ACS 2026 Surgeons and Engineers: A Dialogue on Surgical Simulation

O-09

Research Abstracts

Evaluating Surgical Reasoning in Multi-Agent Artificial Intelligence Systems using the ENTRUST Learning and Assessment Platform

Benjamin Liu; Kiran Nijjer; Raghav Thallapragada; Derek Jiu; Adnan Ahmed; Jason Tsai; Jeffrey K. Jopling, MD; Edward Franklin Melcer; Dana Tsing-Yip Lin, MD, FACS; and Cara A. Liebert, MD, FACS

Stanford University School of Medicine, Palo Alto, CA; The Johns Hopkins University School of Medicine, Baltimore, MD; University of California, Santa Cruz, CA

Introduction: Entrustable Professional Activities (EPAs) in surgery provide a framework for assessing complex, independent clinical decision-making. ENTRUST is a virtual patient simulation platform that assesses clinical reasoning and decision-making across comprehensive patient case vignettes. We developed and tested a multi-agent AI system for navigating ENTRUST and evaluated the system using case-specific MCQs.

Methods: Approximately 200 cases were selected from the ENTRUST platform across a diverse array of surgical sub-specialties. Each case included a 1) comprehensive patient history, 2) multifaceted menu of ~200 diagnostic and management actions encompassing physical exams, imaging orders, laboratory tests, and interventional tests, and 3) multiple choice questions (MCQs) querying applied surgical knowledge. The simulation environment was constructed with multi-turn interactions between separate clinical actors, modeled by instances of LLM Gemma-12B with distinct context windows. Specifically, for each case, the primary AI agent engaged in ten turns of dialogue with a patient agent, four stages of actions with separate diagnostic agents, and five turns of consultation with a specialist agent before taking the final MCQ exam.

Results: Across all cases, our system achieved accuracies of 58%, 63%, and 60% on composite, contextual, and non-contextual MCQs, respectively, which is comparable with medical student proficiency. On average, 12 diagnostic orders were taken in each case with a 50-30-20 spread between indicated, harmful, and neutral clinical action grades, respectively. Analysis stratified across surgical sub-specialties revealed consistent performances with scores above 0.50 in all MCQ categories and top scores above 0.71 for urology cases.

Conclusions: Our study highlights the current limitations of multi-agent AI systems in navigating complex surgical decision-making. Future work will focus on improving multi-agent systems so that they may serve as a powerful and reliable complement to surgical training and practice.

System performance across virtual patient case scenarios stratified by surgical sub-specialty.

Specialty	No. Cases	Composite MCQ Accuracy	Contextual MCQ Accuracy	Contained MCQ Accuracy
General Surgery	166	0.596	0.648	0.601
Vascular Surgery	26	0.613	0.696	0.583
Gynecology and Obstetrics	23	0.512	0.701	0.574
Otolaryngology Surgery	20	0.548	0.572	0.609
Orthopaedic Surgery	19	0.580	0.623	0.567
Gynecologic Oncology	19	0.553	0.679	0.510
Cardiothoracic Surgery	18	0.598	0.560	0.699
Pediatric Surgery	17	0.638	0.713	0.549
Urology	15	0.747	0.796	0.719