



Proposal Defense
Doctor of Philosophy in information Science

“From Fragments to Framework: Benchmarking the Medical Competency of Large Language Models in the Context of Emergency Medicine Practice” by Zhimeng Luo

Date: October 28, 2026

Time: 11:30 a.m. to 1:30 p.m.

Place: Room 828, Information Sciences Building, 135 N.
Bellefield Ave, Pittsburgh PA 15260

Committee:

- Daqing He, Professor and Advisor, Department of Informatics and Networked Systems, School of Computing and Information
- Adam Frisch, Assistant Professor, Department of Emergency Medicine, University of Pittsburgh
- Peter Brusilovsky, Professor, Department of Informatics and Networked Systems, School of Computing and Information
- Na Du, Assistant Professor, Department of Informatics and Networked Systems, School of Computing and Information

Abstract:

Medical benchmarks for large language models (LLMs) measure fragments of practice: answering exam questions, writing clinical notes, advising a patient in conversation. However, medical competency in actual healthcare practice is defined by a comprehensive framework comprising six competencies (medical knowledge, patient care, communication, professionalism, systems-based practice, and practice-based learning), according to the Accreditation Council for Graduate Medical Education. In emergency medicine (EM), practice is defined as managing a patient across the whole encounter, following the model of EM practice by the American Board of Emergency Medicine (ABEM). Within this encounter, the competencies are exercised together, but fragments observe only part of it. Fragment scores therefore capture medical competency only in part, without making explicit which competencies they reflect, at what level, or which remain unassessed.

The goal of this dissertation is to evaluate LLMs for medical competency as the profession defines it. To reach this goal, the dissertation begins with an adapted version of the EM competency framework, because evaluating an LLM differs from assessing a physician in two ways. First, physicians are assessed by observing their practice over years, but an LLM must be evaluated before it has any clinical practice. Second, the profession's criteria describe what physicians show when observed in person, not what an LLM's written responses must contain. The adapted framework keeps the profession's competencies and their definitions. For each competency, it states what an LLM's response must show at the profession's five levels, from novice to expert.

The dissertation has three research questions. RQ1 asks how the profession's framework is adapted for LLMs, and at which levels existing benchmarks cover each competency. The adaptation also marks what text cannot show, such as hands-on procedures. RQ1 thus yields the adapted framework and a map of existing benchmarks against it, which together set what RQ2's benchmark must cover. RQ2 examines how a benchmark is built to cover the adapted



University of
Pittsburgh

School of Computing
and Information

framework. Its cases are simulated whole encounters, following the format of the clinical care cases in ABEM's certifying exam and extended toward competencies those cases do not elicit. Each encounter requires the model to manage the patient proactively, with the case unfolding in response to its information requests and clinical decisions. The model's actions are evaluated with rubrics derived from the adapted framework, yielding scores for each competency at the profession's levels, which RQ3 uses to rank models. Finally, RQ3 asks whether evaluating models on this benchmark by the profession's competencies changes their ranking relative to existing benchmarks, and which ranking physicians' own judgment agrees with. Emergency physicians rate the models' outputs to supply that judgment.

Together, the work moves LLM evaluation from benchmarks that measure fragments of practice to a benchmark built to the adapted competency framework, so that its scores are explicitly interpretable in terms of medical competencies. The claims concern rankings and physician agreement, not deployment or patient outcomes.