Feasibility of utilizing AI to enhance audiovisual surgical learning: proof-of-concept

Julie Hong MD, Victoria Ende, MD, Alyssa Greenwood-Francis MD, Amber Chen-Goodspeed, MD; Omar Mahboob MD, Catherine Qiu MD, Stephen Merola MD, Steven Y Chao, MD, FACS,

NewYork-Presbyterian Queens

Weill Cornell Medicine

BACKGROUND

General use-large language models (GLLM) are transforming medical education by serving as a customizable and interactive learning tool.

Two index cases for surgical training, cholecystectomies and appendectomies, teach residents how to visualize key structures for safe surgery.

This proof-of-concept study aimed to evaluate how well proprietary GLLMs can identify important structures in two common laparoscopic general surgeries compared to early-career surgical residents.

METHODS

- Screenshots from videos of laparoscopic cholecystectomies and appendectomies were selected from the SAGES video library depicting the critical view of safety for each surgery
- To be included (when visible):
  - Cholecystectomy: gallbladder, liver, cystic duct, cystic artery, and cystic plate.
  - Appendectomy: appendix, cecum, taenia coli, mesoappendix, and terminal ileum
- Screenshots were uploaded to an untrained model of ChatGPT-5,and Gemini 3 Pro using the “thinking” method in a temporary chat session with prompts to label the image with transparent colored overlays as a Matplotlib Chart
- These were then independently graded by senior and intern surgical residents in a binary fashion for each item labeled. The image quality and difficulty of the view was also rated (1-5)

RESULTS

17 appendectomies and 24 cholecystectomies

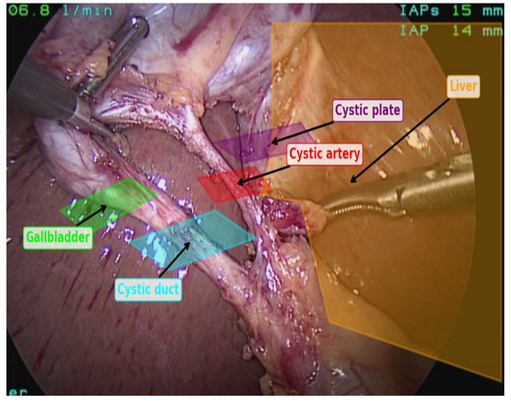
Structures identified by GLLM
- For cholecystectomy: average **2.5 structures**. liver most common 20/24, gallbladder 10/24.
- Appendectomy: **1.6 structures**. cecum 12/17

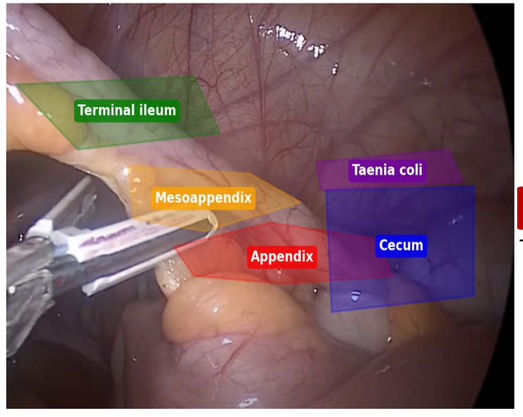
Interns: 410 structures identified. 88% accuracy with 96% sensitivity 45% specificity

ChatGPT: 205 structures identified had 38% accuracy, 100% sensitivity, and 0% specificity

Gemini: 205 structures 55% accuracy, 88% sensitivity, and 15% specificity.

Overall images were of medium quality (2.1 chole, 2.3 appy) and medium difficulty (2.3 chole, 2 appy).





Examples of malipot charts generated by Chat GPT for cholecystectomy and appendectomy (good and bad)

DISCUSSION

Early-career surgical residents outperformed the current iterations of widely available general-use large language models.

Despite promise, GLLMs are not yet a reliable adjunct for supplementing the visual component of surgical learning. While the GLLM in this study was sometimes able to accurately identify the most basic of anatomic structures, it was unable to appreciate the core structures to cholecystectomies and appendectomies, making it ineffective for learners seeking to improve their understanding of key surgical anatomy.

REFERENCES

Bai, L., et al (2023). Surgical-VQLA: Transformer with gated vision-language embedding for Visual Question Localized-answering in Robotic Surgery. In *arXiv*. http://arxiv.org/abs/2305.11692

Brunt, L. M. et al and the Prevention of Bile Duct Injury Consensus Work Group. (2020). Safe cholecystectomy multi-society practice guideline and state of the art consensus conference on prevention of bile duct injury duringcholecystectomy. *Annals of Surgery*, *272*(1), 3–23. https://doi.org/10.1097/SLA.0000000000003791

Moris, D. et al (2021). Diagnosis and management of acute appendicitis in adults: A review: A review. *JAMA: The Journal of the American Medical Association*, *326*(22), 2299–2311. https://doi.org/10.1001/jama.2021.20502

Rau, A.,et al (2025). Systematic evaluation of large Vision-Language Models for surgical artificial intelligence. In *arXiv [cs.CV]*. http://arxiv.org/abs/2504.02799

ACKNOWLEDGEMENTS

Thank you to New York Presbyterian Queens surgery team and SAGES public video library

For more information, please contact:
Victoria Ende mcp9026@nyp.org.