Yale University
School of Medicine

ASSESSING THE REASONING CAPABILITIES OF STATE-OF-THE-ART LARGE LANGUAGE MODELS IN COMPLEX SURGICAL DECISION-MAKING FOLLOWING ADMISSION FOR INGUINAL HERNIA REPAIR

Lee Ying, MD, PhD1; Elizabeth Schaefer2; Benlu Wang2; Xinyue Li2; Samuel Butensky, MD1; Emily Flom, MD1; Randal Zhou, MD, FACS, FASMBS1; Andrew Duffy, MD, FACS, FASMBS1; John Morton, MD, FACS, FASMBS1, Karen Gibbs, MD, FACS, FASMBS1; Eric Schneider, PhD1; Arman Cohan, PhD2; Chin Siang Ong, MBBS, PhD, MPH1; 1Yale School of Medicine; 2Yale University

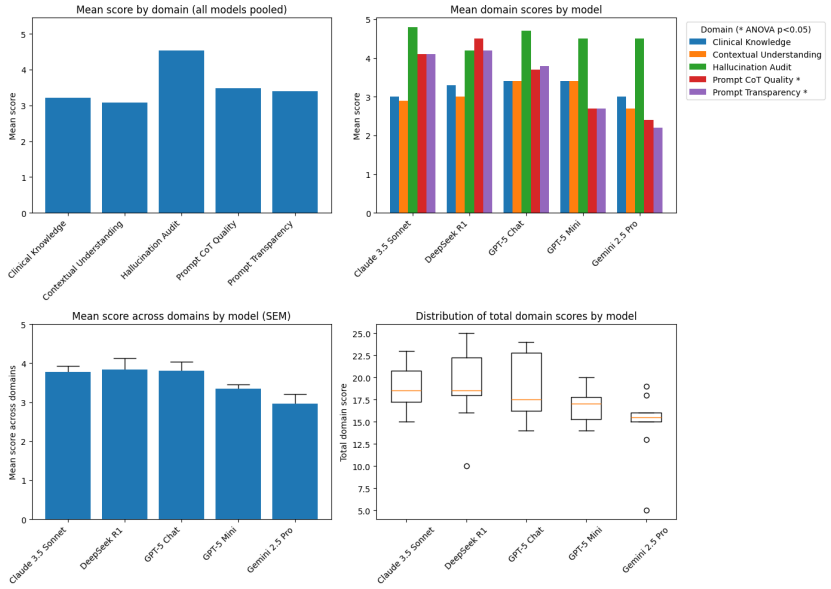
Background

- Large language models (LLMs) may augment surgical decision-making by reducing cognitive burden and supporting adherence to current guidelines
- Inguinal hernia repair is a commonly performed surgical procedure in both emergent and elective settings and requires careful consideration to determine the optimal surgical approach (i.e., open versus laparoscopic, with or without mesh).
- We evaluated the clinical reasoning capabilities of state-of-the-art models in patients admitted following inguinal hernia repair.

Methods

- Five advanced large language models were assessed: Anthropic Claude 3.5 Sonnet, DeepSeek R1, Google Gemini 2.5 Pro, OpenAI GPT-5 Chat, and OpenAI GPT-5 Mini.
- Models were prompted using a standardized case-specific clinical reasoning schema
- Outputs were evaluated across five domains using a 5-point scale with pre-defined anchor points: Clinical Knowledge, Contextual Understanding, Hallucination Audit, Prompt Chain-of-Thought (CoT) Quality, and Prompt Transparency.

Figures



Discussion

- The models accurately summarized a wide range of patient contexts, from healthy patients undergoing elective laparoscopic bilateral inguinal hernia repair with mesh to patients with cirrhosis presenting emergently with incarcerated bowel undergoing primary repair without mesh.
- Between the models, there were no significant differences in Clinical Knowledge, Contextual Understanding, or Hallucination Audit.
- Significant differences were observed for Prompt CoT Quality ($p<0.01$) and Prompt Transparency ($p<0.01$). Claude 3.5 Sonnet demonstrated the highest overall reasoning performance.
- Pairwise comparisons showed Claude 3.5 Sonnet outperformed Gemini 2.5 Pro ($p=0.01$) and GPT-5 Mini ($p=0.03$). GPT-5 Chat and DeepSeek R1 also outperformed Gemini 2.5 Pro ($p=0.02$ and $p=0.03$, respectively).
- Within each model, Hallucination Audit scores were higher than Clinical Knowledge and Contextual Understanding ($p<0.01$ for all models).

Conclusion

- With their low hallucination rates, all models were able to evaluate patient notes and provide succinct, accurate summaries, which may have utility in creating new hand-off tools to ensure safe transitions in patient care.