Improving Large Language Models via Heart Team Simulation: A Prompt Design Analysis

Authors: *Dorian Garin,[1] Stéphane Cook,[1] Charlie Ferry,[1] Wesley Bennar,[1] Mario Togni,[1] Pascal Meier,[1] Peter Wenaweser,[1] Serban Puricel,[1] Diego Arroyo[1]

1. Department of Cardiology, University and Hospital Fribourg, Switzerland

*Correspondence to dorian.garin@icloud.com

Disclosure: The authors have declared no conflicts of interest.

Keywords: AI, aortic stenosis, Heart Team, large language model (LLM), prompt design.

Citation: EMJ Int Cardiol. 2025;13[1]:37-38. https://doi.org/10.33590/emjintcardiol/EZQC1200

BACKGROUND

Large language models (LLM) show promise in supporting clinical decision making, yet the influence of prompt design on performance in complex cardiology scenarios remains unclear.[1] This study introduces Forest-of-Thought (FoT), a novel prompting technique that allows the LLM to simulate a multidisciplinary Heart Team discussion. The authors compared FoT with four other common prompting approaches to determine its impact on LLM treatment decision accuracy in severe aortic stenosis.

METHODS

The authors evaluated five prompting techniques with a single LLM (GPT-4o [OpenAI, San Francisco, California, USA], version 2024-05-13): zero-shot (0-shot), Chain-of-Thought (CoT), few-shot Chain-of-Thought (fs-CoT), FoT, and few-shot FoT (fs-FoT). Clinical vignettes were developed for 231 patients with severe aortic stenosis, for whom a Heart Team had recommended transcatheter aortic valve implantation, surgical aortic valve replacement, or medical management. Each vignette was submitted 40 times under each prompting technique, yielding 46,200 total queries. The self-consistency method determined each technique's final recommendation by selecting the most frequently generated answer from the 40 outputs. The primary outcome was the mean correct response rate, defined as agreement with the Heart Team's recommendation. Secondary outcomes included sensitivity, specificity, area under the curve, degree of treatment invasiveness, and relative weighting of clinical variables.

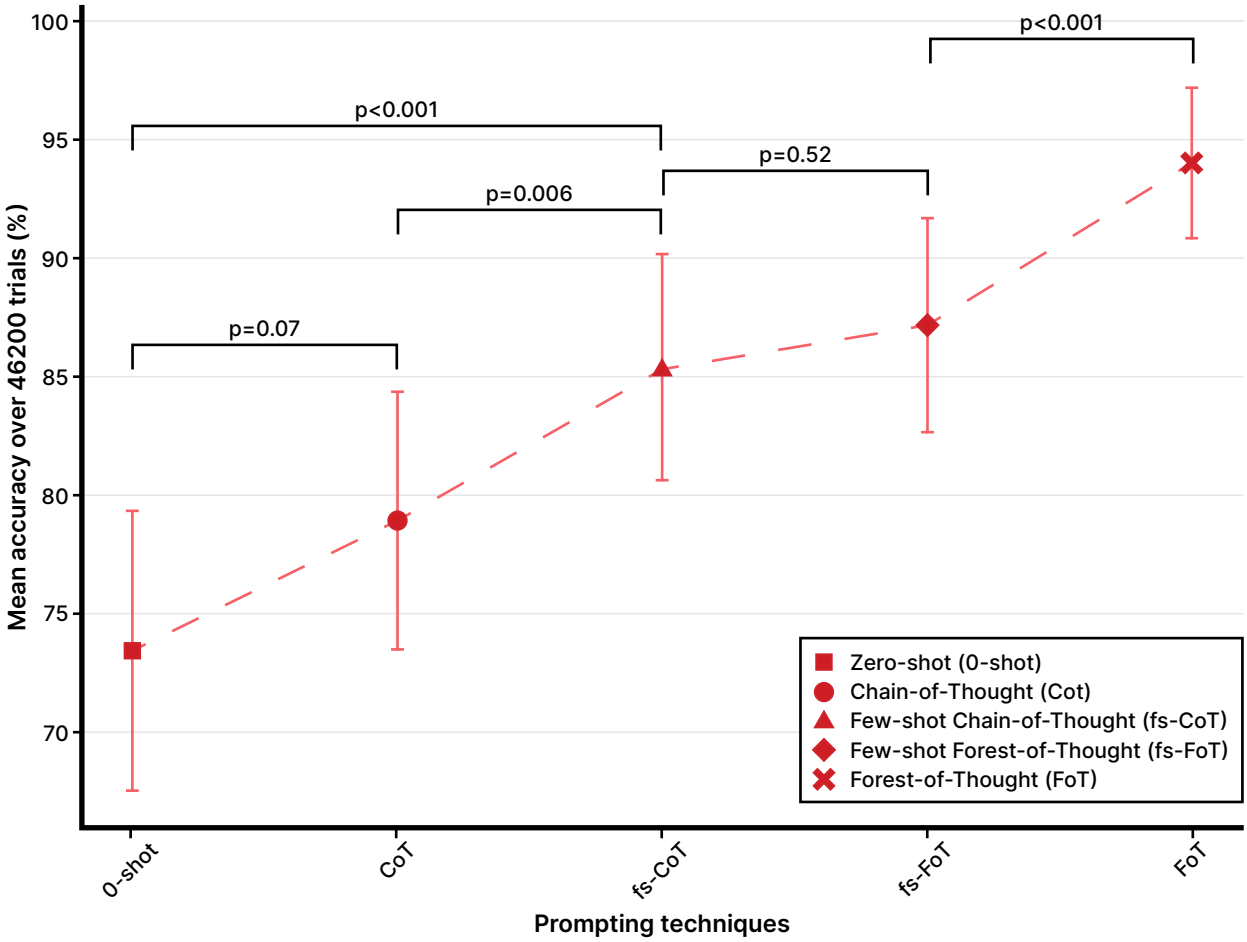
RESULTS

FoT achieved the highest mean correct response rate (94.04%; 95% CI: 90.87–97.21), outperforming all other techniques (fs-FoT: 87.16%; fs-CoT: 85.32%; CoT: 78.89%; 0-shot 73.40%; p<0.001; Figure 1). It also demonstrated the highest sensitivity, specificity, and area under the curve (0.96–0.97). Compared to the Heart Team, the LLM's recommendations were slightly more conservative, favouring less invasive options (mean invasiveness score: –0.0884; 95% CI: -0.1255– -0.0516; p<0.001). Additionally, the model assigned greater weight to non-cardiac comorbidities (Cliff's Delta: –0.231; p=0.04), whereas the Heart Team did not exhibit this preference (Cliff's Delta: +0.12; p=0.75).

CONCLUSION

Prompt design significantly affects LLM performance in managing severe aortic stenosis. The FoT approach, which simulates a multidisciplinary Heart Team, markedly improves decision-making accuracy and produces recommendations closely aligned with expert opinions. However, the LLM demonstrated a tendency towards more conservative treatment plans, underscoring the importance of carefully designed prompts and clinician oversight when deploying LLM-based decision support systems.

Figure 1: Mean accuracy of prompting techniques.



Reference

1. Garin D et al. Improving large language models accuracy via Heart Team simulation: a prompt design analysis. Abstract A62792DG. EuroPCR, 20-23 May, 2025.