you check the reasoning. In a randomised study of radiologists, step-by-step LLM explanations improved diagnostic accuracy by 12.2 points compared with no LLM support, and by 7.2 points compared with a standard answer giving no explanation,[9] because a reasoning chain you can see is one you can challenge.

Across all three pillars, the message is the same: with one and the same model, how you brief it determines the answer you get, and thus how safely you can act upon it. These levers also outlast any single model: framing the problem, grounding it in context, and demanding a verifiable format do not date.

A RESEARCH AGENDA FOR THE INTERVENTIONAL COMMUNITY

The honest limitation that runs through almost all of the aforementioned studies is that it is scenario-based, not bedside. Specifically, the interventional comparison[6] used cases, not real patients; the hypertension[7] and Heart Team[8] studies were simulations or retrospective. A pragmatic randomised trial of LLM assistance in real interventional workflows has never been done. Herein lies the opportunity for the European interventional community. A multicentre, pragmatic trial, embedding an LLM in Heart Team preparation or peri-procedural decision support, with a human in the loop, retrieval grounded in local and European guidelines, and a documented audit trail consistent

Table 1: Frame, ground, verify: briefing the co-pilot.

Pillar	What it means/what to ask yourself	A plausible clinical failure when it is skipped
Frame	Define the decision, the patient, and the objective before anything else. What decision, for whom, to what end, and could the premise itself be wrong?	Asking 'which statin dose?' when perhaps no statin is indicated. The model answers the premise it is given; it does not challenge it.
Ground	Supply the relevant context: comorbidities, current drugs, prior imaging, files, and your local and European pathways.	Omitting HIV infection on boosted antiretroviral therapy because it 'felt irrelevant', yet interactions and risk category change the LDL management.
Verify	Demand sources, assumptions, a counter-argument, and a reasoning chain you can actually inspect and challenge.	Acting on a confident answer without checking which guideline, and which version, it rests on; especially dangerous where the recommendation has recently changed.

A practical three-step structure for any clinical query to a large language model.

LDL: low density lipoprotein.

with the EU AI Act, would move us from exam evidence to the bedside evidence we actually need.

Until then, the message is simple. Use these tools for the tasks where they already help: knowledge retrieval, structuring reports, drafting discharge letters, and preparing the Heart Team. Always with review, and as a supervised component of care rather than an autonomous decision-maker.[10] Treat the model as a co-pilot, never an autopilot, and learn to brief it: frame, ground, and verify. The model will change every 6 months; the discipline of asking the right question, and doubting the answer, is what lasts.

References

1. Bedi S et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2025;333(4):319-28.
2. Yao X et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial (EAGLE). Nat Med. 2021;27(5):815-9.
3. O'Sullivan JW et al. A large language model for complex cardiology care. Nat Med. 2026;32(2):616-23.
4. Goh E et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969.
5. Qazi I et al. Automation bias in large language model–assisted diagnostic reasoning among physicians trained in AI literacy: a randomized clinical trial. NEJM AI. 2026;DOI:10.1056/Aloa2501001.
6. Lauretti A et al. Performance of large language models in interventional cardiology: the ILLUMINATE blinded model-comparison study. J Invasive Cardiol. 2026;DOI:10.25270/jic/25.00104.
7. Li Z et al. The effects of multitype prompt engineering for large language models in hypertension treatment decisions. npj Digit Med. 2026;DOI:10.1038/s41746-026-02645-y.
8. Salihu A et al. A study of ChatGPT in facilitating Heart Team decisions on severe aortic stenosis. EuroIntervention. 2024;20(8):e496-503.
9. Spitzer P et al. The effect of medical explanations from large language models on diagnostic accuracy in radiology. NPJ Digit Med. 2026;9(1):333.
10. Desroche LM et al. Artificial intelligence in cardiovascular care for internal medicine: from promising algorithms to useful clinical services. Eur J Intern Med. 2026;DOI:10.1016/j.ejim.2026.106944.