HHS Public Access

Author manuscript

Hamlyn Symp Med Robot. Author manuscript; available in PMC 2025 July 21.

Published in final edited form as:

Hamlyn Symp Med Robot. 2025 June ; 2025: 129–130.

Evaluation of Large Language Models to Detect Team Mental Model Misalignments During Cardiac Surgery

Roger D. Dias[1,2], **Vaibhav Unhelkar**[3], **Kenneth Shann**[4], **Allison Weinberg**[5], **Geoff Rance**[6], **Rithy Srey**[7], **Paul O'Gara**[8], **Alexander Shapeton**[8], **Jamie M. Robertson**[1,2], **Paulo Borges**[1,2], **Robson Verly**[1,2], **Sanjana Mendu**[4,8], **Marco A. Zenati**[4,8]

[1]Department of Emergency Medicine, Harvard Medical School

[2]Medical AI & Cognitive Engineering (MAICE) Lab, Mass General Brigham

[3]Department of Computer Science, Rice University

[4]Division of Cardiac Surgery, Mass General Brigham

[5]Department of Cardiopulmonary Sciences, RUSH College of Health Sciences

[6]Cape Cod Healthcare

[7]InvoCirc, Inc

[8]Division of Cardiac Surgery, Veterans Affairs Boston Healthcare System and Medical Robotics and Computer Assisted Surgery (MRCAS) Lab

INTRODUCTION

Prior research has demonstrated that communication breakdowns and mental model (MM) misalignments among cardiac surgery team members may lead to adverse events and compromise patient safety [1-4]. However, the potential of using large language models (LLM) to analyze team communications and detect these misalignments remains unexplored, presenting an opportunity to develop innovative tools for improving team communication and patient outcomes in the operating room (OR). The primary aim of this study was to evaluate the performance of an LLM in generating synthetic team communication exchanges during critical phases of cardiac surgery, demonstrating instances of team MM alignment and misalignment. A second aim was to evaluate the correct classification rates of two different LLMs in detecting team MM alignment/misalignment based on the LLM-generated synthetic communication scripts.

MATERIALS AND METHODS

This study did not involve human research subjects and was conducted entirely with synthetic data generated by LLMs. Based on five previous studies in cardiac surgery [1-5], we selected three critical intraoperative phases in which mental model misalignments between surgeons, anesthesiologists, and perfusionists appear most significant and impactful

rdias@bwh.harvard.edu .

as they relate to patient safety and surgical outcomes. These phases were: *1. Cannulation and Initiation of Cardiopulmonary Bypass (CPB)*; *2. Cross-clamp Application and Cardioplegia Delivery*; and *3. Separation from CPB*.

We used expert-generated hand-crafted prompting [7] using the previous studies mentioned above [1-5] and instructing the *Claude 3.5 Sonnet* LLM to generate realistic scenarios of communication exchanges (i.e., scenario scripts) between a surgeon, anesthesiologist, and perfusionist that demonstrate team MM alignments (four scripts per phase) and misalignments (four scripts per phase), totaling 24 scenarios.

An interprofessional team of cardiac surgery experts (1 surgeon, 1 anesthesiologist, and 4 perfusionists), part of our research group, evaluated each scenario generated by the LLM based on the level of realism (*Not Realistic at All* / *Moderately Realistic* / *Very Realistic*), expert's determination of team MM Alignment/Misalignment, and the level of misalignment (*Low, Moderate, High*) when applicable. Each scenario was evaluated by 3 experts, independently, using an online survey platform (REDCap). The final expert evaluation metrics for each scenario were determined by the majority of the three experts' evaluations. Scenarios that were determined by experts as "*Not Realistic at All*" or that the experts disagreed with the LLM Alignment/Misalignment determination (i.e., discrepant scenarios) were excluded from the LLM performance evaluation.

To evaluate the LLM performance, we used the LLM self-generated prompting method [7] to instruct both *Claude Sonnet 3.5* (*Normal style*) and *GPT-4o* to create their own optimized evaluation prompt to analyze each scenario for Alignment/ Misalignment determination and level of misalignment when applicable. The same five studies [1-5] used for scenario generation were uploaded to the LLMs for prompt creation. After the removal of "*Not Realistic at All*" and discrepant scenarios based on the expert evaluation, the remaining scenarios were analyzed on both LLMs using their respective optimized evaluation prompts, and the correct classification rate was reported for each model. All LLM procedures, including scenario generation, prompt generation, and evaluation, were conducted in separate LLM instances.

RESULTS

A total of 24 synthetic scenarios were generated by Claude 3.5 Sonnet, eight per critical phase, with 50% (four per phase) demonstrating team MM alignment and 50% (four per phase) MM misalignment. Cardiac surgery experts evaluated five (20.8 %) scenarios as **Very Realistic** - *all communications are possible to occur in the OR*, 15 (62.5%) as **Moderately Realistic** - *minor issues that may not reflect real intraoperative communications*, and four (16.7%) as **Not Realistic at All** - *present major misrepresentations*. Experts agreed with the *Claude 3.5 Sonnet* determination of MM Alignment/Misalignment in 23 (95.8%) scenarios and disagreed only in one scenario (discrepant scenario ID: 1.2 *in* Fig. 1).

After excluding the four "*Not Realistic at All*" scenarios and one discrepant scenario (Fig. 1 - *in red*), a total of 19 scenarios (Fig. 1 - *in green*) were used to evaluate the LLMs' performance on detecting team MM alignment/misalignment. Both LLMs *Claude 3.5*

Sonnet and *GPT-4o* correctly classified 19 (100%) of the scenarios as either mental model alignment (10 scenarios) or misalignment (9 scenarios). Among the nine MM misalignment scenarios, all of which were correctly classified as MM misalignment by both LLMs, the level of misalignment (low, moderate, or high) was correctly identified by *Claude 3.5 Sonnet* for three (33.3%) scenarios and *GPT-4o* for five (55.5%) scenarios (Fig. 2).

DISCUSSION

The findings of this study suggest that LLMs can generate team communication scenarios in cardiac surgery with a moderate-to-high degree of realism, accurately demonstrating examples of team MM alignments and misalignments. The ability of both *Claude 3.5 Sonnet* and *GPT-4o* to correctly classify all expert-validated scenarios demonstrates the potential of AI-driven tools in assessing team MM alignment and identifying potential misalignments that could compromise patient safety. However, the discrepancies in the assessment of misalignment levels highlight areas where further refinement and model fine-tuning are needed. Future research should explore how these models perform in real-time intraoperative settings based on real-life team communication exchanges among cardiac surgery team members. LLMs fine-tuned to the cardiac surgery field could be used to drive an intraoperative AI Coach [6] that monitors OR communications in real time and recommends patient safety interventions such as a *safety-pause timeout* when a high MM misalignment is detected. Such a system has the potential to improve surgical team coordination, shared mental models, and patient safety.

Acknowledgment:

This work was supported by the National Institute of Health (NHLBI - R01HL126896 and R01HL157457) and the National Science Foundation (NSF - IIS2310187).

REFERENCES

[1]. Dias RD, Zenati MA, Conboy HM, et al. Dissecting Cardiac Surgery: A Video-based Recall Protocol to Elucidate Team Cognitive Processes in the Operating Room. Ann Surg. 2021;274(2):e181–e186. [PubMed: 31348036]

[2]. Hazlehurst B, McMullen CK, Gorman PN. Distributed cognition in the heart room: how situation awareness arises from coordinated communications during cardiac surgery. J Biomed Inform. 2007;40(5):539–551. [PubMed: 17368112]

[3]. Brown EKH, Harder KA, Apostolidou I, et al. Identifying Variability in Mental Models Within and Between Disciplines Caring for the Cardiac Surgical Patient. Anesth Analg. 2017;125(1):29–37. [PubMed: 28537973]

[4]. Wadhera RK, Parker SH, Burkhart HM, et al. Is the "sterile cockpit" concept applicable to cardiovascular surgery critical intervals or critical events? The impact of protocol-driven communication during cardiopulmonary bypass. J Thorac Cardiovasc Surg. 2010; vol. 139,2:312–9. [PubMed: 20106395]

[5]. Harari R, Dias RD, Salas E, Unhelkar V, Chaspari T, Zenati M. Misalignment of Cognitive Processes within Cardiac Surgery Teams. Hamlyn Symp Med Robot. 2024;16:33–34. [PubMed: 39081305]

[6]. Seo S, Kennedy-Metz LR, Zenati MA, Shah JA, Dias RD, Unhelkar VV. Towards an AI Coach to Infer Team Mental Model Alignment in Healthcare. IEEE Conf Cogn Comput Asp Situat Manag. 2021;2021:39–44. [PubMed: 35253018]

[7]. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. 2023. Available: http://arxiv.org/abs/2311.16452

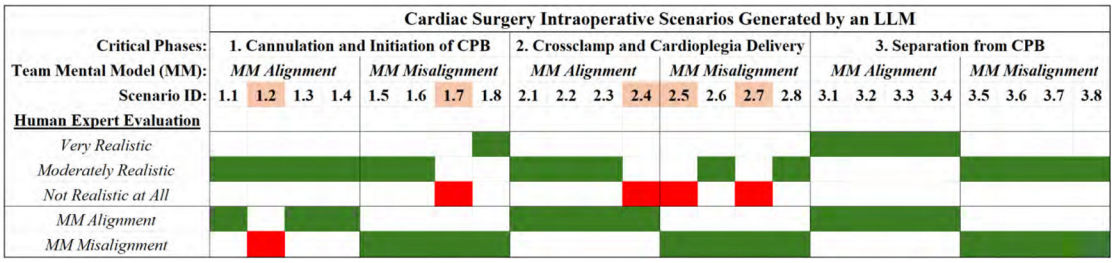

Fig. 1 –. Human Expert Evaluations of LLM-generated Cardiac Surgery Scenarios.

**Red indicates excluded and Green included scenarios based on human expert evalutation.*

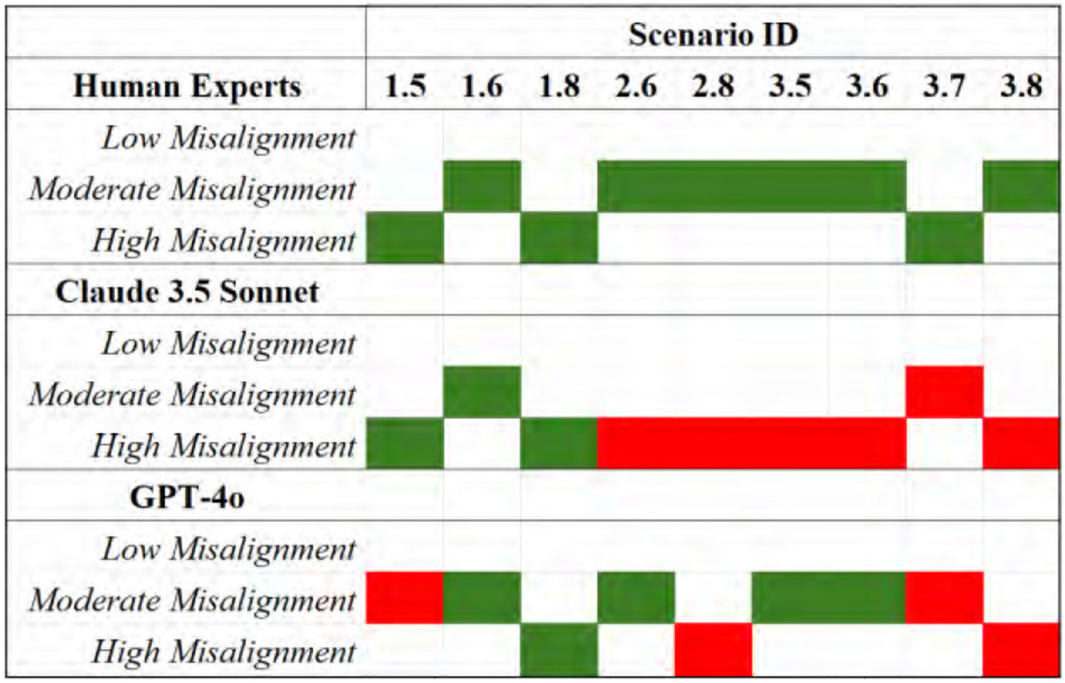

Fig. 2 –. Agreement (green) and Disagreement (red) between Clinical Experts and LLMs on MM misalignment.