Wharton University of Pennsylvania | William and Phyllis MACK INSTITUTE for INNOVATION MANAGEMENT

Evaluating LLMs for Dynamic, Multimodal Clinical Decision-Making

by Lennart Meincke, Christian Terwiesch, and Arnd Huchzermeier [1]

January 2026

The Wharton School, University of Pennsylvania

Abstract

In this study, we evaluate the ability of a multimodal LLM to autonomously manage a virtual end-to-end diagnostic workflow. We test an autonomous agent in a high-fidelity medical simulation across four acute care scenarios. We compare the AI's policy against over 14,000 simulation runs by medical students and an expert emergency room physician benchmark. We find that (1) a multimodal LLM can function as a competent virtual physician, successfully stabilizing patients and solving complex cases that require interpreting text, audio, and imaging in real time, closely mirroring most of the actions of the expert physician; (2) the AI agent matches or exceeds medical students in case completion rates and secondary outcomes such as time and diagnostic accuracy, though it engages in less patient communication than both students and the expert physician; and (3) the agent's evolving diagnostic beliefs exhibit value-of-information properties—front-loading high-yield tests, experiencing diminishing belief revisions over time, and producing well-calibrated confidence estimates. These findings suggest that LLMs can orchestrate complex clinical workflows rather than merely execute isolated tasks. They also offer design implications for human-AI collaboration and diagnostic pathway design in low-resource and time-critical healthcare operations.

Watch a supplemental video: https://tinyurl.com/llm-dynamic-medical-eval

[1] **Lennart Meincke**: WHU – Otto Beisheim School of Management; Mack Institute for Innovation Management, The Wharton School, University of Pennsylvania; lennart@sas.upenn.edu; **Christian Terwiesch**: The Wharton School, University of Pennsylvania; Perelman School of Medicine, University of Pennsylvania; Mack Institute for Innovation Management, The Wharton School, University of Pennsylvania; terwiesch@wharton.upenn.edu; **Arnd Huchzermeier**: WHU – Otto Beisheim School of Management; arnd.huchzermeier@whu.edu

Introduction

Expert systems and machine learning models have long been used to support clinical decision-making. For example, many ICUs have implemented sepsis warning systems which employ predictive models trained on past ICU patients that alert the care team when a patient is likely to transition to a septic state. Similarly, image recognition models trained on libraries of X-rays can predict the likelihood of a new X-ray showing signs of cancer. A common characteristic across such early models is that they were built for very specific tasks and can therefore only be applied to a narrow set of clinical use cases.

With the arrival and rapid improvement of large language models (LLM) pre-trained on vast corpora of data, Artificial Intelligence (AI) now has the potential to move beyond specialized models and have one single model become universally applicable across a wide range of domains. A recent review by Su et al. (2025) identified 95 articles that describe the usage of standard LLM-based models for diagnostics in various clinical settings, the most common being radiology, psychiatry, and neurology. Despite the identification of some biases (e.g., gender or ethnicity), the overwhelming number of these studies demonstrate a strong performance of the LLMs.

As scholars of Operations Management, we observe and attempt to overcome two common limitations across this rapidly growing body of research. First, with few exceptions, these prior evaluations of LLMs in clinical decision-making focus on static task-level assessments rather than dynamic workflows—for example generating a single diagnosis from a textual medical vignette (e.g. Ayers et al. (2023) or diagnosing an X-ray and turning it into a textual description (Huang et al. 2025). Real clinical decision-making, in contrast, is dynamic. Be it in an emergency department or a routine visit for chronic care, in the context of clinical decision-making, not all information about a patient is available at the beginning of the patient-provider encounter. Rather, providers need to take actions to obtain new information by talking to patients, by examining them, or by ordering tests (e.g., labs or imaging). The resulting new information has the potential to improve the accuracy of diagnosis and guiding the right treatment decision for the patient. Yet the new information also comes at a cost. This cost might be an opportunity cost of time. A patient with an epidural hematoma (typically a post traumatic brain injury) likely requires a rapid surgical intervention (in the form of a craniotomy) and hence the provider would be ill advised to wait 30 minutes for an ultrasound of the heart. The cost of the new information also might be a financial one.

Second, clinicians must integrate information from diverse sources including text, images, sounds, and patient-reported symptoms. The data available for the decision-making process thus typically is multimodal. Most of the prior work evaluating LLMs and their ability to engage in clinical decision-making have relied on text as the input to the LLM. In the context of radiology, prior studies have also used images produced by X-rays, CT-scans, or MRIs. In clinical practice, however, information is often represented in many other modalities beyond text and images, including sound (e.g. listening to the

sounds of the lungs obtained via a stethoscope), touch (e.g. palpating the liver to detect a hepatomegaly), and smell (e.g. detecting a ketone odor in diabetic patients). Whether modern multimodal LLMs can weave together diverse data streams—interpreting a chest X-ray, listening to lung sounds, reading lab values—within a coherent diagnostic workflow remains largely untested.

Given these two limitations of prior work, our primary research aim is to evaluate the efficacy of an LLM to engage in clinical reasoning in a research setting requiring sequential (dynamic) decisions based on multimodal information. Rather than evaluating performance on individual tasks, we study the complete diagnostic workflow.

The diagnostic process can be thought of as a dynamic information-gathering problem under uncertainty: the decision maker must repeatedly trade off the cost and delay of additional information against its value for improving downstream treatment decisions. Drawing on previous work on diagnostic pathways, test adoption, and AI integration (Dai and Singh 2025; Hopp et al. 2018; Shi et al. 2021; Somanchi et al. 2022), we adopt this perspective to compare human and AI policies on dimensions including service time, diagnostic cost, and process quality.

Empirically validating the capabilities of an LLM in a real-world situation poses major quality and ethical challenges. This is likely the reason why almost all prior studies that have evaluated the LLMs in decision-making settings were retrospective (i.e., the real clinical decisions have long been made) making it hard to reproduce the dynamic decision-making situation faced by the provider. Our evaluation takes a different approach. We evaluate the LLM by putting it into the role of the provider in a clinical simulation model. Such simulation models have recently emerged as high-fidelity environments mimicking real clinical settings (Diaz-Navarro et al. 2024) and are used to train and test current and future providers. Specifically, we use the simulation package BodyInteract that provides a library of clinical settings in a virtual reality format and allows the decision maker to take actions that would be available to them in a real clinical setting. We developed a test environment that connects an off-the-shelf LLM (Gemini Pro 2.5) with the BodyInteract simulation. This allows us to evaluate an AI agent based on the LLM in the same manner as medical students in a class or providers at their time of (re)certification are evaluated.

Beyond assessing the quality of the AI agent's decision, we also want to compare its behavior with how humans behave when faced with the same clinical decision problems. Our secondary research aim is thus to compare the decisions and actions of the AI agent with that of human decision makers. Towards that aim we empirically look at a large number of (human) users of the simulation and compare their decisions and actions with what our AI agent does. In particular, we compare humans and AI agents in their ability to stabilize and appropriately treat a patient ("solving the case") and secondary outcomes

such as the time taken, the associated costs of care, and the style of practice. This novel setup allows us to establish the following three contributions:

- **Medical competence at the workflow level.** First, we show that a modern multimodal LLM can function as an autonomous virtual physician in high-fidelity clinical simulations, stabilizing virtual patients and succeeding in solving challenging medical cases that are used for physician certification and training.
- **End-to-end, multimodal orchestration.** Second, we demonstrate that a multimodal LLM can select the actions through the same interface as human users, using screenshots, audio and case text. It can order and interpret tests and take all other actions needed to handle the workflow end-to-end. We benchmark this end-to-end policy against an emergency physician and medical students, showing that it closely mimics the expert and often times outperforms the medical students in the proportion of cases solved and along the secondary outcomes.
- **Calibrated reasoning and value-of-information behavior.** Third, we "open the black box" of the LLM's decision process by logging the agent's evolving diagnostic beliefs, showing that its information-gathering strategy exhibits properties consistent with value-of-information reasoning and approximately calibrated confidence, a finding that contrasts with recent work documenting LLM miscalibration in other settings (Geng et al. 2024).

Our findings have important managerial implications. As we imagine the future role of AI in medicine, we need to evaluate what current models can and cannot do. For that, it is important to understand that physicians do more than completing a specific set of tasks. Using Clayton Christensen's "Jobs to be Done" framework (see Christensen et al. 2005), the job to be done by a physician is not to interpret an X-ray, but to stabilize and heal the patient. A narrow and specialized AI model would be sufficient to interpret a given X-ray. However, the harder challenge is knowing *when* to order which test and *what* to do with the results—orchestrating a workflow of tasks rather than executing any single one. Our finding that LLMs can manage this orchestration suggests their usefulness extends beyond the task level to the workflow level. The evidence for calibrated beliefs provides hope that future models may recognize when to act autonomously and when to defer to human judgment.

The remainder of this article is organized as follows. Study 1 introduces a simple at-home hypoglycemia case. Study 1a asks whether an autonomous LLM-based agent can navigate the interface in real time and stabilize the patient, while Study 1b compares its timing, action sequence, and style of practice to that of medical students and a medical expert. Study 2 turns to a more demanding emergency room case that requires ordering, interpreting, and acting on multiple diagnostic tests. Study 2a examines whether the agent can successfully solve this richer, higher-stakes scenario, and Study 2b compares its diagnostic and treatment strategy to human decision makers. Study 3 generalizes this analysis to a bundle of three emergency room cases of similar complexity, allowing us to assess how

robust the observed patterns are across different clinical problems. Study 4 returns to these emergency room cases to examine the process the AI uses to solve the case including its testing strategy (its process of uncertainty resolution) and its belief updating.

Theoretical Framework

Prior Research

Similar to what has been reported in Operations Management (Terwiesch 2023), LLM's have shown impressive skills on academic medical exams. For example, Chen et al. (2023) report that out of the 509 eligible questions in the BoardVitals test bank of neurology questions, ChatGPT correctly answered 335 questions (65.8%) on the first attempt/iteration and 383 (75.3%) over three attempts/iterations, scoring at approximately the 26th and 50th percentiles of human test takers, respectively. Eisemann et al. (2025) reported that AI-supported radiologists achieved significantly higher cancer detection rates compared to control groups consisting of two human experts.

Almost all the existing medical studies can be classified as "static" or "one-shot". A clinical case study is presented to the decision maker (AI or human), who then generates a diagnosis based on the data.

Two recent exceptions to this static approach are Tu et al. (2025) and Nori et al. (2025). Tu et al. (2025) present AMIE (Articulate Medical Intelligence Explorer), an LLM-based system optimized for conducting a diagnostic dialogue between patient and provider. This model takes the role of a provider who engages in a back-and-forth discussion with a patient and shows improved diagnostic accuracy (as judged by physicians) and better conversation quality (as judged by patients). Closest to our work, a team of Microsoft researchers (Nori et al. 2025) use 304 diagnostically challenging cases from the *New England Journal of Medicine* to develop the Sequential Diagnosis Benchmark which tests the ability of a decision-maker to iteratively request additional information from a gatekeeper model that only reveals information when explicitly queried thereby simulating the process of requesting tests. This study also considers the financial costs of information gathering, which allows for a cost-quality analysis. The authors show that a combination of human physicians and off-the-shelf LLMs can improve diagnostic accuracy while also reducing costs. While both Tu et al. and Nori et al. allow for sequential (multi-round) iterations of their AI model, both of them implicitly assume that (1) There is no penalty for delays in information processing (the decision maker, human or AI model, is really under no time pressure and can de facto take endless time to contemplate their next steps) (2) The patient is stable (i.e., the patient is neither getting better nor getting worse with time) (3) There are no lead times associated with tests (test results are always available in the next 'period').

The Operations Management literature has a long history of analyzing the dynamics of decision-making problems in general and diagnostic processes in particular. Laker et al. (2018) demonstrate that information overload and framing effects can degrade both diagnostic quality and timeliness, highlighting the cognitive costs of acquiring too much data. Somanchi et al. (2022) analyze the trade-off between acting early on limited information and waiting for richer data in emergency department admission prediction, showing that the optimal stopping point depends on case acuity and downstream capacity. Shi et al. (2021) develop a framework for evaluating new diagnostic tests that accounts for both clinical accuracy and operational value—emphasizing that tests affect not only patient outcomes but also service times, congestion, and resource utilization.

As far as AI and machine learning is concerned, prior work has shown how machine learning can improve risk prediction and treatment selection while also affecting patient flow, capacity utilization, and cost (Feng and Shanthikumar 2022; Guha and Kumar 2018; Hopp et al. 2018). This stream reframes diagnosis not merely as a prediction task but as a sequence of information-gathering and treatment decisions made under resource constraints.

More recently, (Dai and Singh 2025) study how AI should be positioned within diagnostic pathways—as a gatekeeper, a second opinion, or not at all—finding that the optimal role depends on case risk and that abstaining from AI can dominate for intermediate-risk patients. Related work examines the conditions under which clinicians adopt AI recommendations, emphasizing trust, workflow integration, and the design of human-AI collaboration (Dai and Tayur 2022; Kyung and Kwon 2022).

Clinical Decision-Making as POMDP

Our analysis relaxes these assumptions by treating diagnosis as a partially observable Markov decision process (POMDP) in continuous time. Bravo et al. (2019) use a similar framework for search-and-rescue operations, balancing information-gathering flights against immediate search actions—analogous to providers balancing diagnostic tests against treatment. Bensoussan et al. (2020) model dynamic maintenance via an MDP that trades off upgrade costs against failure risk as a system deteriorates. Xia (2020) and Xia et al. (2023) extend this to risk-sensitive formulations incorporating outcome variability. The novel feature of our analysis is bringing this framework to bear on an empirical comparison of human and AI diagnostic policies in a realistic simulation environment by treating diagnostic decision-making as a dynamic, multimodal, partially observable decision problem. Rather than evaluating static one-shot diagnoses, we study a continuous-time process in which the decision maker alternates between information-gathering and treatment actions, and where both information and treatment have non-trivial lead times. In the case when an LLM takes the decision, such time delays capture the response time of the LLM, which tends to be small, but certainly not zero. In the case of a human decision maker, this captures the time of cognitive processing.

Conceptually, the underlying patient trajectory can be described by a latent state St that captures the true physiological condition of the patient at time t (e.g., “severe pneumonia with hypoxia,” “hypoglycemic,” “stabilized post-treatment”), including unobserved comorbidities and disease severity. The decision maker cannot observe St directly but instead receives partial observations Ot such as vital signs, lab and imaging results, and patient verbal responses. At each decision point, they choose an action At from a finite set that includes information-gathering actions (dialogue, physical examination, diagnostic tests), treatment actions (medications, oxygen, fluids, calls to specialists or emergency services), and a terminal “stop and diagnose” action. The simulation engine then updates the latent state and generates new observations according to a Markovian transition kernel.

Formally, the encounter can be summarized by a tuple (S, A, R) where S is the (unobserved) state space, A the set of available actions, and R the reward function. Because the true state is hidden, the decision maker can be viewed as maintaining a belief state bt, a probability distribution over S induced by the history of past actions and observations. A policy $\boldsymbol{\pi}$ maps observable histories—or equivalently, belief states—into actions. In our setting, $\boldsymbol{\pi}$AI denotes the stochastic policy implemented by the LLM-based agent interacting with the simulation through our multimodal perception and control harness, while $\boldsymbol{\pi}$H denotes the policies implemented by medical students and the expert physician.

The reward function R combines terminal and running components. At the end of each case, decision makers receive a large positive reward for successful stabilization and penalties for timeout or critical failure. We further evaluate diagnostic accuracy—whether the final diagnosis matches the simulator's reference—as a terminal outcome. During the case running costs accrue along three dimensions: (i) the opportunity cost of time, reflecting delayed treatment and tied-up staff and equipment capacity; (ii) financial costs for diagnostic tests and procedures; and (iii) process quality, measured as the share of recommended patient engagement actions (e.g., talking to the patient) that were taken. Although these actions are not strictly required to complete a case, they are integral to real clinical care, and each simulated case provides a list of reference actions. Rather than imposing a particular weighting on our dimensions, we report them separately and view AI and human policies as occupying different points on a cost-quality frontier.

In principle, an optimal POMDP policy balances these elements by acquiring just enough information to support effective treatment while avoiding unnecessary delay and diagnostic expense. This leads to a cost-quality and efficiency-thoroughness frontier similar to value-of-information models in OM (Bavafa et al. 2021; Shi et al. 2021; Somanchi et al. 2022). In practice, we do not attempt to solve or estimate the underlying POMDP. Instead, we use this framework as a conceptual lens to interpret observed policies. Our empirical analysis compares π_{AI} and π_H on the same simulated decision problems, asking: how do AI and human policies differ in their trade-offs among completion, time, accuracy, communication, and cost; and to what extent does the AI’s information-gathering behavior resemble a

value-of-information policy? The objective of the decision maker is to select a policy π that maximizes the expected cumulative reward. Rather than solving for an optimal policy, our empirical analysis explores how πAI and πH differ in how they value information and manage costs.

We operationalize this framework in four studies that progressively increase case complexity and analytic depth. Study 1 establishes feasibility of π_{AI} in a simple at-home hypoglycemia case with a constrained state and action space. Study 2 turns to a more complex emergency-room pneumonia case that requires ordering, interpreting, and acting on multiple diagnostic tests. Study 3 generalizes the analysis to a bundle of three complex emergency-room cases and considers communication behavior and diagnostic test expenditures as components of R, allowing us to study time-cost-quality trade-offs across cases. Study 4 "opens the black box" by logging the AI agent's evolving diagnostic beliefs and examining whether its information-gathering strategy exhibits patterns consistent with value-of-information reasoning and approximately calibrated confidence.

Research Setting

This research has been approved by the Institutional Review Board at the University of Pennsylvania Protocol under Protocol #859387. All studies were conducted using BodyInteract, a proprietary virtual patient simulation (www.bodyinteract.com). For the present research, we focused on four pre-existing cases from the BodyInteract library. To ensure that they reflect current clinical standards and represent realistic acute care scenarios, we recruited an emergency medicine physician (from here on referred to as "the expert") who independently reviewed and vetted all selected cases prior to data collection. The final case set comprised three emergency room scenarios and one at-home scenario. Together, they cover multiple medical specialties (respiratory, neurology, cardiology, and endocrinology) and patient demographics (ages 30-75 years, both male and female patients). The index conditions include pneumonia, ischemic stroke, congestive heart failure, and hypoglycemia. In each case, users must conduct a systematic assessment, make diagnostic decisions, and implement appropriate interventions under explicit time pressure. The default time limit for all four cases is 20 minutes. Detailed case descriptions are provided in Appendix A.

In this paper we deliberately treat the AI agent as an off-the-shelf, static model rather than a finely tuned, bespoke decision support tool. Concretely, the agent is a single multimodal LLM (Gemini Pro 2.5) that we access via an API and connect to the BodyInteract simulator through a lightweight "harness" that clicks buttons and retrieves screenshots, audio, and text. We deliberately avoid mimicking physician behavior or optimizing for any particular objective function; instead, we observe the emergent policy the model adopts when left to its own devices. We argue that establishing what a capable but unconstrained AI agent can do is a necessary first step before exploring how its behavior might be shaped through incentives, guardrails, or human-AI collaboration protocols. The technical

architecture is modular and could readily accommodate such extensions; for instance, one could penalize diagnostic expenditures, require minimum communication thresholds, or integrate real-time physician oversight. We view the present study as a baseline that future work—by ourselves and others—can build upon.

Simulation Environment

The simulation is a UnityEngine-based application that features several different 3D environments (emergency room, consultation room, street, home) depending on the scenario. The main user interface (UI) combines (i) a realistic 3D rendering of the patient and environment, and (ii) a set of menu buttons and submenus that allow users to perform actions such as talking to the patient, ordering tests, administering treatments, and monitoring vital signs and (iii) a variety of data, including X-rays and recent vital signs (see **Fig. 1**).

Fig. 1 The main user interface for emergency-room based cases.

Notes. The agent has already turned on multiple monitoring options (e.g., heart rate and blood pressure) and has requested and received a head CT scan result.

Each case includes a structured case briefing that presents initial information about the patient (e.g., age, sex, weight, presenting symptoms) at the start of the simulation and remains accessible throughout. As new information is obtained (e.g., test results, vital signs), it is displayed on the screen. Users interact with the patient primarily via interface buttons that spawn nested menus (e.g.,

"Dialogues," "Monitoring," "Tests," "Treatment," "Calls"), but certain operations—such as using a stethoscope or palpating extremities—require direct interaction with the 3D patient model.

This simulation environment and evaluation logic (success, timeout, critical failure) underlies all four studies; case-specific details are described in the respective study sections.

AI Agent and Technical Workflow

We implemented an autonomous AI agent based on Google's Gemini Pro 2.5 large language model (from here on referred to as "the agent"). The agent controls BodyInteract cases end-to-end, from perception of the current simulation state to issuing actions that operate the user interface in real time. In all studies, the same agent architecture, perception pipeline, and control loop are used; only the clinical scenario, prompting and evaluation metrics differ by study.

Our agent uses multiple modalities to parse the current simulation state. The primary understanding is derived from continuously captured screenshots of the full-screen game interface. Specific operations allow the agent to switch to video or audio capture to better evaluate the patient, such as interpreting CT scans (video) or stethoscope sounds (audio). The agent also has access to a text-based case summary (the same summary the user sees when selecting the case). It interacts with the simulation by selecting from the same menu-based actions available to human players, including dialogue options, monitoring functions, diagnostic tests, treatments, and calls (e.g., to emergency medical services).

Since the primary goal of the present work lies in evaluating the clinical reasoning ability of our agent and not its proficiency in using a mouse and keyboard to control an application, we developed a harness that allows the agent to directly issue commands to user interface elements. That is, the agent does not have to use its image recognition abilities to calculate the screen coordinates of a button, move the cursor, click and validate whether the click was successfully performed, but can instead request the harness to perform the click on a specific button (e.g., the Dialogues button). After the action has taken place, the agent can validate it by inspecting a screenshot. This also greatly improves the latency of the agent as it does not need to spend time on locating and interacting with UI elements but instead is primarily concerned with the clinical case. However, it is in principle possible to remove this harness and rely on native image-based navigation capabilities at the cost of speed and click accuracy. Further technical details are shown in Appendix B.

Comparing the Agent with Students and the Expert

We use our human expert as the gold standard. The human expert engaged in each of our four clinical cases, and we recorded the expert's decisions alongside their timestamps.

To obtain a wider set of human engagements, we assembled a dataset comprising 17,436 sessions from 9,273 individual user accounts who interacted with the four selected cases between October 29, 2024, and October 29, 2025. The human data was not collected by us and instead generously provided by BodyInteract. No additional user-level information (e.g., age, gender, prior training, or institution) was available. It is therefore possible that some sessions were not conducted in a formal educational setting and that, in a subset of runs, instructors or users may have modified non-default parameters such as the time limit. **Table 1** summarizes, for each case, the number of student sessions and their outcomes (success, timeout, and failure). We refer to one such simulation engagement, be it by the agent or a human, as a run.

Sample exclusion criteria are described in Appendix C. Each data point in our sample corresponds to one run.

Table 1 Human Student Data After Exclusions

Case	*N*	Success	Out of Time	Failure
Pneumonia	2175	1,694	454	27
Stroke	11,553	10,892	587	74
Congestive Heart Failure	885	463	411	11
Hypoglycemia	78	74	4	0
Total	14,691	13,123	1,456	112

Notes. The table shows human data with student accounts only and cancelled sessions removed.

While one might ideally prefer a deterministic AI system, contemporary LLM-based agents exhibit intrinsic randomness. Even with the sampling temperature set to 0, repeated queries can yield different completions. In our setting, this randomness is further amplified by the agent's use of screenshots as input. Because each next action is conditioned on a full-frame image of the simulation, even micro-differences in the screenshot at a given time (e.g., subtle timing differences in animations or UI state) can alter the token probability distribution for the subsequent action. As a result, running the "same" case multiple times does not produce perfectly identical action sequences. To properly characterize this stochasticity, we collected 60 runs ($N_{AI} = 60$) of our agent for each of the cases as opposed to relying on a single exemplar trajectory.

Study 1: Proof of Concept

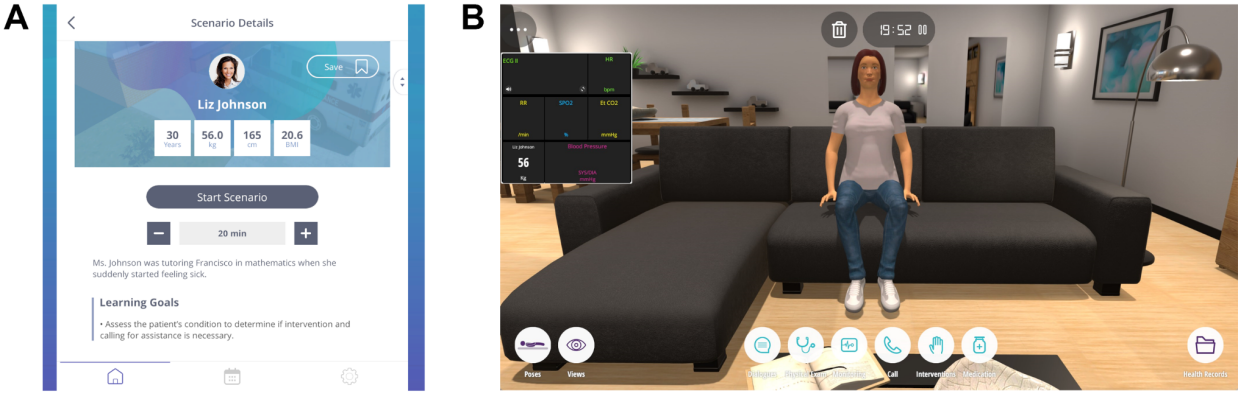
We begin by analyzing a low-acuity at-home hypoglycemia case. The simplicity of the case corresponds to a relatively small state space *S,* and a constrained action set *A*. While the clinical risk is

limited, the case still requires integration of history taking, focused examination, monitoring, and treatment under time pressure. Study 1 is organized into two parts. Study 1a asks whether an autonomous LLM-based policy π_{AI} can successfully complete the case at all—that is, whether it can understand the scenario, select appropriate actions through the technical harness, and stabilize the patient without committing critical errors. Study 1b then uses the same case to compare the agent's decisions and action sequences to our human policies π_H in the form of medical students and an experienced clinician, providing an initial lens on similarities and differences their clinical performance. The AI agent interacts with each case in multiple independent runs to account for variability.

Study 1a: Can AI play a medical simulation?

Study 1a instantiates this setup in a single BodyInteract scenario. We selected an at-home consultation in which a virtual patient suddenly becomes unwell in her home (see **Fig. 2**). To complete the case, the decision maker (a bystander in the home) must recognize the hypoglycemia, administer fast-acting carbohydrates, and arrange emergency follow-up. All these actions are required to successfully complete the case though there is no unique correct sequence of steps.

Fig. 2 Scenario Briefing (A) and Scenario Environment (B)



Notes. The scenario briefing (A) provides basic information about the patient, such as their age and weight. The scenario environment (B) renders a realistic 3D environment of the scenario (specifically for this case a home environment) in which the user can interact with the patient through the menu buttons at the bottom of the screen.

Method: To solve this case, there are multiple paths a user can choose with actions from different categories. While the primary objective is to treat the patient and call for help, there is no specific order of steps required nor does the case mandate that specific tests be performed. For instance, one user might immediately choose to look at the patient's heart rate or temperature (monitoring category), while another player might first talk to the patient to better understand their ailment (dialogues category). In our specific case, the patient is diabetic, an information that can be revealed by talking with her ("How do you feel?" -> "I feel weak and without strength… I am diabetic. Could you please check my glucose

level?"). However, a user can also monitor the patient's blood sugar level immediately and conclude that her blood sugar level is low (hypoglycemia, 57.0 mg/dL). There are two main treatment options available: glucose gel and a sugary drink, which the user can choose to administer at any time, in any quantity and in any order. Once the patient's blood sugar level stabilizes from the fast-acting carbohydrates and rises above 70 mg/dL, the user can conclude the scenario by calling for help from emergency medical services (calls category). After a few seconds, the case ends successfully because the two minimally required steps were taken (administer fast reacting carbohydrates and call for emergency medical services). While dialogue and testing are encouraged, they are not strictly necessary to successfully treat and thus complete the scenario. As the final step, the user is prompted to provide the correct diagnosis for the case from four choices.

We ran the AI agent 60 times under identical conditions ($N_{AI} = 60$). To illustrate the qualitative behavior, we describe one representative run in detail and then summarize aggregate performance across all 60 runs.

Results: In one exemplary run, the agent first remarked that while the patient appears conscious and sitting upright, her facial expression suggests that she may feel unwell. Given the lack of any further information, it speculated about broad possibilities from something like hypoglycemia to more serious conditions like a cardiac or neurological event. Given that the patient was conscious, the agent began to talk to her to gather further information and learned that she is diabetic. Then, the agent proceeded to take her blood glucose level (reading is 55 mg/dL) and remarked that "a blood glucose level below 70 mg/dL is considered low, and 55 mg/dL is significant enough to cause her symptoms of weakness". It concluded that oral treatment for her low blood sugar is the best course of action given that she was alert and able to protect her own airway and thus gave her a sugary drink. The agent continued to monitor her blood glucose level and noted that it was only slowly improving, thus it also administered glucose gel. Then, it called emergency medical services which concluded this case and successfully solved it.

This qualitative pattern was representative of the broader sample. Across all 60 AI runs, the agent successfully completed the case in 100% of sessions without triggering a critical failure and the agent selected the correct diagnosis in 97% of runs.

Discussion: In this simple at-home hypoglycemia scenario, the agent behaved in a clinically plausible and guideline-concordant manner. Starting from minimal contextual information, it generated an appropriate differential diagnosis, prioritized clarification of the patient's diabetes status, obtained a point-of-care glucose measurement, and selected oral carbohydrate therapy consistent with standard recommendations for an awake patient with low blood sugar. It also demonstrated basic closed-loop behavior by re-checking glucose values and escalating from a sugary drink to concentrated glucose gel

when the initial intervention produced only a partial response, before arranging definitive follow-up by calling emergency medical services.

From a systems perspective, this single-case experiment provides an initial validation of our end-to-end architecture. The agent was able to perceive the evolving simulation state, issue appropriate high-level commands through the harness, and update its internal plan based on feedback, all within the real-time constraints of the simulation. In our notation, π_{AI} behaves like a policy on a small (S, A, R) problem that reliably drives the process to favorable terminal rewards. The absence of critical errors and the successful completion of the case suggest that, at least in low-complexity settings with a constrained action set, an LLM-based agent can function as a safe and effective virtual provider. At the same time, Study 1a is deliberately limited. It focuses on a single, relatively straightforward case. This design is sufficient to establish feasibility but does not yet characterize how its performance compares to human learners facing the same task. These questions motivate Study 1b.

Study 1b: How does AI differ from human players?

Study 1b moves from a single feasibility demonstration to a systematic comparison between the AI agent and human decision makers on the same at-home hypoglycemia case. Here, we examine how the agent's trajectories line up with those of medical students and a medical expert. LLMs are trained on large corpora that encode extensive medical knowledge, but they lack the lived experience and tacit judgment that clinicians acquire through practice. LLMs might therefore behave differently, especially in (simulated) high-stakes scenarios. Building on the feasibility evidence from Study 1a, we now analyze the AI agent's behavior in more detail by directly comparing it to medical students and a medical expert on the same at-home hypoglycemia case. Specifically, we are interested in how success rates, timing, and action choices differ between the AI and human players.

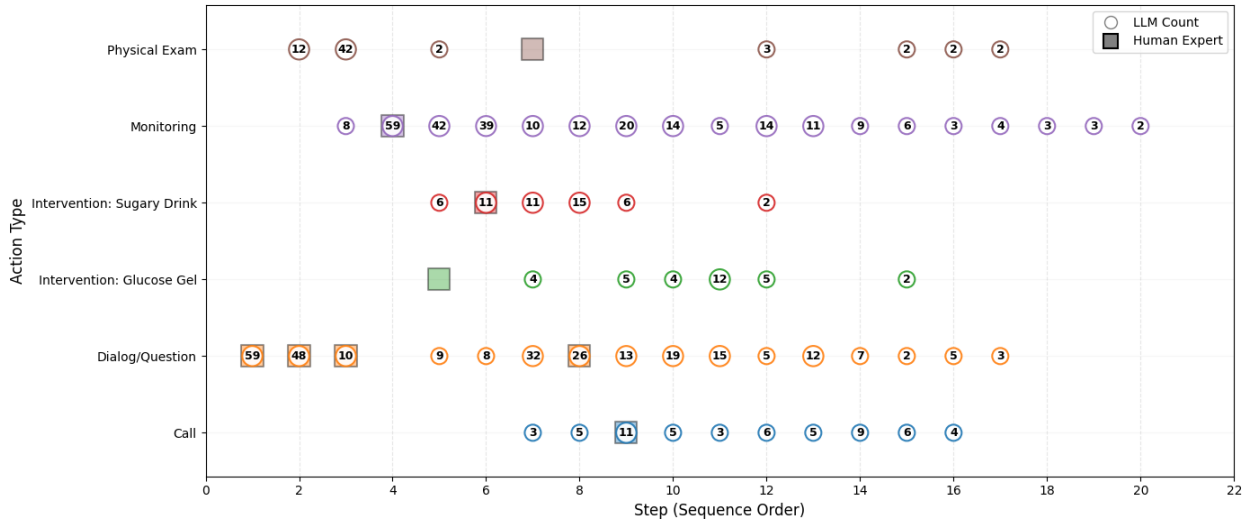
Method: We compared the 60 AI runs from Study 1a to 78 medical student runs (N_{Human} = 78) from 36 unique users, and one medical expert run. The expert session was conducted on November 12, 2025. For all AI-human comparisons, we estimated ordinary least squares (OLS) regression models with standard errors clustered at the user level, using an indicator for "AI vs. human" as the main predictor. For binary outcomes (completion), these OLS models can be interpreted as linear probability models; results are robust to logistic specifications. Lastly, we compared the sequence of actions between the AI runs and the medical expert.

Results: The presented case was successfully solved in all AI agent runs (100%; see Study 1a). In the medical student sample, four participants ran out of time with an average success rate of 94.9%, modestly and weakly significantly lower than the AI performance (B = 0.051; 95% CI [0.000, 0.102]; $t(136) = 1.98$, $p = .050$; see **Tables S1 and S2**). The medical expert successfully completed the run.

Medical students took 285 seconds on average, compared to 177 seconds for the AI agents (B = -107.897; 95% CI [-196.685, -19.110]; $t(136) = -2.40$, $p = .018$, see **Table S3**). The medical expert took 263 seconds. In 97% of all AI agent sessions, the agent correctly diagnosed the patient's condition, compared to 91% of all human participants. The medical expert correctly diagnosed the patient.

Comparing AI and human solutions in more detail reveals a nuanced picture. AI and humans differ with respect to the order of operations (see **Fig. 3**) for this case. While both usually start with a dialogue, the AI then proceeds with a physical exam much more often than with monitoring, the human expert's second step. The human expert deduces the low blood-sugar scenario from monitoring alone and immediately proceeds with an intervention, whereas the AI spends more time gathering information before intervening. The human expert also solves the case in fewer steps (6 compared to 12.73 for AI), though as mentioned above takes more time in total.

Fig. 3 Actions Taken by AI and Medical Expert

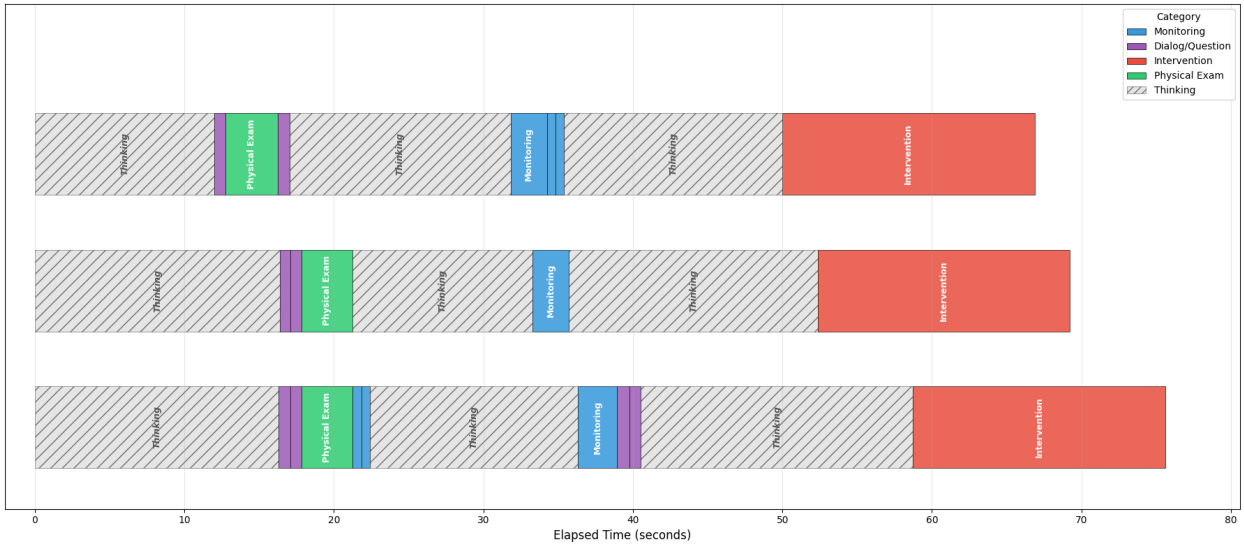


Notes. The figure shows the number of times the agent took a specific action at a specific step across all 60 runs inside the circle, excluding actions with just one observation across all runs for readability. The square shows the human expert's actions. Individual action data was not available for medical student runs.

Zooming in on the AI runs reveals their stochasticity (see **Fig. 4**). The figure shows the first 60 seconds of the action sequence for three randomly selected AI runs (time-limited for readability, see full **Fig. S1** for full sequence), highlighting how the same case can unfold in different ways. Each horizontal bar is one run; colored segments mark periods when the agent is executing a Monitoring action (blue), Intervention (red), or Physical Exam (green), and the hatched gray segments indicate "thinking time," when no action is taken while the model is generating its next decision (the simulator time continues and patient condition can worse during thinking periods). The top run begins with an extended period of thinking, then briefly switches to dialogue, performs a physical exam, returns to dialogue, and pauses

again to think before moving through a cluster of monitoring actions and finally a single intervention. The middle run follows a different pattern: it starts with thinking, then takes two consecutive dialogue actions, performs a physical exam, thinks again, does one monitoring action, pauses once more, and only then intervenes. In the bottom run, the agent begins with thinking, followed by two dialogue actions, proceeds to a physical exam, then alternates monitoring, thinking, additional monitoring, and further dialogue before a final thinking phase and intervention. Together, these three exemplars show that the agent's policy is not deterministic: it varies across runs in the order in which it chooses dialogue, exam, monitoring, and treatment, in how long it dwells in each mode, and in how much time it leaves between actions to "think" before deciding what to do next.

Fig. 4 Action Timelines for Three Representative AI Runs



Notes. The figure shows the first 60 seconds of three randomly selected AI runs (time-limited for readability, see Fig. S1 for full sequence). Each row corresponds to one complete AI run of the same case, read from left to right as elapsed time in seconds. Hatched gray segments ("Thinking") are periods with no simulator action (though time continues and the patient condition can worse), reflecting LLM deliberation and response latency. Differences in the color order across rows illustrate variation in action sequence (e.g., some runs might use more dialogue actions before a physical exam). Differences in the length of segments show variation in duration of both actions and thinking time across nominally identical runs.

Discussion: Overall, Study 1 shows that an autonomous LLM-based agent can reliably operate a realistic medical simulation and safely solve a straightforward at-home hypoglycemia case. The agent consistently reached high success rates and diagnostic accuracy and did so faster than medical students, demonstrating that our technical setup is sufficient for end-to-end control in a simple but realistic clinical scenario. In terms of policies, π_{AI} and π_{H} occupy very similar points on the reward frontier, with π_{AI} mainly outperforming on the time component of R. Study 1 thus illustrates both the promise of AI for efficient, guideline-concordant management in low-complexity situations and the early

signs of underinvestment in patient interaction and broader assessment that become more consequential in the complex emergency case examined in Study 2.

Study 2: AI in a Complex Medical Case

Study 2 is based on a more complicated patient case, specifically a 58-year-old male patient with thoracic pain for three days, fever and cough. Unlike Study 1, this is an emergency room scenario where the user has a lot more actions at their disposal, but that also requires a lot more steps to solve successfully. That is, relative to Study 1, both the latent state space *S* and the action set *A* expand substantially. Our human panel (N = 2,175 sessions) failed to solve the case within the allotted time in 21% of the cases and in a few rare occasions even made severe medical mistakes that abruptly ended the simulation. In Study 2a, we test whether π_{AI} can still stabilize the patient and complete the case. In Study 2b, we compare this AI policy to human policies π_H across case completion, clinical diagnosis, timing and action sequences.

Study 2a: Can AI solve a complex case?

As in Study 1, solving this case requires some basic steps, such as communicating with the patient. However, the user now also must order and interpret several diagnostic tests and then choose appropriate actions based on the results. In addition to an initial assessment of the patient's airway, breathing (including lung auscultation), circulation, disability, and exposure, the user must review medical tests such as blood and sputum cultures (to identify infection), a chest X-ray (to inspect the lungs), an arterial blood gas test (to assess oxygen levels), and a complete blood count (to evaluate infection and overall status). To treat the patient's pneumonia, the user needs to administer antibiotics; to manage the fever, they must provide antipyretic medication (fever reducers). For low oxygen levels (hypoxia), the patient requires oxygen therapy, for example through a nasal tube or a face mask. Lastly, fluids and electrolytes should be administered through an intravenous (IV) line before the patient is turned over to a pulmonologist (lung specialist) for further treatment. If the user takes too long to act, the patient's condition can deteriorate quickly, potentially leading to a very fast heart rate (severe tachycardia), dangerously low oxygen, low blood pressure (hypotension), and a complete stop in urine production (anuria).

Method: Study 2a applies the same autonomous AI agent and harness from Study 1a to a more complex emergency room pneumonia case. In this setting, the available action space includes a wide range of assessments, diagnostic tests, and interventions typical for an emergency department. To enable the agent to use multimodal clinical information, we augmented its inputs with audio and video from the simulation. Specifically, when the agent requested actions such as lung auscultation, chest X-

ray, or transthoracic echocardiography, the corresponding audio signals and image frames were passed directly to the Gemini Pro 2.5 model, which can interpret these modalities natively without the need for textual captions.

We again collected 60 autonomous AI runs for this case ($N_{AI} = 60$) and evaluated whether the agent could successfully stabilize the patient and complete the scenario within the allotted time.

Results: Looking at a randomly selected run, the agent first identified the patient's presentation—three days of thoracic pain, fever, and productive cough—as most consistent with pneumonia and immediately prioritized a full set of vital signs and ECG monitoring. On review, it noted high fever (39°C), tachycardia (heart rate around 116 bpm), tachypnea (rapid breathing), hypoxia (SpO_2 89% on room air), and elevated blood pressure, and explicitly flagged the constellation as a likely severe lower respiratory tract infection with impending respiratory failure. It started oxygen via nasal cannula at 4 L/min and established peripheral IV access, but when the oxygen saturation improved only minimally (to about 90%) and the heart rate rose further, it interpreted this blunted response as evidence of significant ventilation-perfusion mismatch and escalating disease severity.

The agent then escalated respiratory support to a non-rebreather mask with high-concentration oxygen and used its multimodal capabilities to auscultate the lungs and interpret their sounds. It then ordered and interpreted a chest X-ray, complete blood count, and arterial blood gas from the corresponding images. It recognized a marked leukocytosis with neutrophilia (WBC 16,000/μL), radiographic consolidation consistent with pneumonia, and an arterial blood gas showing hypoxemia (PaO_2 71 mmHg on high-flow oxygen) with mild metabolic acidosis but normal lactate. On this basis, it concluded that the patient was suffering from severe pneumonia with type 1 respiratory failure and early sepsis, explicitly warning that the persistent and worsening tachycardia (up to ~136 bpm) was an important sign of ongoing physiological stress.

In response, the agent initiated treatment: it administered intravenous antibiotics, IV fluids and electrolytes, antipyretic medication for the fever, and analgesics for chest discomfort. Follow-up vital signs showed a clear improvement in the respiratory dimension: oxygen saturation increased to 100% on the mask, respiratory rate normalized, and the temperature fell to 37.5°C. However, the heart rate remained dangerously elevated. The agent interpreted this pattern as successful reversal of hypoxia but persistent systemic inflammation and therefore arranged escalation of care by notifying pulmonology which successfully concluded the case.

Study 2b: How does AI differ from human players?

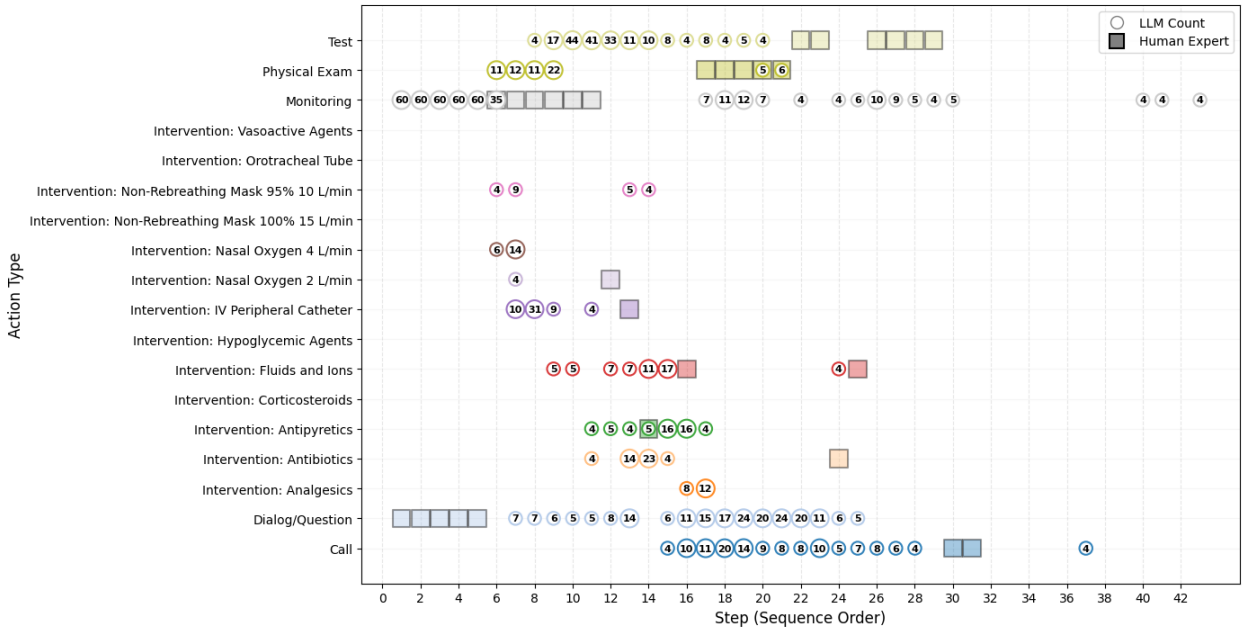
Similar to Study 1, we are interested in comparing the chosen actions of the AI agent to medical students and a medical expert to understand potential differences and blind spots.

Method: We compare AI and human performance on the same pneumonia emergency-room case. On the AI side, we again used the AI agent runs (N_{AI} = 60) from Study 2a. On the human side, we analyzed all 2,175 runs (N_{Human} = 2,175) from 640 unique users, and one medical expert run. Outcome variables mirror Study 1b and include success vs. failure, total time elapsed and diagnostic accuracy. For all AI-human comparisons, we estimated OLS regressions with source (AI vs. human) as the main predictor and reported heteroskedasticity-robust standard errors clustered at the user level to allow for arbitrary correlation across runs from the same user. For binary outcomes (completion), these OLS models can be interpreted as linear probability models; results are robust to logistic specifications. We also again compared the sequence of actions between AI and humans.

Results: The AI agent successfully completed the case in 88.3% of all cases, compared to 77.8% of human participants (B = 0.104; 95% CI [0.018, 0.191]; $t(2233) = 2.37$, $p = .018$; see **Tables S4 and S5**). Both groups suffered from timeouts as the primary cause of failing to complete the case. The medical expert successfully completed the run. The AI agent solved the case in 443 seconds on average, compared to 620 seconds for the medical students (B = -196.884; 95% CI [-276.085, -117.684]; $t(2233) = -4.87$, $p < .001$; see **Table S6**). The medical expert took 725 seconds. Interestingly, the AI agents only diagnosed the patient correctly in 55% of all runs (compared to 94% of human participants), choosing sepsis as the primary diagnosis over pneumonia in many cases. Inspecting the reasoning traces of the agent revealed that while it agreed that the patient suffered from pneumonia, it argued that sepsis was the better fitting diagnosis based on the symptoms. The medical expert diagnosed the patient correctly as suffering from pneumonia.

The order of operations taken by AI and humans again differed for this case (see **Fig. 5**). While the AI agent usually started with monitoring, the human expert first talked to the patient, something the AI did not do until after starting multiple interventions, such as providing the patient with oxygen. The treatment patterns overall were comparable, with both AI and the medical expert first providing oxygen before administering fluids and medications. Average step count was also comparable, with the human expert taking 31 steps compared to 31.22 for the AI agent.

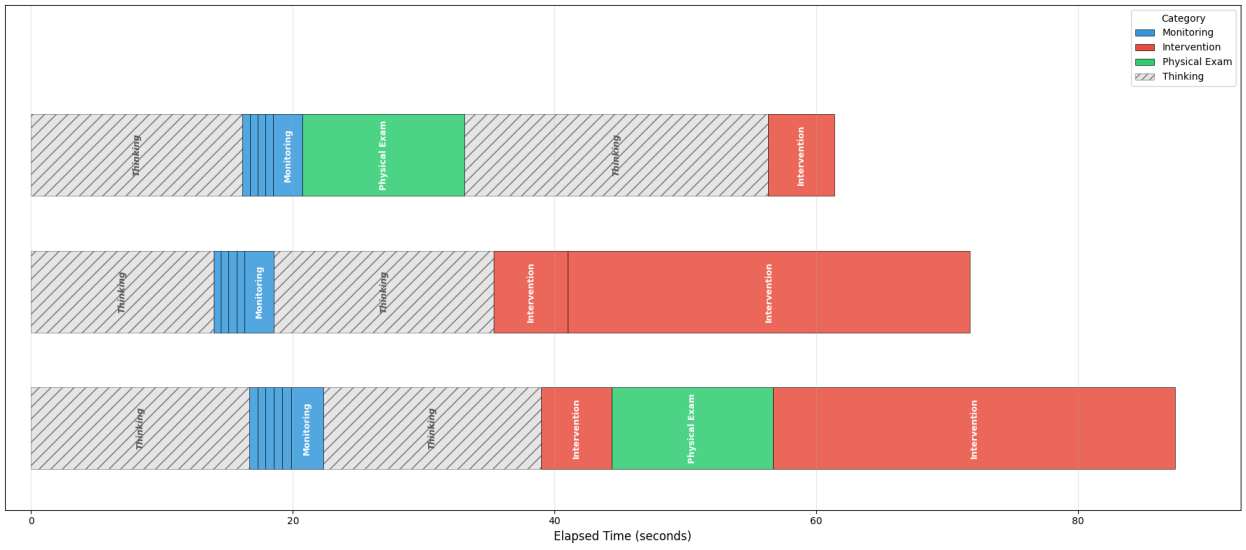
Fig. 5 Actions Taken by AI and Medical Expert



Notes. The figure shows the number of times the agent took a specific action at a specific step across all 60 runs inside the circle. We exclude the bottom 10% of least frequent actions taken by AI across all 60 runs as well as steps with less than five observations for readability. The square shows the human expert's actions.

Similar to the simpler case, we again observe variability across AI runs for action sequence order, time and thinking duration (see **Fig. 6**).

Fig. 6 Action Timelines for Three Representative AI Runs in a Complex Case



Notes. The figure shows the first 60 seconds of three randomly selected AI runs (time-limited for readability, see Fig. S2 for full sequence). Each row corresponds to one complete AI run of the same case, read from left to right as elapsed time in seconds.

Hatched gray segments ("Thinking") are periods with no simulator action (though time continues and the patient condition can worse), reflecting LLM deliberation and response latency. Differences in the color order across rows illustrate variation in action (e.g., some runs examine first, others intervene earlier). Differences in the length of segments show variation in duration of both actions and thinking time across nominally identical runs.

Discussion: Overall, Study 2 shows that AI can completely handle a complex emergency room case that involves dynamic test ordering and hypothesis updating. In addition, AI can consult multiple modalities, such as lung audio auscultation, to further enhance its diagnostic accuracy. In the richer (*S*, *A*, *R*) environment of this case, π_{AI} still outperforms the average human policy π_{H} on completion and time, but at the cost of weaker alignment with the diagnostic accuracy. The medical expert's behavior sits between these extremes: slower and more exploratory than the AI, more focused and less redundant than the students. Our results highlight both the promise and the current blind spots of AI assistance in acute care settings.

Study 3: AI across multiple complex cases

Study 3 broadens the analysis from a single emergency-room case to a portfolio of three complex scenarios. We extend the reward function *R* to also include patient engagement performance and the economic expenses associated with diagnostic testing. These additions allow us to study how different policies π trade off components of *R*, such as time, accuracy and economic efficiency, across complex cases.

Method: We pooled three complex emergency room cases into a multi-case benchmark. Specifically, we included the pneumonia case from Study 2 together with an acute stroke case and a congestive heart failure case. For our comparison we analyzed 180 AI agent runs (N_{AI} = 180; 60 per case), 14,613 medical student runs (N_{Human} = 14,613) from 8,615 students across the same three cases, and one emergency physician (medical expert) run per case. As before, we report our established measures of (a) whether the case was successfully completed, (b) total time elapsed (in seconds), (c) whether the final diagnosis matched the simulator's reference diagnosis. In addition, we now include (d) the share of recommended communication actions taken (talking to the patient and (e) the cost for each test and procedure. Specifically, we mapped each diagnostic test and imaging procedure to a cost using a fee schedule from BlueCross BlueShield (https://payerprice.com/rates/71046-CPT-fee-schedule) to obtain a per-session test-cost index. This mapping yielded a simple test-cost proxy per run. All dollar values should therefore be interpreted as approximate and primarily useful for relative comparisons rather than as estimates of real-world spending. We then estimated OLS regressions with heteroskedasticity-robust standard errors clustered at the user level, using source (AI agent vs. human student/expert) as the main predictor and include case fixed effects to control for case difficulty across the three scenarios. For binary outcomes (completion), these OLS models can be interpreted as linear probability models; results are robust to logistic specifications.

To characterize policy consistency, we drew on the detailed action logs from the AI runs. For each case, we treated each run as a set of distinct actions (e.g., obtaining vital signs, ordering a specific test, administering a particular medication) and computed the mean Jaccard similarity between action sets across all pairs of runs. We then focused on the first ten actions of each run and compared them to the modal sequence using the Levenshtein distance, which counts the minimum number of insertions, deletions, or substitutions needed to transform one sequence into another. Finally, we conducted a stepwise convergence analysis over the first 15 actions, identifying at each step the most common action and the proportion of runs that choose it.

Results: Across the three complex cases, the AI agent was more likely than humans to complete the cases successfully. Medical students succeeded in 89.3% of sessions (13,049/14,613), with 9.94% timing out and 0.77% failing due to critical mistakes. The AI agent succeeded in 95.0% of sessions, with the remaining 5.0% ending in timeout or failure, a significant improvement (controlling for case: B = 0.198; 95% CI [0.155, 0.240]; $t(14791) = 9.09$, $p < .001$; see **Tables S7 and S8**). This adjusted effect is larger than the raw difference (5.7 percentage points) because human sessions are heavily concentrated in the relatively easier stroke case, whereas the AI runs are evenly distributed across cases; controlling for case difficulty removes this favorable human case mix. In our expert benchmark, the emergency physician successfully completed the pneumonia and acute stroke cases but ran out of time in the congestive heart failure case. This is likely because despite our expert's decades of clinical experience, the case happens on a very condensed timeline of 20 minutes compared to hours in the real world, most of which the expert has to spend navigating the unfamiliar simulation user interface instead of directing clinical staff in an ER (see limitations for more discussion).

The AI agent was also markedly faster. Medical student sessions took on average 484 seconds, whereas AI sessions took 303 seconds on average, a large and significant time advantage for the AI (controlling for case: B = -326.090; 95% CI [-365.105, -287.075]; $t(14791) = -16.38$, $p < .001$; see **Table S9**). Again, this adjusted difference is larger than the raw 181-second gap because students disproportionately appear in the faster stroke case, while AI runs are evenly distributed across cases. The medical expert took 547 seconds on average across the two successfully completed cases. In terms of diagnostic accuracy, the agent assigned the correct primary diagnosis in 82.78% of runs, compared to 83.56% for human participants, and the emergency physician correctly diagnosed both completed cases.

Process-level analyses revealed systematic differences in how AI agents and humans used communication. Aggregated across the three cases, humans carried out 61.88% of recommended dialogue actions, compared to 22.15% for the AI. Thus, in this multi-case setting, AI agents again tended to prioritize key treatments and move quickly through the scenario, while humans invested more in talking to the patient.

Cost analyses, based on our test-cost proxy, indicate that the AI agent generally ordered more tests than our human expert. This suggests that the expert can treat confidently on a leaner set of high-yield datapoints, whereas the AI still tends to purchase more information before acting. For cost, AI sessions incurred an average of $608 for the three complex cases. Considering only the two cases the medical expert successfully completed, AI sessions cost $660 versus $346 for the human expert.

Turning to action-sequence consistency, **Table 3** reports Jaccard indices over action sets and Levenshtein distances over the first 10 actions. Across runs within each case, mean Jaccard indices range from 0.687 to 0.757, indicating substantial overlap in which actions the AI takes. At the same time, only 3-8% of runs exactly match the modal 10-step sequence, and mean Levenshtein distances around 5 suggest notable variability in ordering. In other words, AI runs tend to involve similar sets of diagnostic and treatment actions, but the precise order in which these actions are taken differs across runs (the AI exhibits set consistency but sequence variability).

Table 3 Action Sequence Consistency Across AI Runs

Case	Mean Jaccard	SD	Modal Sequence Match	Mean Levenshtein Distance
Pneumonia	0.687	0.111	8.3% (5/60)	5.17 (SD: 1.68)
Stroke	0.736	0.106	3.3% (2/60)	5.55 (SD: 1.65)
CHF	0.757	0.108	6.7% (4/60)	5.25 (SD: 1.88)

Notes. Jaccard index ranges from 0 (no overlap) to 1 (identical action sets). Modal sequence match indicates the proportion of runs whose first 10 actions exactly match the most common sequence. Levenshtein distance measures edit distance between action sequences.

A qualitative convergence analysis further illustrated this pattern. While no single action achieves >90% agreement at any step, several actions show moderate consensus (50-75%) in the early steps—particularly vital sign monitoring (heart rate, O2 saturation) and case-specific key interventions. This suggests the AI has learned robust "anchor actions" that appear across most runs, even as the surrounding sequence varies.

Discussion: Study 3 extends our single-case findings from Studies 1 and 2 to a small bundle of complex emergency-room scenarios. Across three distinct acute presentations—pneumonia, ischemic stroke, and congestive heart failure—the AI agent consistently achieved higher case completion rates than medical students and did so substantially faster, while matching human learners on overall diagnostic accuracy. That the agent maintains this performance edge across heterogeneous conditions

suggests that its ability to operate the simulation end-to-end is not confined to a particular disease or workflow pattern but generalizes to a broader class of time-critical emergency cases.

At the same time, Study 3 reinforces a central qualitative pattern from the earlier studies: AI and humans appear to occupy different points on a process-level "thoroughness-efficiency" spectrum. The AI agent invested much less in dialogue than students and typically ordered a narrower set of diagnostic tests. This more minimalist, treatment-focused style translated into lower or comparable diagnostic test expenditures relative to human learners, but it also meant that the agent collected less contextual and longitudinal information about patients than humans typically did. Compared with the medical expert, $\boldsymbol{\pi}_{\text{AI}}$ selects action sequences that incur higher diagnostic costs, indicating that their test selection does not yet match the human expert's parsimonious use of diagnostics. Taken together, our results suggest that current LLM-based agents can function as fast and effective stabilizers in complex acute care simulations, but they may underinvest in the broader information-gathering, communication, and cost-aware test selection that characterize expert human practice.

The action-sequence analysis shows why. Across runs, the AI reliably converges on a common set of high-yield actions but the order in which these actions are taken varies (only 3-8% of runs matching the modal 10-step sequence). The policy is therefore consistent in *what* it does and flexible in *how* it gets there.

Study 4: Understanding the agent's reasoning process

Studies 1-3 treated the AI agent as a black box: we observed its actions and outcomes across runs and compared them to human policies. In Study 4, we "open the box" within runs. Specifically, we exploited a unique feature of our experimental design. At each step, we prompted the agent to state its current predictions for the patient's diagnosis. This makes the belief state b_t for $\boldsymbol{\pi}_{\text{AI}}$ directly observable, allowing us to study how its policy updates beliefs over S in response to new observations and actions. Motivated by recent work showing that LLMs are often overconfident and miscalibrated (Geng et al. 2024; Kapoor et al. 2025; Wang et al. 2024; Xiong et al. 2024), we ask whether the policy π_{AI} exhibits properties consistent with an approximately optimal value-of-information strategy, including Bayesian-like belief updating, sequencing of tests by declining marginal information value, and calibrated confidence in its final diagnoses.

Method: Each case presents four possible diagnoses at the end of the scenario. During AI runs, we intermittently prompted the model (through a separate logging channel) to report a probability distribution over these four diagnoses, based on everything observed so far. This secondary channel was isolated from the main control loop so that belief logging did not influence the agent's actions. We combined the four cases presented in Studies 1-3, yielding 240 AI sessions. In 13 out of 240 sessions,

the AI agent failed to follow the prediction instructions at least once and returned an incomplete response, resulting in 227 AI sessions with 1305 steps and 1078 transitions in total. We computed:

- **Belief shift:** the absolute change in the probability assigned to the true diagnosis between consecutive predictions within a session
- **Entropy:** Shannon entropy (in bits) of the full four-way probability distribution, capturing overall diagnostic uncertainty
- **Probability of the correct diagnosis:** the probability assigned to the simulator's reference diagnosis for that case

For interpretability, we also flagged a transition as producing a "non-trivial revision" if the belief shift was at least 5 percentage points, treating smaller changes as noise-level fluctuations. To link beliefs to performance, we recorded at each step whether the agent would already be correct if forced to stop and diagnose at that moment (i.e., whether the highest-probability label matched the reference diagnosis).

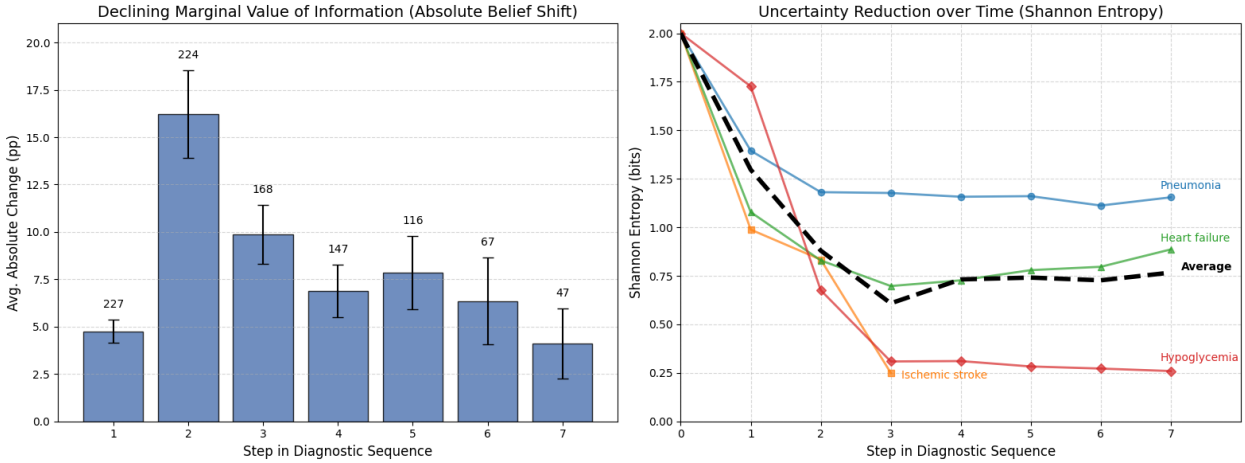
For the value-of-information analysis, we pooled transitions across all four cases and focused on prediction steps 1-7, as steps beyond that showed very few observations per step (a few long-running runs with 1-2 observations per step). We defined steps 1-2 as "early" and steps 6-7 as "late." At the transition level, we summarized, by step, the mean belief shift and the fraction of non-trivial revisions. At the session level, we computed, for each run, the mean absolute change in probability on the true diagnosis ("belief shift"), the mean probability on the true diagnosis, and Shannon entropy over the early (steps 1-2) and late (steps 6-7) windows. We restricted the analysis to sessions with predictions in both early and late windows (N = 67), which necessarily over-represents runs with longer trajectories.

Finally, we conducted a case-specific analysis for the pneumonia case. For this case we repeated the early-vs-late session-level summaries (N = 20 sessions with data in both windows) and, separately, used the first and last predictions in each session (N = 60) to examine (i) final probabilities on pneumonia and sepsis for sessions that did vs. did not end with the correct primary diagnosis, and (ii) entropy reduction (first minus final entropy) as a function of accuracy.

Results: If the agent approximately follows a value-of-information policy, it should front-load high-yield diagnostic actions and experience diminishing belief updates over time. This is what we observe (see **Fig. 7**). At the transition level, in the first two diagnostic steps, 59.2% of actions produced a meaningful belief revision (≥5 percentage points), with an average belief shift of 10.5 percentage points. The second step is especially strong on average (16.2%), most likely because the first step often includes monitoring and dialogue, whereas the second step includes tests such as CT scans or X-rays which

provide more information. By steps 6-7, only 28.9% of transitions produced meaningful revisions, with an average belief shift of 5.4 percentage points. Step-level mean shifts tend to be smaller at later steps (Spearman ρ ≈ -0.43 over steps 1-7), but with only seven step-level averages this trend is imprecisely estimated. The stronger evidence comes from the within-session early late comparison.

Fig. 7 Declining Marginal Value of Information (left) and Uncertainty Reduction over Time (right) Across Four Cases.



Notes. Left: The chart shows absolute belief change of the true diagnosis at every step across all four cases with error bars, averaged across all transitions. The number above each bar shows the number of observations at each step. Right: The chart shows the reduction in Shannon entropy over the course of all four cases, starting at step 0 where the probability for all four diagnostic options is equally likely. We exclude steps beyond 7 for both charts as the number of observations becomes small and estimates noisy (see Method).

At the session level, early steps produce substantially larger belief shifts than late steps. The mean belief shift over steps 1-2 is 14.9 percentage points, compared to 6.2 percentage points over steps 6-7. The average difference (late - early) is -8.657 percentage points (95% CI [-12.610, -4.704]), $t(66) = -4.37$, $p < .001$). In other words, the actions the agent chooses early in the case tend to be precisely those that move its beliefs about the true diagnosis the most.

As uncertainty declines, beliefs generally move toward the true diagnosis. For the same sessions, the mean probability of the reference diagnosis increases from 66.3% in the early window to 78.1% in the late window. The average increase is 11.746 percentage points (95% CI [5.737, 17.756]; $t(66) = 3.90$, $p < .001$). Entropy exhibits an even stronger pattern: mean Shannon entropy declines from 1.20 bits in the early window to 0.61 bits in the late window, a reduction of 0.61 bits (95% CI [-0.695, -0.481]; $t(66) = -10.94$, $p < 0.001$). Thus, across sessions with non-trivial trajectories, early steps are both more informative and followed by substantial consolidation of belief onto the correct label.

These window-based results are consistent with a simpler first-last comparison. Across all 217 sessions, entropy declines from 1.38 bits at the first snapshot to 0.74 bits at the last (Δ = -0.639 bits;

95% CI [-0.717, -0.560]; $t(216) = -16.10$, $p < .001$)—and entropy decreases in 90.8% of sessions. The magnitude of this reduction varies by case: hypoglycemia shows strong convergence (1.78 → 0.31 bits), heart failure is similar (1.24 → 0.75), the stroke case starts with relatively high confidence and tightens modestly (1.00 → 0.81 bits), and the pneumonia case shows more modest entropy reduction (1.35 → 1.12 bits), consistent with a more ambiguous labeling problem. In the hypoglycemia case, the probability assigned to the correct label increases from 39.6% initially to 92.8% at the final snapshot. In the heart failure case, it increased from 67.0% to 80.9% and in stroke case, the agent starts with high confidence in the correct diagnosis (76.1%) and finishes at 80.6%.

The pneumonia case presents an exception. Here, the probability on the reference diagnosis (pneumonia) *decreased* from 63.9% to 47.0%, while probability on sepsis increased from 9.8% to 44.9%. Examination of the agent's reasoning traces reveals that this shift reflects clinically sophisticated updating rather than diagnostic error. As the agent acquired information about the patient's vital signs, laboratory values, and arterial blood gas, it observed evidence of systemic inflammatory response and organ dysfunction—hallmarks of sepsis secondary to pneumonia. The agent's reasoning explicitly noted: "Sepsis is not a separate disease but a systemic response to infection... he has sepsis secondary to pneumonia." A closer look at final beliefs shows that the agent's probabilities encode meaningful differences in diagnostic stances, even when multiple labels remain plausible. Among the 60 pneumonia sessions, 31 runs end with pneumonia as the top-probability label and 24 end with another label (typically sepsis, 5 excluded due to bad agent prediction results). In sessions that ultimately choose pneumonia, the final probability on pneumonia averages 61.4% compared to 34.9% in sessions that do not ($\Delta = 0.266$; 95% CI [0.212, 0.320]; $t(53) = 9.88$, $p < .001$).

To assess calibration, we grouped belief snapshots by the probability the agent assigned to the true diagnosis (p_{true}) and computed how often it would already be correct if forced to decide at that step (see **Table 4**). When p_{true} was between 0-40%, the agent would have been correct in 45.83% of steps; for 40-60%, in 86.31%; and for 60-80% and 80-100%, in 100% of steps. Final-step calibration shows a similar monotone pattern: when the agent ends a case with 80-100% probability on the reference diagnosis, it is correct 100% of the time; in the 60-80% range, 100%; in the 40-60% range, 78.95%; and in the 0-40% range, 0%. Thus, higher stated confidence consistently tracks higher actual accuracy, both along trajectories and at decision time.

Table 4 Trajectory-Based Calibration of the True Diagnosis (Step Level)

Probability range on true diagnosis	Mean p_{true}	Accuracy if stopped now	N (steps)
0-40%	34.42%	45.83%	168

40-60%	53.51%	86.31%	241
60-80%	73.04%	100.0%	476
80-100%	92.42%	100.0%	420

Pooling across cases, sessions that end with a correct diagnosis show substantially larger entropy reductions than incorrect sessions. Across all four cases, correct sessions (N = 200) reduce entropy by an average of 0.641 bits, compared to 0.242 bits for incorrect session ($\Delta = 0.399$; 95% CI [0.274, 0.523]; $t(78) = 6.39$, $p < .001$). This pattern primarily reflects the relative ease of the hypoglycemia and stroke cases, where the agent can nearly resolve diagnostic uncertainty by the end of the encounter, whereas pneumonia remains ambiguous. Within Case 150, however, entropy reduction does not significantly distinguish correct from incorrect sessions (0.258 vs. 0.181 bits; $\Delta = 0.077$; 95% CI [–0.033, 0.186]; $t(52) = 1.41$, $p = .165$), reinforcing the view that, in this more ambiguous case, entropy reduction is as much a marker of case difficulty as of diagnostic skill. Together, these patterns suggest that the agent's policy behaves like a value-of-information policy in straightforward cases—front-loading high-yield actions, increasing probability mass on the true label, and sharply reducing entropy.

Discussion: Taken together, these analyses provide evidence that the AI agent's implicit policy is calibrated to information value. Diagnostic actions produce genuine belief revisions rather than mere confirmation; the magnitude of revision declines over time consistent with diminishing marginal returns; and entropy—a direct measure of diagnostic uncertainty—falls substantially over the course of each session. Critically, the agent's stated probabilities are well-calibrated to actual accuracy, both along the trajectory and at the final decision, suggesting it can appropriately distinguish cases where it has reached diagnostic confidence from cases where genuine uncertainty remains.

The calibration finding is particularly important for potential clinical deployment. If an AI agent's confidence estimates are meaningful—if high confidence reliably predicts accuracy—then these estimates could inform human-AI collaboration strategies. This is noteworthy given converging evidence that large language models are not naturally well calibrated: they tend to be overconfident in incorrect answers, and their verbalized probabilities often deviate substantially from empirical accuracy unless explicitly tuned for calibration (Xiong et al. 2024; Kapoor et al. 2024; Wang et al. 2024; Geng et al. 2024). Physicians might defer to AI recommendations when confidence is high and provide closer oversight or additional testing when confidence is low.

These results also have implications for interpreting the AI agent's tendency to order fewer tests than human learners, documented in Study 3. They suggest that π_{AI} behaves more like a value-of-information policy than a case of underinvestment in diagnosis: given a belief state b_t it is more likely than π_H to select actions that produce large belief revisions and to stop testing once the expected gain in reward R from further information becomes small, while human learners often order additional tests that merely confirm existing beliefs. We cannot test this directly without comparable belief-state data from human sessions, but the AI-side evidence is consistent with this interpretation.

Limitations

These contributions need to be interpreted in light of several limitations. First, our evidence is based on a high-fidelity simulation environment rather than real clinical practice. BodyInteract encodes clinically vetted disease progressions, test results, and success criteria in a high fidelity simulation. Though we treat this simulation as a “digital twin” of real patients we acknowledge that our analysis is built on a model of how a patient’s condition presents and developed, rather than on real patients.

Second, our human baseline consists predominantly of medical students and a single emergency physician per case, rather than a broad panel of attending physicians. Students are a natural comparison group because BodyInteract cases are designed for education and assessment, but they are not representative for clinical practice. The expert’s behavior offers a more mature benchmark but is limited to one individual per case. This design allows us to compare AI, learners, and one expert under identical conditions, yet it also means we cannot characterize the full distribution of expert policies or inter-physician practice variation. Moreover, we have complete information for all actions and outcomes for the AI model and the expert, but for the students we only have the final outcomes. Given that the simulation has a complex user interface with many menu entries, it is also likely that some of the speed advantages of the AI agent over humans can be explained by interface friction, especially given the short case durations. In the congestive heart failure case, the expert performed more than one action every minute before timing out, but spent most of the time navigating the menu instead of actively treating the patient.

Third, our case library covers four acute conditions including three emergency room scenarios. These are important, time-critical problems, but they do not span all specialties, acuity levels, or chronic-care workflows. To increase the external validity to other diseases, settings, and future research is required.

Fourth, on the AI side we evaluate a single vendor family (Gemini) in a specific architecture that uses a harness to control the simulation. The harness allows the agent to invoke UI elements directly rather than relying on low-level mouse and keyboard actions, which isolates clinical reasoning from interface

navigation but also makes the environment more forgiving than many real clinical IT systems. We also constrain the agent's access to a particular prompt and cost structure.

Fifth, we use the POMDP framework as a conceptual lens to interpret differences in information use and timing, but we do not solve the underlying POMDP nor estimate belief states or value functions directly. As such, our conclusions about implicit value-of-information thresholds or objective functions should be read as interpretations supported by the evidence, not as identified structural parameters. Our calibration analysis is also restricted to a four-option diagnostic choice in a controlled environment; it should not be interpreted as evidence that LLMs are generally well calibrated in open-ended clinical reasoning, especially given recent work documenting substantial miscalibration in more naturalistic settings. In contrast, a stream of work develops fully specified MDP and POMDP models to derive optimal policies under uncertainty, including applications in humanitarian search-and-rescue (Bravo, Leiras, and Oliveira 2019), security maintenance (Bensoussan, Mookerjee, and Yue 2020), and risk-sensitive control (Xia 2020; Xia, Zhang, and Glynn 2023). A promising avenue for future research is to combine our simulation approach with structural estimation or risk-sensitive MDP formulations, bringing the richer objective functions used in this literature to bear on the multi-dimensional time-cost-quality trade-offs we document.

Discussion

This paper examines whether a modern multimodal large language model can move beyond static, single-task evaluation and function as an autonomous diagnostic agent in a dynamic, high-fidelity clinical simulation. Across four clinical cases of varying complexity, we compare an LLM-based agent's policy $\boldsymbol{\pi}_{\mathrm{AI}}$ to those of medical students and an emergency physician $\boldsymbol{\pi}_{\mathrm{H}}$ in a continuous-time, partially observable decision environment. Our results establish three primary contributions.

First, we show that a modern multimodal LLM can act as an autonomous virtual provider at the workflow level rather than only at the task level. Within a set of simulated cases, the agent engages in a dynamic sequence of information-gathering and treatment actions under explicit time pressure, rather than solving a single vignette in one shot. It consistently stabilizes patients and successfully completes clinical cases. Across the four scenarios (from a simple at-home hypoglycemia case to complex emergency-room presentations) $\boldsymbol{\pi}_{\mathrm{AI}}$ recognizes the problem, orders and interprets relevant tests, initiates appropriate treatments, and closes the loop through monitoring and escalation. The agent achieved higher completion rates than medical students and completed cases substantially faster, while matching human learners on diagnostic accuracy.

Importantly, this workflow-level competence arises in a realistic simulation environment with nontrivial lead times and evolving patient trajectories, not from static question answering. Returning to

Christensen's "Jobs to be Done" lens, the agent in our simulations can execute a substantial portion of the job of patient stabilization: it orders tests, interprets multimodal data, and implements treatment sequences that reliably move patients from unstable to stable states.

Second, our comparison with human decision makers highlights how π_{AI} orchestrates multimodal workflows relative to π_{H}. The agent integrates visual, textual, and audio streams—reading case briefings, interpreting chest X-rays and monitor screens, and listening to lung sounds—and uses these inputs to guide a sequence of tests and treatments in real time. The comparisons reveal that the AI occupies a distinct point on the cost-quality and efficiency-thoroughness frontier. Relative to students, π_{AI} behaves like a fast stabilizer: it prioritizes a focused set of high-yield tests and treatments, completes cases considerably faster, and attains at least comparable diagnostic accuracy and better overall case completion. Students, by contrast, tend to spend more time and, in many cases, order a broader set of tests, including those with limited marginal information value. The expert practices a high-yield diagnostic style with outcomes comparable to or better than the agent while ordering fewer diagnostic tests and engaging more in patient communication. This suggests that further shaping of the agent's objective (e.g., penalizing diagnostic expenditures more strongly) could move π_{AI} closer to expert-like test selection, achieving not just speed but also cost efficiency.

Interestingly, the agent invests considerably less in patient engagement than humans do, a meaningful limitation of its current policy that might require explicit modification for real-world deployment. While these interactions are not required to complete cases in the simulator, they are integral to real clinical care—building rapport, eliciting nuanced history, and ensuring patient understanding. This divergence in communication behavior suggests a complementary human-in-the-loop division of labor. In our simulations, the agent functions as a diagnostic engine or "clinical logistician": it manages orders, monitors vitals, and executes high-yield tests and treatments quickly. The human clinician, by contrast, naturally takes the role of an empathetic interface: gathering rich contextual history, communicating plans, and providing reassurance. This pattern points to human-AI configurations in which the AI handles stabilization, logistics, and information management—most naturally in roles such as rapid triage, over-the-shoulder second-opinion support, or guidance in physician-scarce settings—while the clinician remains responsible for the relational and contextual aspects of care that the current agent largely neglects but are central to real-world medicine.

Third, by logging the agent's evolving diagnostic beliefs, Study 4 "opens the black box" of π_{AI}. We find that the agent's internal belief dynamics exhibit several hallmarks of value-of-information reasoning. Early in a case, the tests the agent selects tend to induce large shifts in the probability assigned to the true diagnosis and substantial reductions in diagnostic entropy. As the encounter progresses, both belief shifts and entropy reductions diminish, indicating that the agent front-loads high-yield information and then experiences declining marginal informational returns from additional tests.

Moreover, in contrast to prior work, the agent's stated confidence is meaningfully calibrated: higher reported probabilities on the reference diagnosis are associated with higher empirical accuracy. When the agent finishes a case with high confidence, it is almost always correct; when it remains uncertain, errors are more likely.

Our results have several managerial implications for the design of diagnostic pathways and human-AI collaboration. Crucially, they should not be interpreted as an argument for leaving patients alone with an unsupervised LLM. Rather, they suggest how an agent like ours can be positioned inside human-centered workflows, with clear boundaries on what roles the AI takes on and what clinicians retain. The agent's speed, calibrated confidence, and VOI-consistent behavior make it useful in specific roles where human oversight remains central.

First, the agent can serve as a rapid triage engine in high-volume or mass-casualty settings. In crowded emergency departments or disaster scenarios, the primary constraint is often physician attention and time. The agent's ability to quickly stabilize straightforward cases and produce calibrated confidence estimates makes it well-suited for initial sorting: high-confidence, stable presentations can be flagged for streamlined physician review, while low-confidence or deteriorating cases are escalated immediately for hands-on evaluation. This reduces passive costs (time, staffing burden) while retaining physician agency.

Second, the agent can function as a real-time auditor or "second pair of eyes" during human-led encounters**.** Rather than driving the clinical workflow, the AI observes in parallel—processing the same multimodal inputs the physician sees—and flags discrepancies: a diagnosis the physician may not have considered, a high-yield test not yet ordered, or a trajectory that diverges from the agent's evolving differential. The calibration findings from Study 4 are particularly relevant here. When the agent assigns high confidence to a diagnosis the physician appears to be missing, that signal warrants attention; when confidence is low, the alert is correspondingly softer. This positions the AI as a cognitive safety net rather than an autonomous decision-maker.

Third, the agent can extend diagnostic capability to settings where no physician is physically present—battlefields, remote or rural locations, or under-resourced healthcare systems. In such contexts, a medic, nurse, or even a trained bystander could interact with the simulation-like interface while the agent guides information gathering and treatment in real time. The at-home hypoglycemia case in Study 1 illustrates this: a non-clinician bystander, assisted by the agent, would be able to successfully stabilize the patient and arrange appropriate follow-up. Remote physician oversight via telemedicine can provide an additional supervisory layer when connectivity permits, but the agent ensures that time-critical stabilization is not delayed by the absence of on-site expertise.

Acknowledgements

We thank the Mack Institute for Innovation Management for their generous research funding. We thank Hummy Song and Martin Bittner for their helpful comments.

References

Ayers, J. W., A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Longhurst, M. Hogarth, D. M. Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern. Med.* **183**(6): 589–596.

Bavafa, H., S. Savin, C. Terwiesch. 2021. Customizing Primary Care Delivery Using E-Visits. *Prod. Oper. Manag.* **30**(11): 4306–4327.

Bensoussan, A., V. Mookerjee, W. T. Yue. 2020. Managing Information System Security Under Continuous and Abrupt Deterioration. *Prod. Oper. Manag.* **29**(8): 1894–1917.

Bravo, R. Z. B., A. Leiras, F. L. Cyrino Oliveira. 2019. The Use of UAVs in Humanitarian Relief: An Application of POMDP-Based Methodology for Finding Victims. *Prod. Oper. Manag.* **28**(2): 421–440.

Chen, T. C., E. Multala, P. Kearns, J. Delashaw, A. Dumont, D. Maraganore, A. Wang. 2023. Assessment of ChatGPT's performance on neurology written board examination questions. *BMJ Neurol. Open* **5**(2): e000530.

Christensen, C. M., S. Cook, T. Hall. 2005. Marketing malpractice: the cause and the cure. *Harv. Bus. Rev.* **83**(12): 74–83, 152.

Dai, T., S. Singh. 2025. Artificial Intelligence on Call: The Physician's Decision of Whether to Use AI in Clinical Practice. *J. Mark. Res.* **62**(5): 854–875.

Dai, T., S. Tayur. 2022. Designing AI-augmented healthcare delivery systems for physician buy-in and patient acceptance. *Prod. Oper. Manag.* **31**(12): 4443–4451.

Diaz-Navarro, C., R. Armstrong, M. Charnetski, K. J. Freeman, S. Koh, G. Reedy, J. Smitten, P. L. Ingrassia, F. M. Matos, B. Issenberg. 2024. Global consensus statement on simulation-based practice in healthcare. *Adv. Simul. Lond. Engl.* **9**(1): 19.

Eisemann, N., S. Bunk, T. Mukama, H. Baltus, S. A. Elsner, T. Gomille, G. Hecht, S. Heywang-Köbrunner, R. Rathmann, K. Siegmann-Luz, T. Töllner, T. W. Vomweg, C. Leibig, A. Katalinic. 2025. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nat. Med.* **31**(3): 917–924.

Feng, Q., J. G. Shanthikumar. 2022. Developing operations management data analytics. *Prod. Oper. Manag.* **31**(12): 4544–4557.

Geng, J., F. Cai, Y. Wang, H. Koeppl, P. Nakov, I. Gurevych. 2024. A Survey of Confidence Estimation and Calibration in Large Language Models. K. Duh, H. Gomez, & S. Bethard, eds. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico, Association for Computational Linguistics, 6577–6595.

Guha, S., S. Kumar. 2018. Emergence of Big Data Research in Operations Management, Information Systems, and Healthcare: Past Contributions and Future Roadmap. *Prod. Oper. Manag.* **27**(9): 1724–1735.

Hopp, W. J., J. Li, G. Wang. 2018. Big Data and the Precision Medicine Revolution. *Prod. Oper. Manag.* **27**(9): 1647–1664.

Huang, J., M. T. Wittbrodt, C. N. Teague, E. Karl, G. Galal, M. Thompson, A. Chapa, M.-L. Chiu, B. Herynk, R. Linchangco, A. Serhal, J. A. Heller, S. F. Abboud, M. Etemadi. 2025. Efficiency and Quality of Generative AI–Assisted Radiograph Reporting. *JAMA Netw. Open* **8**(6): e2513921.

Kapoor, S., N. Gruver, M. Roberts, K. Collins, A. Pal, U. Bhatt, A. Weller, S. Dooley, M. Goldblum, A. G. Wilson. 2025, August 17. Large Language Models Must Be Taught to Know What They Don't Know. arXiv.

Kyung, N., H. E. Kwon. 2022. Rationally trust, but emotionally? The roles of cognitive and affective trust in laypeople's acceptance of AI for preventive care operations. *Prod. Oper. Manag.* **n/a**(n/a).

Laker, L. F., C. M. Froehle, J. B. Windeler, C. J. Lindsell. 2018. Quality and Efficiency of the Clinical Decision-Making Process: Information Overload and Emphasis Framing. *Prod. Oper. Manag.* **27**(12): 2213–2225.

Nori, H., M. Daswani, C. Kelly, S. Lundberg, M. T. Ribeiro, M. Wilson, X. Liu, V. Sounderajah, J. Carlson, M. P. Lungren, B. Gross, P. Hames, M. Suleyman, D. King, E. Horvitz. 2025, July 2. Sequential Diagnosis with Language Models. arXiv.

Shi, P., J. E. Helm, H. S. Heese, A. M. Mitchell. 2021. An Operational Framework for the Adoption and Integration of New Diagnostic Tests. *Prod. Oper. Manag.* **30**(2): 330–354.

Somanchi, S., I. Adjerid, R. Gross. 2022. To Predict or Not to Predict: The Case of the Emergency Department. *Prod. Oper. Manag.* **31**(2): 799–818.

Su, H., Y. Sun, R. Li, A. Zhang, Y. Yang, F. Xiao, Z. Duan, J. Chen, Q. Hu, T. Yang, B. Xu, Q. Zhang, J. Zhao, Y. Li, H. Li. 2025. Large Language Models in Medical Diagnostics: Scoping Review With Bibliometric Analysis. *J. Med. Internet Res.* **27**: e72062.

Terwiesch, C. 2023, January 17. Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course. Mack Institute for Innovation Management, The Wharton School, University of Pennsylvania.

Tu, T., M. Schaekermann, A. Palepu, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, Y. Cheng, E. Vedadi, N. Tomasev, S. Azizi, K. Singhal, L. Hou, A. Webson, K. Kulkarni, S. S. Mahdavi, C. Semturs, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, A. Karthikesalingam, V. Natarajan. 2025. Towards conversational diagnostic artificial intelligence. *Nature* **642**(8067): 442–450.

Wang, C., G. Szarvas, G. Balazs, P. Danchenko, P. Ernst. 2024, October 9. Calibrating Verbalized Probabilities for Large Language Models. arXiv.

Xia, L. 2020. Risk-Sensitive Markov Decision Processes with Combined Metrics of Mean and Variance. *Prod. Oper. Manag.* **29**(12): 2808–2827.

Xia, L., L. Zhang, P. W. Glynn. 2023. Risk-sensitive Markov decision processes with long-run CVaR criterion. *Prod. Oper. Manag.* **32**(12): 4049–4067.

Xiong, M., Z. Hu, X. Lu, Y. Li, J. Fu, J. He, B. Hooi. 2024, March 17. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. arXiv.

Appendix A: Case Selection

The following information was available to the agent (and the humans) at the start of each scenario:

Case "Hypoglycemia"
Ms. Johnson was tutoring Fransico [a student] in mathematics when she suddenly started feeling sick.
Age: 30
Weight: 56 kg
Height: 165 cm
BMI: 20.6

Case "Congestive Heart Failure"
Mr. Clayton has been feeling more shortness of breath than usual. He can hardly get any sleep and needs to be sitting the whole time; otherwise, he cannot breathe.
Age: 66
Weight: 78 kg
Height: 176 cm
BMI: 25.2

Case "Stroke"
Melyssa was found lying on the floor by her daughter at home. She complained about a lack of strength in her

right arm, her speech was confused and it was difficult to understand what she was saying. Her daughter took her immediately to the emergency room.
Age: 75
Weight: 75 kg
Height: 160 cm
BMI: 29.3

Case "Pneumonia"
Mr. Garry has had a fever and a cough with sputum for the past three days. He also complains of chest pain and, because he felt no relief, he decided to go to the Emergency Department.
Age: 58
Weight: 80 kg
Height: 180 cm
BMI: 24.7

Appendix B: Agent Technical Implementation

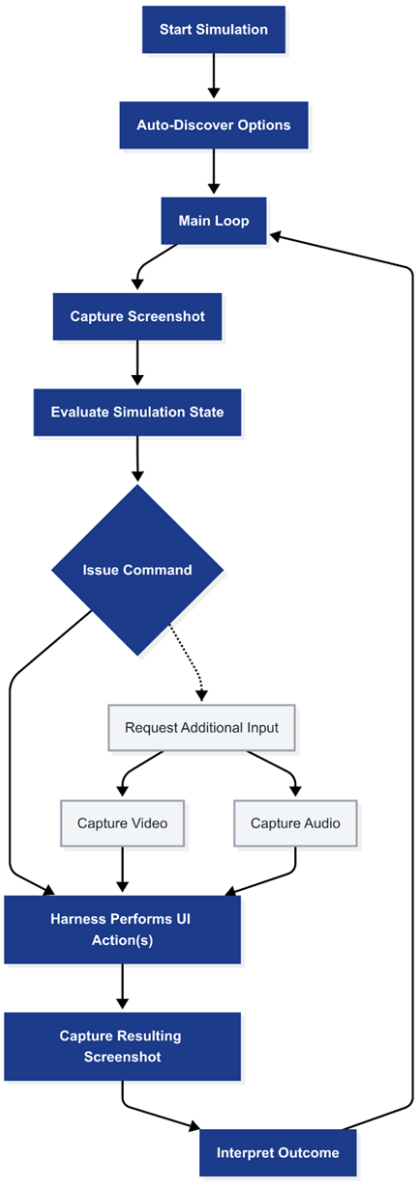
To further streamline operations, we developed small programs that bundle multiple clicks in one action for the agent to take. For instance, to order a specific test, the agent does not have to open each menu, take a screenshot, look at the options and continue. Instead, it can rely on its autodiscovery at the start and instead issue a command to click the test button. A small program handles the submenu navigation. The agent is, however, aware of the submenu structure of the menus and will occasionally issue commands to navigate them to get to the correct option. These commands are ignored in favor of program-based navigation.

At the start of the simulation, the agent performs automatic discovery of all available menu options by navigation through all main UI elements and their submenus to learn what activities are available. This is necessary since many scenarios have custom options, such as dynamic dialogue elements, that are not present in other scenarios. In addition to this dynamic list, a list of fixed operations that are always available (such as opening health records) is provided to the agent in text form.

Once the auto-discovery phase is complete, the agent enters a repeated perception-reasoning-action loop (see **Fig. B1**). In each cycle, the agent first receives an updated screenshot and, when relevant, associated audio or video snippets. Conditional on this multimodal input, the agent uses the LLM to generate a plan for the next step. The plan can include information-gathering actions, such as asking a question, starting a monitoring procedure, or ordering a diagnostic test, as well as interventional actions, such as administering a treatment or calling emergency medical services. The selected actions are passed to the harness, which executes them by invoking the corresponding primitives in the simulation. After execution, the agent receives a fresh screenshot and updated information and verifies whether the intended changes in the simulation state have occurred—for example, whether a monitoring function is active or a medication has been administered by inspecting the new game state (screenshot). If the desired effect has not been achieved, the agent can issue follow-up commands in the next cycle.

This loop continues until one of three case-level outcomes occurs. The simulation succeeds if all case-specific success criteria are met, such as stabilization of vital signs and completion of required calls. It times out if the predefined time limit is reached without satisfying the success criteria, and it fails if a critical medical error is made that severely endangers the virtual patient. The same control loop, harness, and outcome logic are used in all four studies; the subsequent sections describe how they are instantiated in specific clinical scenarios and how performance is evaluated relative to human decision makers.

Fig. B1 Agent workflow



Appendix C: Exclusion Criteria

We apply two exclusion criteria to our human student set. First, we remove sessions played under teacher accounts, which may have been used for demonstration or supervisory runs. Second, we exclude cancelled sessions because they do not reflect full clinical attempts and terminate before a meaningful outcome is reached. After these exclusions, our main analysis sample consists of $N = 14{,}691$ sessions from 8,516 unique students

Appendix D. Supplemental Tables and Figures

This section contains supplemental tables and figures.

Table S1. Ordinary Least Squares Regression Predicting Case Completion Success for Study Case 'Hypoglycemia'. Errors Are Clustered by User ID.

	Case Completion Success		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.949	0.898 – 1.000	< .001
AI	0.051	0.000 – 0.102	0.050
Observations		138	
R^2 / R^2 adjusted		0.023 / 0.016	

Table S2. Logistic Regression Predicting Case Completion Success for Study Case 'Hypoglycemia'. Errors Are Clustered by User ID.

	Case Completion Success		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	2.918	1.879 – 3.957	< .001
AI	17.648	16.579 – 18.718	< .001
Observations		138	
R^2 Tjur		0.023	

Table S3. Ordinary Least Squares Regression Predicting Case Completion Time for Study Case 'Hypoglycemia'. Errors Are Clustered by User ID.

	Case Completion Time (seconds)

Predictors	*Estimates*	*CI*	*p*
(Intercept)	284.731	200.063 – 369.398	< .001
AI	-107.897	-196.685 – -19.110	0.018
Observations		138	
R^2 / R^2 adjusted		0.057 / 0.050	

Table S4. Ordinary Least Squares Regression Predicting Case Completion Success for Study Case 'Pneumonia'. Errors Are Clustered by User ID.

	Case Completion Success		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.779	0.750 – 0.808	< .001
AI	0.104	0.018 – 0.191	0.018
Observations		2235	
R^2 / R^2 adjusted		0.002 / 0.001	

Table S5. Logistic Regression Predicting Case Completion Success for Study Case 'Pneumonia'. Errors Are Clustered by User ID.

	Case Completion Success		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	1.259	1.090 – 1.428	< .001
AI	0.765	-0.041 – 1.572	0.063
Observations		2235	
R^2 Tjur		0.002	

Table S6. Ordinary Least Squares Regression Predicting Case Completion Time for Study Case 'Pneumonia'. Errors Are Clustered by User ID.

	Case Completion Time (seconds)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	619.701	591.252 – 648.150	< .001
AI	-196.884	-276.085 – -117.684	< .001
Observations		2235	

R^2 / R^2 adjusted	0.008 / 0.007

Table S7. Ordinary Least Squares Regression Predicting Case Completion Success for Three Complex Cases. Errors Are Clustered by User ID.

	Case Completion Success		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.776	0.748 – 0.805	< .001
AI	0.198	0.155 – 0.240	< .001
Stroke Case	0.166	0.137 – 0.194	< .001
CHF Case	-0.238	-0.285 – -0.190	< .001
Observations		14793	
R^2 / R^2 adjusted		0.124 / 0.124	

Table S8. Logistic Regression Predicting Case Completion Success for Three Complex Cases. Errors Are Clustered by User ID.

	Case Completion Success		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	1.246	1.078 – 1.414	< .001
AI	2.025	1.303 – 2.748	< .001
Stroke Case	1.557	1.371 – 1.742	< .001
CHF Case	-1.134	-1.363 – -0.904	< .001
Observations		14793	
R^2 Tjur		0.126	

Table S9. Ordinary Least Squares Regression Predicting Case Completion Time for Three Complex Cases. Errors Are Clustered by User ID.

	Case Completion Time (seconds)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	623.170	595.030 – 651.309	< .001
AI	-326.090	-365.105 – -287.075	< .001

Stroke Case	-191.636	-221.001 – -162.271	< .001
CHF Case	208.230	160.788 – 255.671	< .001
Observations		14793	
R^2 / R^2 adjusted		0.118 / 0.118	

Fig. S1 Action Timelines for Three Representative AI Runs in the Hypoglycemia Case (Study 1).

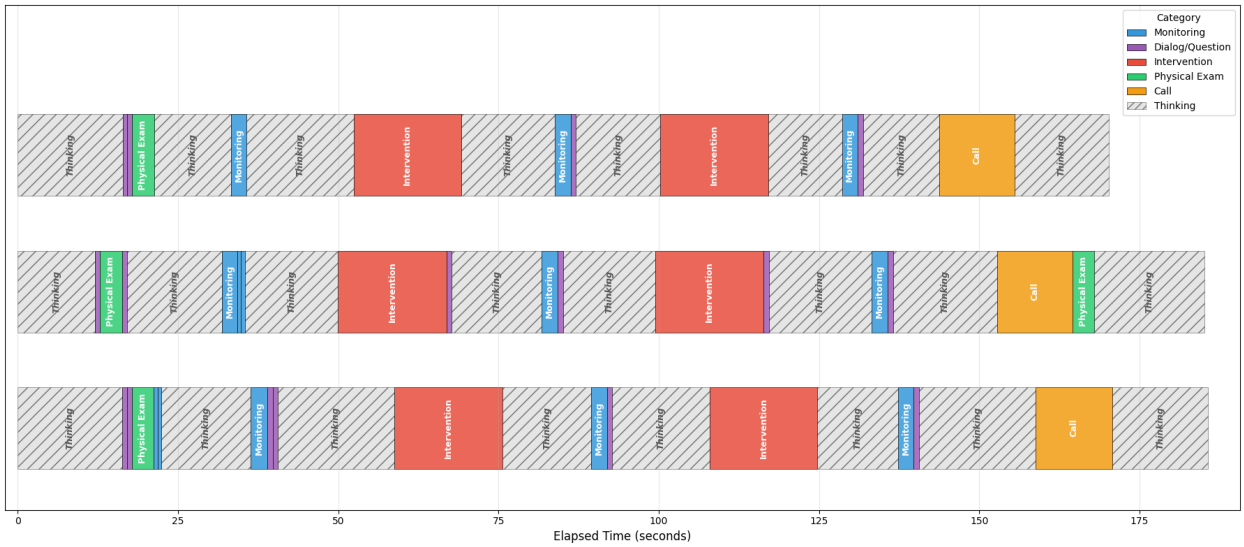


Fig. S2 Action Timelines for Three Representative AI Runs in the Pneumonia Case (Study 2).