npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01896-5

Usability and adoption in a randomized trial of GutGPT a GenAI tool for gastrointestinal bleeding

Sunny Chung[1] ✉, Mauro Giuffrè[1], Niroop Rajashekar[1], Yuan Pu[1], Yeo Eun Shin[2], Simone Kresevic[1,3], Colleen Chan[4], Shinpei Nakamura-Sakai[4], Kisung You[5], Theo Saarinen[6], Allen Hsiao[7,8], Ambrose H. Wong[7], Leigh Evans[7], Terika McCall[9], Rene F. Kizilcec[10], Jasjeet Sekhon[4,13], Loren Laine[1,13] & Dennis L. Shung[1,11,12,13] ✉

Generative AI (GenAI) may enhance clinical decision support systems (CDSS), but its impact on adoption remains unclear. We conducted a simulation-based randomized trial to evaluate whether a GenAI-enhanced CDSS, "GutGPT," improves adoption compared to an AI dashboard in acute upper gastrointestinal bleeding management. Clinical trainees were randomized to either GutGPT or a comparator dashboard across three cases. The primary outcome was Behavioral Intention, from the Unified Theory of Acceptance and Use of Technology (UTAUT). Secondary measures included additional constructs and decision accuracy. A total of 106 participants participated (52 GutGPT, 54 comparator). GutGPT users reported higher Effort Expectancy. Behavioral Intention had no significant difference. Qualitative analysis highlighted trust and workflow concerns. These findings suggest that usability alone is insufficient to drive adoption. As this study was conducted in a simulation without real-world integration or patient outcomes, further studies are needed. (Trial Registration: ClinicalTrials.gov; Identifier: NCT05816473; Registered March 6, 2023).

Artificial intelligence tools, including machine learning (ML) models, have shown growing potential in clinical prediction tasks across a wide range of specialties. In areas such as breast cancer risk assessment, cardiovascular disease prediction, and stroke risk stratification, AI-based models have demonstrated superior or comparable performance to traditional clinical risk scores in retrospective evaluations[1–3]. Despite this progress, the integration into real-world clinical workflows remains limited[4].

A central barrier to clinical adoption of these systems lies not only in model performance, but in their usability, interpretability, and workflow compatibility[5–7]. Clinicians often report difficulties understanding how predictions are generated, concerns about over-reliance on opaque models, and frustration with tools that add rather than reduce cognitive or documentation burden[8–12]. These concerns can be particularly pronounced in acute care settings, where time-sensitive decision-making and cognitive overload are common[13,14]. Consequently, AI tools that offer high predictive accuracy may still fail to be adopted or trusted if their design does not align with clinical priorities or user expectations.

Generative AI (GenAI) introduces new possibilities for clinical decision support through its natural language interfaces and conversational reasoning capabilities. These systems can synthesize clinical guidance, explain predictions in natural language, and respond to user queries in real time[15,16]. However, these benefits remain largely theoretical: few studies have empirically evaluated how GenAI affects clinician behavior, trust, or intention to adopt AI tools in actual clinical workflows. The real-world impact of GenAI based clinical decision support systems (GenAI-CDSS) remains poorly understood, particularly in urgent care scenarios where adoption barriers may be most acute.

[1]Yale School of Medicine Section of Digestive Diseases, New Haven, CT, USA. [2]UC Berkeley, Berkeley, CA, USA. [3]Department of Engineering and Architecture, University of Trieste, Trieste, Italy. [4]Department of Statistics and Data Science, Yale University, New Haven, CT, USA. [5]Department of Mathematics, Baruch College, The City University of New York, New York, NY, USA. [6]Department of Statistics, UC Berkeley, Berkeley, CA, USA. [7]Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, USA. [8]Department of Pediatrics, Yale School of Medicine, New Haven, CT, USA. [9]Yale School of Public Health, New Haven, CT, USA. [10]Department of Information Science, Cornell University, Ithaca, NY, USA. [11]Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA. [12] Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, USA. [13]These authors contributed equally: Jasjeet Sekhon, Loren Laine, Dennis L. Shung. ✉e-mail: sunny.chung@yale.edu; shung.dennis@mayo.edu

To address this gap, we conducted a simulation-based randomized controlled trial to evaluate whether a GenAI-CDSS interface, *GutGPT*, improves usability and adoption compared to a structured AI dashboard for risk stratification in acute upper gastrointestinal bleeding (UGIB) that has been validated previously in the literature[17]. We selected UGIB as our clinical use case due to its high acuity, time-sensitive triage needs. Importantly, UGIB decisions often require rapid integration of guideline-based treatment thresholds, making it an ideal setting to evaluate whether GenAI can facilitate interpretability and real time clinical support[18–20]. Our primary outcome was Behavioral Intention (BI), assessed using the Unified Theory of Acceptance and Use of Technology (UTAUT) framework, which has been widely applied to study health IT and CDSS adoption[21–24]. Secondary measures included other UTAUT constructs, decision accuracy, and qualitative analysis of clinician trust and system interaction.

This study aims to advance understanding of how GenAI interfaces may influence early stage adoption of AI-based clinical decision support systems and to identify the usability, trust, and workflow factors that enable or hinder their integration into high-acuity clinical practice.

Results

After excluding participants who did not complete the study, there were 106 participants, with 54 in the comparator arm and 52 in the intervention arm. We conducted 41 simulation sessions with 2–5 participants each, most commonly involving 2 or 3 clinicians. Most participants reported limited familiarity with AI. Baseline demographics and survey scores were similar across the two arms (Table 1).

Table 1 | Participant demographics and baseline survey scores

Demographics	Dashboard (n = 54)	Dashboard + GutGPT (n = 52)
Female	6 (11%)	9 (17%)
Male	48 (89%)	43 (83%)
Race		
White	26 (48%)	34 (65%)
Black	7 (13%)	6 (12%)
Asian	17 (31%)	7 (13%)
Training Level		
Residency		
PGY1 or PGY2	29 (54%)	20 (38%)
PGY3 or PGY4	12 (22%)	20 (38%)
Medical Student	13 (24%)	12 (23%)
AI Familiarity		
AI/ML Coursework	4 (7%)	3 (6%)
Not at All or Slightly	41 (76%)	42 (81%)
Baseline survey score		
Behavioral Intention	3.5 ± 0.1	3.3 ± 0.1
Performance Expectancy	3.6 ± 0.1	3.4 ± 0.1
Effort Expectancy	3.0 ± 0.1	2.8 ± 0.1
Social Influence	3.7 ± 0.1	3.4 ± 0.1
Facilitating Conditions	2.7 ± 0.1	2.8 ± 0.1
Trust	3.1 ± 0.1	2.9 ± 0.1
Benefit	3.4 ± 0.1	3.3 ± 0.1
Risk	3.2 ± 0.1	3.4 ± 0.1

Participant characteristics were balanced between the two study arms. Baseline survey scores across UTAUT constructs (Behavioral Intention, Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions) and additional measures (Trust, Benefit, Risk) were similar between groups. Survey scores are reported as mean and standard error, calculated as the average of individual item scores within each UTAUT construct on a 1–5 Likert scale, where higher scores indicate stronger agreement with the measured construct.

Primary outcome

BI showed no differences in either between-arm comparisons or within-arm analyses. The median change in BI scores was 0.0 (95% CI: [0.0, 0.3]) in both the intervention and comparator arms (between-arm analysis, $p = 0.657$). These findings suggest that neither system resulted in a change exceeding the predefined clinically meaningful threshold of ±0.5 (based on 0.5 SD of the baseline distribution). To complement this, we analyzed the proportion of participants showing $a \geq 1$-point increase on the Likert scale. A higher percentage of participants in the intervention arm demonstrated increases compared to the comparator arm: 28.8% vs. 19.6% (difference 9.2%, 95% CI: [−7.3%, 25.7%]). However, this difference was not statistically significant, indicating that while individual-level increases were more frequent in the intervention arm, they did not translate into statistically significant between-group effects.

Secondary outcomes

Both groups demonstrated high decision accuracy, with overall averages of 91.7 ± 27.9% in the intervention arm and 92.1 ± 27.2% in the comparator arm. Scenario-specific accuracy for the intervention arm was 85.0 ± 36.6% (Low-Risk), 100.0 ± 0.0% (Medium-Risk), and 90.0 ± 30.7% (High-Risk), while the comparator arm achieved 95.2 ± 22.8%, 80.9 ± 40.2%, and 100.0 ± 0.0%, respectively.

Secondary outcomes suggested meaningful improvements in Effort Expectancy (EE). The intervention arm showed a median change of 0.6 (95% CI: [0.3, 1.0], exceeding the clinically meaningful threshold of ±0.4), while the dashboard arm showed a smaller improvement of 0.3 (95% CI: [0.0, 0.5], threshold = ±0.3). When considering proportions, 40.4% of participants in the intervention arm achieved $a \geq 1$-point increase in EE compared to 17.6% in the comparator arm (difference 22.7%, 95% CI: [5.2%, 40.3%]).

Other UTAUT constructs did not meet the predefined threshold for a meaningful change, except for Performance Expectancy (PE) in the comparator arm (median change: 0.3, 95% CI: [0.0, 0.5]). Minimal changes were observed for Social Influence (SI) and Facilitating Conditions (FC). Trust increased in both arms (median change: 0.2 for intervention, 0.4 for comparator). There were no notable between-arm differences for these constructs in either the median comparison or proportion analysis. Full between-arm comparisons are presented in Table 2.

To further explore potential heterogeneity in system usability and adoption, we conducted subgroup analyses stratified by training level (medical student vs. resident), PGY year (PGY1 vs. PGY2+), sex, and race (White vs. non-White). These analyses examined both median changes and the proportion of participants experiencing ≥1-point improvements in each UTAUT construct. Results are summarized in Supplementary Tables 5–8, including median changes, 95% confidence intervals, and proportions with ≥1-point increases. Notably, medical students and non-White participants reported greater improvements in EE with GutGPT (median difference +1.1 and +1.0, respectively). Across subgroups, the percentage of participants with meaningful increases varied considerably. For example, 66.7% of medical students in the GutGPT arm showed $a \geq 1$-point increase in EE compared to 15.4% with the dashboard alone. While these findings were exploratory and not powered for interaction testing, they provide insight into the contexts and user characteristics that may modulate perceived usability and willingness to adopt GenAI tools in clinical practice.

We also conducted a post hoc analysis comparing participants who interacted with the GPT-3.5 vs. GPT-4 versions of GutGPT, as the underlying LLM was updated partway through the study. Among participants in the intervention arm, GPT-3.5 users showed a greater median improvement in EE (+0.7 [95% CI: 0.2, 1.2]) compared to GPT-4 users (+0.5 [95% CI: 0.00, 1.00]), and a higher proportion achieved ≥1-point increases (45.2% vs. 33.3%). For BI, the median improvement was greater for GPT-3.5 (+0.3 [95% CI: 0.0, 0.7]) than GPT-4 (0.0 [95% CI: 0.0, 0.3]). These findings suggest that observed usability gains were not driven solely by backend model improvements. Full results are presented in Supplementary Table 9.

Table 2 | Comparison of UTAUT constructs, trust, perceived risk, and perceived benefit between the GutGPT + dashboard and dashboard arms

	Median (95% CI)			Change in Likert scale		
	GutGPT + dashboard	Dashboard	Meaningful difference	GutGPT + dashboard	Dashboard	Difference (95% CI)
BI	0.0 (0.0, 0.3)	0.0 (0.0, 0.3)	0.4	28.8%	19.6%	9.2% (−7.3, 25.7)
EE	0.6 (0.3, 1.0)	0.3 (0.0, 0.5)	0.5	40.4%	17.6%	22.7% (5.2, 40.3)
PE	0.0 (0.0, 0.3)	0.3 (0.0, 0.5)	0.4	15.4%	13.7%	1.7% (−12.0, 15.3)
FC	0.1 (0.0, 0.3)	0.0 (0.0, 0.3)	0.4	13.5%	7.8%	5.6% (−6.3, 17.5)
SI	0.0 (0.0, 0.3)	0.0 (0.0, 0.3)	0.3	11.5%	9.8%	1.7% (−10.2, 13.7)
Trust	0.2 (0.1, 0.6)	0.4 (0.2, 0.8)	0.3	26.9%	29.4%	−2.5% (−19.9, 14.9)
Risk	−0.1 (−0.3, 0.0)	−0.1 (−0.4, 0.0)	0.3	0.0%	3.9%	−3.9% (−9.3, 1.5)
Benefit	0.2 (0.0, 0.5)	0.2 (0.2, 0.5)	0.3	9.8%	11.8%	−2.0% (−14.0, 10.1)

This table presents a comparative analysis of Behavioral Intention (BI) and Unified Theory of Acceptance and Use of Technology (UTAUT) constructs across the GutGPT + dashboard and dashboard arms. Median changes with 95% confidence intervals (CIs) are reported alongside the meaningful difference, which was defined as the greater of 0.5 SD of the baseline distribution from either arm. The percentage of participants achieving $a > 1$-point improvement on the Likert scale is shown for each construct, with the 95% CI of the difference between arms provided for further comparison.

Thematic analysis of interviews

Quantitative results indicated that while GutGPT + dashboard improved usability (higher EE), it did not significantly impact BI. Interviews identified key themes: trust, integration challenges, information presentation, the natural language interface, and differences by experience level.

Trust as a barrier and mediating factor. Trust critically influenced participants' reliance on GutGPT, with concerns about accuracy, liability, and the role of human judgment. "I would probably be much more dismissive because I'm using my human experience." Transparency was appreciated, however, noting that "It cites its sources, which is nice." Overall, trust was context-dependent, particularly when outputs aligned with clinical intuition: "If it's being more cautious than I am, I trust it more. But if it's saying something counterintuitive, I'm not sure I would."

EHR integration is crucial. Participants emphasized the importance of seamless EHR integration to enhance usability. "A button on the EMR that makes it just like pop up" would simplify access. Without integration, GutGPT was perceived as disruptive, particularly in fast-paced settings: "It's hard to implement just because [the emergency department is] so fast-paced."

Information overload. GutGPT's verbosity posed challenges, with lengthy outputs seen as impractical. One participant described the responses as "way too long" and suggested, "Provide the directions and then… a button that said like do you want to know why it's choosing this." More experienced clinicians preferred concise outputs over detailed explanations.

Natural language interface and usability. The natural language interface contributed to positive usability perceptions. Participants found it intuitive, noting, "I thought it was easy…just like other chatbot interfaces," and "I immediately knew what to do when I started using it." While usability was praised, concerns about trust and integration persisted.

Experience-level differences. Junior clinicians valued GutGPT as a learning tool, describing it as "exciting" for providing "useful references and clinical guidance." Another noted, "I would use it even just as part of studying." Senior clinicians, in contrast, prioritized efficiency and actionable outputs, explaining, "I think your level of training really matters here."

While the interviews revealed perceptions of GutGPT, analyzing their actual engagement with the system provides insights. Thus, we conducted an interaction analysis of query complexity, usage patterns, and behaviors during the simulation sessions.

Interaction analysis of GutGPT queries

A total of 171 clinician queries were directed to GutGPT. 42.7% of queries explicitly included patient information, while 57.3% were simple and relied on contextual inference. Query styles were predominantly general and exploratory (60.2%), with 39.8% specific to the current patient. Interaction patterns included follow-up queries (26.3%) and rephrased queries (12.9%), reflecting an iterative style. Most queries sought data retrieval (62.6%), while 36.8% aimed to confirm clinical decisions. As part of the query-coding exercise, we manually reviewed all 171 logged queries and confirmed that the response by GutGPT was appropriate for the query's intent; no misrouted questions were identified. Detailed interaction patterns are presented in Table 3.

Discussion

Risk stratification plays a key role in managing UGIB, particularly for identifying very-low-risk patients who may be safely discharged. Although ML models have outperformed traditional clinical scores, their clinical adoption remains limited[17]. In this randomized simulation study, supplementing a structured AI dashboard with a GenAI interface (GutGPT) did not significantly improve BI or clinical decision accuracy. However, the GutGPT arm showed improved EE, suggesting that GenAI may enhance perceived usability without directly influencing adoption or accuracy.

This study intentionally excluded a non-AI comparator, as it was not powered to assess performance against standard care. Instead, it aimed to isolate the impact of GenAI (specifically natural language interaction) on usability and adoption intent. This approach aligns with simulation-based learning methods that evaluate one simulation form against another[25]. Both study arms included AI-CDSS components to focus on interface effects rather than model performance. Future studies comparing AI tools with standard workflows could contextualize clinical impact more fully.

These findings underscore a key tension in clinical AI deployment: usability alone does not ensure adoption. While participants in the GutGPT arm found the tool easier to use and helpful for navigating guidelines, this did not translate into greater adoption intent. Clinician-facing GenAI tools must go beyond improving user experience; they must integrate into clinical workflows and build trust through consistent, transparent, and context-aware behavior. This aligns with UTAUT theory, which posits that PE, trust, and FC often outweigh usability in shaping adoption behavior[24]. Thus, while the observed difference in EE was noted, its lack of impact on BI suggests that it may not be clinically meaningful in isolation within this simulation context. It also reflects a broader theme in clinical AI: that trust, workflow fit, and perceived relevance often outweigh ease of use when shaping BI.

Qualitative feedback highlighted GutGPT's value in surfacing clinical guidelines and clarifying recommendations, particularly for junior clinicians. Participants often rephrased queries or sought confirmation,

reflecting both the tool's utility and a persistent need for human oversight. Senior clinicians noted workflow friction due to lengthy and unpredictable responses. These observations emphasize the importance of tailoring GenAI interfaces to user expertise and clinical context, balancing explanatory depth with decision-making efficiency. Such concerns mirror broader AI adoption challenges, including model opacity, hallucination risk, and the need for design improvements to build user confidence[26]. One promising approach could be an interface toggle between concise summaries and more detailed explanations, allowing users to self-select the level of depth needed. Such adaptive features may improve workflow compatibility while supporting both novice and expert clinicians. Figure 1 summarizes how these findings map onto the UTAUT framework, highlighting key facilitators and barriers to adoption identified through thematic analysis.

Table 3 | Interaction analysis of clinician queries to GutGPT during simulation

	Measure		Reliability
Patient data	No inclusion	57.3%	0.84
	Included data	42.7%	
Query style	General/ exploratory	69.2%	0.72
	Specific to patient	39.8%	
Interaction patterns	Follow up queries	26.3%	0.80
	Rephrased queries	12.9%	0.78
Query purpose	Data retrieval	62.6%	0.81
	Confirmatory	36.8%	
Model usage	Average query length	12.7 ± 2.6 words	
	Average response length	149.1 ± 112.4 words	
	Average queries per session	6.3	
Token usage	Average input tokens per query	19.7	
	Average output tokens per query	207.8	
	Cost per session	~$0.013	

This table summarizes the interaction patterns, query complexity, and model usage within the GutGPT + dashboard arm ($n = 52$). Queries were classified by whether they included patient data or not. Query styles were predominantly general/exploratory. Follow-up queries and rephrased queries demonstrated iterative interaction patterns. Most queries sought data retrieval. Overall, these codes demonstrated high inter-coder reliability. Average clinician query length was concise, while GutGPT responses were detailed. Token usage analysis indicated 19.7 tokens per input query and 207.8 tokens per output response, resulting in an estimated session cost of ~0.013.

While we did not formally score all 171 simulation queries in this study, GutGPT's responses to these same questions were recently evaluated in a related UGIB dataset using expert grading and a reward model[27]. This companion study provides benchmarked estimates of guideline fidelity across model configurations and informs ongoing refinement. Although it focused on output quality rather than routing accuracy, it confirmed that clinician queries were generally handled appropriately. Our own manual audit also suggested high face validity of routing logic, though we did not conduct a formal quantitative benchmark (e.g., independent dual annotation with κ statistics). Future work will include a larger, blinded accuracy study to assess sensitivity, specificity, and inter-rater reliability.

EHR integration emerged as a critical determinant of adoption. Participants cited platform-switching and lack of embedded workflow as reasons for disengagement, mirroring broader findings from implementation science around digital tool uptake[28]. Embedding GenAI tools directly within EHR systems thus may improve clinician trust and usability by streamlining access and reducing friction. However, this integration raises additional challenges, including interoperability, governance, and resource constraints, which must be addressed to support sustained development[16,29,30].

Based on token usage logs, the average cost per simulation session was approximately $0.013. GutGPT was designed to run efficiently on standard laptops for testing and development. However, deployment in real-world clinical environments, particularly those involving real patient data, would require more robust compute infrastructure and security controls. This includes locally hosted inference, audit logging, and granular access controls to mitigate risks of inadvertent data disclosure. These infrastructure requirements are essential for regulatory compliance, data protection, and

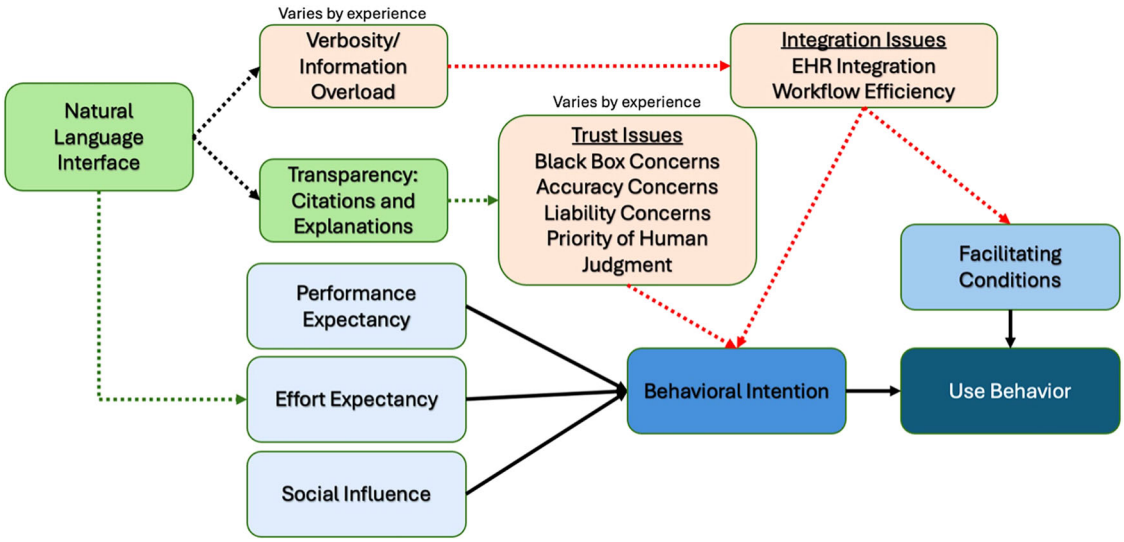

Fig. 1 | Conceptual framework integrating qualitative themes with the UTAUT model. This figure maps the facilitators and barriers influencing clinician adoption of GutGPT onto the Unified Theory of Acceptance and Use of Technology (UTAUT) framework. Blue boxes represent the original UTAUT constructs assessed in the study. Green boxes reflect emergent facilitators of adoption identified through qualitative analysis (e.g., guideline transparency, reduced cognitive burden), while red boxes denote barriers (e.g., verbosity, variable response quality, lack of integration). Solid black arrows reflect established relationships between UTAUT constructs. Green and red arrows represent positive and negative influences derived from participant feedback. Dotted arrows indicate exploratory or experience-dependent pathways not explicitly modeled in the original UTAUT framework. This figure is intended as a conceptual synthesis of study findings, combining qualitative themes with theoretical adoption constructs to highlight areas for future system refinement. It should not be viewed as a formal causal model.

clinician trust. Similar deployment patterns have been successfully implemented for EHR-integrated CDSS systems, including tools for antimicrobial stewardship[31] and EHR-based/AI-based CDSS for antibiotic therapy in sepsis patients[32] demonstrating the feasibility and impact of embedding decision support directly into clinical workflows.

This study coincided with a transition in LLM architecture, as OpenAI upgraded from GPT-3.5-Turbo to GPT-4 during the trial. Prompt structures, interface design, and guidance language were held constant. Post hoc analysis showed that participants who used GPT-3.5 reported greater median improvements in both EE and BI than those who used GPT-4. Although exploratory and not powered for formal inference, this suggests that usability gains were not solely due to model performance, but also reflected interface design and the value of conversational interaction. Over time, clinician trust may evolve with cumulative exposure to updated models, underscoring the need for longitudinal studies of adoption and system drift in clinical AI tools. However, these effects require further study, as the model transition was neither randomized nor controlled. The forced model switch also underscores the risks of relying on proprietary, closed-source LLMs in clinical contexts. Loss of version control and API stability may compromise reproducibility and long-term reliability. Future iterations of GutGPT should explore these alternatives to support greater flexibility, data privacy, and regulatory alignment, key priorities for health systems adopting GenAI tools in privacy-sensitive environments.

This study has several limitations. First, while the simulation-based design allowed for controlled evaluation of human-AI interaction within an established educational framework, it does not fully replicate the complexities of real-world clinical settings[33]. Simulation-based learning is a widely accepted approach for enhancing clinical skills in a structured environment[34], and future work should focus on adapting the system based on lessons learned here for integration into actual clinical care pathways.

Second, this study focused specifically on the management of UGIB, a choice informed by team expertise and the availability of robust, guideline-supported decision pathways. However, this targeted scope may limit generalizability to other clinical domains. The use of three standardized UGIB scenarios was based on feasibility constraints, with each simulation designed to be completed within a two-hour session, comparable to other studies, such as a prior simulation trial evaluating an AI-powered physician agent training nursing students in sepsis care[35]. Future studies with expanded resources should explore a broader array of clinical vignettes to assess generalizability.

Third, although this is the largest study to date evaluating AI's impact on educational outcomes for physician trainees using simulation-based learning[36], the participant sample was relatively small and predominantly male, which would limit generalizability. Participants were scheduled in small-group sessions. While scenario order was randomized between sessions, there remains a theoretical risk of cross-group contamination if participants discussed the cases or shared impressions of the tools outside of the sessions. Group based simulation was intentionally used to reflect real-world inpatient and emergency medicine workflows, where clinical decisions are often made collaboratively during rounds, handoffs, and team-based consults. While this design improves ecological validity, it may also introduce peer influence or consensus bias that could affect decision accuracy or perceptions of the system.

Group size variation represents an additional potential source of bias, as the study measured individual-level UTAUT outcomes following group-based decision-making. Most simulation teams included 2 or 3 participants, with only one group of 5 and a few groups of 4, limiting the feasibility of meaningful subgroup analysis by team size. Nonetheless, peer dynamics and social desirability bias may have influenced individual responses. Future studies should compare individual versus group-based deployment models using larger, more balanced samples to better understand these dynamics.

Additionally, the mid-study transition from GPT-3.5 to GPT-4 introduces a potential confound that may affect internal validity. Although interface elements and prompt structures were held constant, the change in model backend was not randomized and could have influenced participant perceptions or outcomes. At the same time, such backend model updates are common in real-world software and GenAI deployment, where systems are continuously updated and iterated upon. This design choice, while a limitation from a methodological standpoint, also reflects the dynamic nature of clinical AI tools in practice. Future studies may consider prospectively evaluating model version changes within a randomized framework to balance internal validity with real-world applicability.

Finally, our primary and secondary outcomes relied on subjective, self-reported measures guided by the UTAUT[24], a validated framework for assessing user intent. While UTAUT has been used to evaluate provider adoption of electronic medical records, conversational agents in healthcare, and clinical guideline applications[22,37,38], it does not assess clinical effectiveness. Additionally, time-on-task was not systematically recorded, which limited our ability to assess potential efficiency differences between arms. Because simulation environments do not enable real-world patient outcomes, future clinical deployment studies of GutGPT should assess objective clinical outcomes. Specifically, the primary outcome should be the proportion of discharged patients who do not experience adverse clinical events requiring hospital-based intervention, defined as transfusion, urgent procedures, or 30-day all-cause mortality[19]. Secondary outcomes in future clinical studies will include both adherence to guideline-recommended care (e.g., continuation of aspirin for secondary prevention, appropriate antibiotic use in suspected portal hypertensive bleeding)[39] and operational metrics such as task efficiency (e.g., task completion time, interaction duration), documentation accuracy, decision latency, and downstream patient outcomes.

In summary, this study is the first randomized evaluation of GenAI in clinical decision support performed in a clinical simulation setting. While generative interfaces may enhance perceived usability, adoption is shaped by a complex interplay of trust, workflow integration, and clinician experience. Simulation-based testing can offer an important lens for uncovering these dynamics in early development, helping inform safer and more effective integration of GenAI tools into clinical workflows.

Methods

Machine learning model dashboard

The ML Model Dashboard takes a published ML model that outperforms other clinical risk scores for acute gastrointestinal bleeding with additional elements to promote ML interpretability derived from the model, including a contextual measure and visualization of variables that contribute towards the prediction. This dashboard was developed iteratively with feedback from board-certified gastroenterologists, data scientists, software engineers, and human factors experts. The initial dashboard was tested with trainees via an online-based preliminary study with 10 physician trainees (mixture of medical students, internal medicine residents, and gastroenterology fellows) with feedback integrated into the final dashboard deployed in the study[40]. The interface shows the risk prediction for patients presenting with UGIB for in-hospital intervention and/or 30-day mortality (see Supplementary Fig. 1). Key features included dynamic input adjustment, which enabled participants to modify relevant covariates and instantly observe the impact on the prediction[41]. This dashboard is based on a previously validated EHR-integrated gradient boosted decision tree model developed by Shung et al. for real-time risk stratification of patients presenting with overt gastrointestinal bleeding[17]. The model was trained on 2546 patients and externally validated on an additional 926 patients, achieving an area under the receiver operating characteristic curve of 0.92. Compared to traditional risk scores such as the Glasgow-Blatchford Score and Oakland Score, it identified substantially more very-low-risk patients at a 99% sensitivity threshold (37.9% vs. 18.5% for GBS), supporting its use as a clinically relevant comparator system.

The dashboard displayed the predicted risk alongside contextual information, including the comparison of the current patient's risk to that of their 100 most similar patients. This similarity metric is motivated by the

interpretation of Random Forests as adaptive nearest neighbor algorithms, as proposed by Lin and Jeon (2006) and Stone (1977)[42,43]. In this view, observations falling into the same leaf node are considered proximate within feature space. For a given patient, two observations are assigned a similarity score of 1 if they fall into the same terminal leaf node within a decision tree, and 0 otherwise; the final similarity score is averaged across all trees in the ensemble. This structure-adaptive definition of similarity enables intuitive, data-driven risk contextualization without requiring explicit distance functions. Although related work has applied this principle to detect extrapolation and identify low-confidence predictions[44], our application repurposes the underlying mechanism to surface clinically meaningful comparators from the training distribution, thereby enhancing interpretability.

To further enhance interpretability, the dashboard included visual tools such as Individual Conditional Expectation (ICE) plots, illustrating how variations in individual variables influenced the model's prediction for a given case[45]. Unlike SHAP plots, which have been used for interpretability in previous studies with ML models for acute gastrointestinal bleeding to give a general understanding of how variables affect the model predictions in aggregate[46,47], ICE plots have the potential to individualize explanations by providing a graph of individual observation trajectories for each patient showing the functional relationship between the predicted response and variable[45]. This approach has also been applied to promote interpretability for ML models in other clinical settings as well[48,49]. Feature distribution plots were also provided to compare the patient's input values for high-impact features to distributions across the training cohort[50,51]. During the simulation study, patient data were pre-filled from standardized case vignettes but remained editable, enabling participants to explore system functionality without requiring manual data entry. This design closely mimicked the clinical workflow of an EHR-integrated decision support tool and served as a realistic comparator for evaluating the added value of a GenAI interface.

A fully annotated version of the ML dashboard interface is provided in Supplementary Fig. 1, highlighting its adjustable input fields, similarity-based predictions, and interpretability panels.

GutGPT Interface

The dashboard was integrated with a multi-large language model (LLM) add-on system named "GutGPT" using OpenAI's GPT-3.5 ("gpt-3.5-turbo-16k") or GPT-4 ("gtp-4-1106-preview") when we updated the interface. The core of GutGPT decision-making process is implemented in the classify_response() and generate_response() functions in the prompts.py module. When a clinician submits a query, it is first processed by a parser LLM ("gpt-3.5-turbo-16k" or "gtp-4-1106-preview" with the latest version) which employs a few-shot learning approach using a knowledge base of 61 exemplary query-classification pairs structured in the context_multiclassifier.txt file. This parser LLM assigns confidence scores (0–100) across six distinct categories:

- A: Risk prediction for hospital-based interventions
- B: Variable importance in prediction models
- C: Evidence-based clinical management guidelines
- D: General gastrointestinal bleeding information
- E: Tool functionality and methodology
- F: Unrelated or out-of-scope queries

Categories are not mutually exclusive, allowing for multi-dimensional classification of complex clinical questions. Classification is regulated by the following system prompt:

"You are a classifier for clinician queries about UGIB. Using the few-shot examples in "context_multiclassifier.txt", assign a confidence score (0–100) to each of the six categories below. Include every category in your response—even if the score is zero—and format exactly as shown:

- Risk prediction for a hospital-based intervention for a potential GI bleeding patient.
- Variable importance of predicted risk of requiring a hospital-based intervention for a patient. C) Guidelines for patient management options according to current guidelines.
- Symptoms, causes, and other general questions of GI bleeding.
- General questions of this tool and goals.
- Unsure or not a question related to GI bleeding."

The confidence scores are returned as a structured dictionary, which then determines the subsequent routing logic. The routing LLM with the highest assigned score is then chosen by calling the appropriate specialized function:

- For category A (risk prediction): generate_response_mlmodel()
- For category B (variable importance): generate_response_mlmodel()
- For category C (clinical guidelines): generate_response_guidelines()
- For categories D, E, F (general questions): generate_response_general()

The following hyperparameters were applied across all LLMs: *temperature* = 0.8, *top_p* = 0.5, and *max_tokens* = 500. Answer generation was conducted in a local Python (version 3.11) environment, utilizing the OpenAI API (version 1.17) for seamless interaction with the OpenAI's models.

The system consists of three distinct LLMs, each with a specific function (Fig. 2). The first LLM functions as a parser that directs user queries to the appropriate sub-models through few-shot prompting, using 61 examples of question-answer pairs (Supplementary Note 2) to classify queries across six categories (A–F) with confidence scores from 0 to 100. Based on the highest confidence score, queries are routed to one of three pathways: (1) for risk prediction (Category A) or variable importance (Category B) queries, the interpretability LLM provides patient-specific risk assessments - this LLM has access to the model prediction and the names and values of the top three contributing features, but not the full set of raw input variables (e.g., all labs or vitals), predicted outputs, and key variables contributing to the individual patient's risk prediction; (2) for clinical management queries (Category C), the guideline-enhanced LLM generates evidence-based recommendations by retrieving information from a vector database on relevant clinical guidelines; (3) or general information (Categories D, E, F), the parser LLM itself handles the response generation.

Interaction with ML model dashboard. The integration between the clinical decision support dashboard and Gut-GPT is implemented through a structured data pipeline that enables contextualized clinical conversations. In particular, when a user's query is classified as Category A (risk prediction) or Category B (variable importance), the system capture all the input values from the user interface elements (sliders, dropdowns, checkboxes, similarity index, risk of hospital based intervention, and top-3 most influential features from PDP-ranking) corresponding to patient features and transform them into a standardized data frame format using the values_for_df() function and creates a patient profile data frame that matches the structure expected by the predictive model. The run_chatbot() callback function assembles the following contextual information to send to the LLM: user's clinical question (user_input), top risk-contributing features (high_risk_features), intervention statistics among 100 most similar patients (num_interventions), and risk of hospital-based intervention (model_prediction). Complete patient feature values are not retrieved for privacy concerns.

LLM responses for dashboard interaction are regulated by the following system prompt:

"You are GutGPT, a specialized chatbot assistant, linked to a dashboard that contains relevant information regarding patient's data and outputs from a ML model that predict the risk of hospital-based intervention in patients with UGIB. Your main role is to reply to a user query reported here: "user_input" aimed at retrieval of the following information: the model prediction ranging from 0 to 1 for hospital-based intervention "model_prediction" and how many patients among the 100 most similar received hospital-based intervention "num_interventions". In addition, you will receive the name of the three features that contributed the most to the final risk predicted by the model "high_risk_features"."

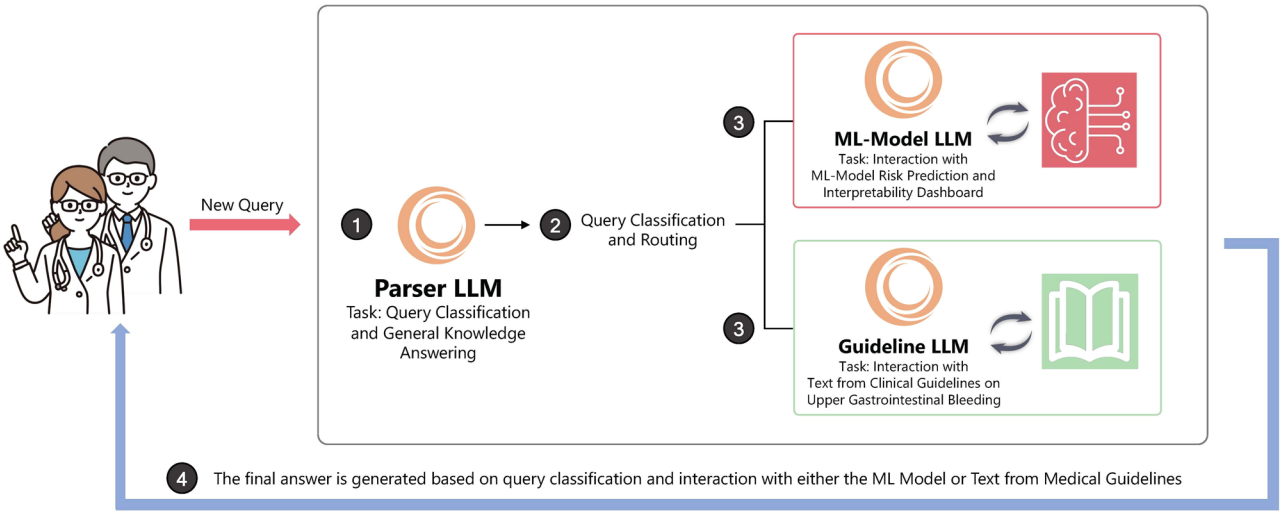

Fig. 2 | Schematic of the GutGPT architecture. This figure depicts GutGPT's three-tiered LLM architecture for gastrointestinal bleeding decision support. The workflow begins with a user new query (step 1), processed by the Parser LLM, which classifies queries across six categories using confidence scores (step 2). Based on classification, queries are routed (step 3) to either the ML-Model LLM for risk prediction and feature importance, the Guideline LLM for evidence-based recommendations, or handled directly by the Parser LLM for general information. The final response (step 4) integrates either machine learning insights or guideline recommendations, ensuring contextually appropriate clinical decision support while maintaining the integrity of the underlying predictive model. *Note: GutGPT never recalculates or changes the ML model's prediction. It only explains it.*

To verify response consistency, we conducted a reproducibility test across 20 repeated runs per query for each of the three simulation scenarios. GutGPT consistently returned the correct model-predicted risk and values for the top three contributing features. Full results are provided in Supplementary Table 11.

Interaction with clinical guidelines. When a query is classified as Category C (guidelines for patient management), GutGPT activates a specialized knowledge retrieval system to provide evidence-based clinical recommendations from the major Northern American UGIB Guidelines[20,52,53]. This functionality is implemented through the generate_response_guidelines() function, which establishes a direct connection to structured guideline documents. We reformatted the original documents from raw PDF formats to ones suitable for LLMs. This involved converting all information, both text and non-text, into a textual format, creating a coherent structure across all guidelines, and dividing each document into three macro sections: pre-endoscopic, endoscopic, and post-endoscopic management. During the usage of "gpt-3.5-turbo-16k", we also introduced artificial separator symbols (\n\n\n\n—four consecutive newlines) to create semantically meaningful boundaries at the paragraph level. This deliberate preprocessing step ensured that chunks would align with natural divisions in the clinical content rather than arbitrary character counts. When the vector database was created, chunks were separated precisely at these symbol boundaries using a specialized character text splitter. This chunking approach preserves the integrity of clinical recommendations that often span multiple paragraphs, ensuring that related content remains cohesive within the retrieval system. Whereas, during the usage of "gpt-4-1106-preview", guidelines were chunked at the document level and not at the paragraph level.

Each document chunk is transformed into a high-dimensional vector using OpenAI's embedding model text-embedding-ada-002, that generates 1536-dimensional vectors that capture the semantic nuances of medical terminology and clinical concepts, enabling precise retrieval of relevant guideline sections. The system utilizes the Chroma vector database to store and retrieve guideline embeddings. This creates a persistent database in the specified directory (db_all_by_section), allowing the system to reuse embeddings across clinical sessions without recalculating them, thereby reducing latency in the clinical workflow. The retrieval mechanism employs cosine similarity as implemented in the Chroma database. Cosine similarity calculates the cosine of the angle between query and document vectors, returning a score between −1 and 1, with 1 indicating perfect similarity. This metric is particularly effective for medical text due to its ability to focus on semantic direction rather than magnitude, making it robust to variations in terminology length and frequency. The selective retrieval approach was designed to identify the five most relevant guideline chunks (when using "gpt-3.5-turbo-16k") or the most relevant guideline document (when using "gpt-4-1106-preview").

LLM responses for guideline-based responses are regulated by the following system prompt:

"You are GutGPT, a specialized chatbot assistant, expert in UGIB. You will be interacting with physicians from the emergency department or digestive disease section at different training levels (student/resident/fellows/attendings). Your role is to reply to the following user's query: "user_input" by following recommendations reported in Northern America Guidelines for Variceal and Non-Variceal UGIB as reported here: "retrieved_guideline_text". The guidelines have been reformatted following this structure:

- ***Paragraph/Subparagraph Title:** Identifies the main topic or section header, setting the context for the subsequent content.*
- ***Paragraph/Subparagraph Recommendations:** Lists bullet-pointed directives or guidelines that serve as actionable items.*
- ***Paragraph/Subparagraph Text:** Offers evidence supporting the recommendations mentioned above.*

Based on the complexity of the queries, you will either provide the information in a list format, with each point on a separate line and bolded for emphasis (when answering queries involving multiple steps in patient management) or simple text (when answering more simple queries). When providing answers you will provide only the recommendations (without mentioning the evidence for each recommendation unless requested). You will need to provide concise answers with a limit of 150–200 words."

Other types of responses. When a query is classified as Category D, E, or F, GutGPT activates a general response answering mechanism implemented through the generate_response_general() function that is activated in the original parser model.

LLM responses for other type of responses are regulated by the following system prompt:

"You are GutGPT, a specialized chatbot assistant, expert in UGIB. You will be interacting with physicians from the emergency department or digestive disease section at different training levels (student/resident/fellows/attendings). Based on the nature of the query, respond according to these guidelines: When the user is asking about symptoms, causes, and general gastrointestinal bleeding questions:

- *Provide evidence-based information about GI bleeding pathophysiology, etiology, symptoms, diagnosis, and epidemiology.*
- *Focus on clinical manifestations, risk factors, diagnostic approaches, and general management principles.*
- *When discussing these topics, maintain clinical precision and use terminology appropriate for healthcare professionals.*

When the user is asking about tool-specific questions:

- *Clarify that the system uses a ML model trained on historical patient data to predict the risk of hospital-based interventions for UGIB patients.*
- *Emphasize that the system has three components: risk prediction interpretation, feature importance explanation, and clinical guideline integration.*
- *Note that the system cannot access physical examination findings or other information not captured in the electronic health record.*
- *Explain that you can provide relevant recommendations based on Northern America's most recent guidelines on Variceal and Non-Variceal UGIB.*

When the user is asking about potential unrelated queries:

- *Determine if the query can be reasonably connected to UGIB or the tool's functionality.*
- *If it can be connected, redirect the conversation toward relevant UGIB topics.*
- *If the query is entirely unrelated, politely state: "As a specialized UGIB assistant, I'm focused on providing information related to gastrointestinal bleeding, its management, and this clinical decision support tool. I don't have expertise in unrelated medical or non-medical topics."*
- *Do not provide information on topics unrelated to UGIB or the tool's functionality.*

Keep responses concise, clinically relevant, and focused on information that would be useful to a medical professional managing a patient with potential or confirmed UGIB."

We have tested the factual accuracy of the system and found that the system accuracy is 84.6% for direct recommendation retrieval from guidelines and 80.3% when tested on a subset ($n = 117$) of questions asked during the scenario simulation in a study that is currently in press at *npj Digital Medicine* (Mauro et al.).

An annotated screenshot of the GutGPT interface is provided in Supplementary Fig. 2, and a recording of the interface can be seen in Supplementary Movie 1.

Ethics approval, informed consent, and protocol

This study was reviewed and approved by the Yale University Institutional Review Board (IRB #2000034521) on March 6, 2023, prior to study recruitment and was conducted in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants prior to their participation. Participation was voluntary with the option to withdraw at any time. The protocol can be found ClinicalTrials.gov, Identifier: NCT05816473.

Study design, participants, and randomization

This RCT compared a GenAI-enhanced clinical decision support system (GutGPT + Dashboard) with a system without GenAI (Dashboard) in a high-fidelity simulation center. Participants were internal medicine/emergency medicine residents and medical students from our institution's medical school (July 2023–November 2024), excluding those with prior experience in the simulation. Informed consent was obtained (Fig. 3).

This study was supported by an NIH K23 grant awarded to the senior author. The funding agency had no role in the design of the study, collection, analysis, or interpretation of data, or preparation of the manuscript. All authors had access to the study data and reviewed and approved the final manuscript.

Participants were scheduled in small-group simulation sessions, each consisting of 2–5 individuals. Although survey responses were completed individually, the simulation component was explicitly designed to reflect team-based clinical decision-making. Within each session, participants collaborated to determine the appropriate disposition for each case scenario. This design was chosen to replicate real-world inpatient and emergency care environments, where clinical decisions are frequently made collaboratively during team-based workflows, such as bedside rounds, handoffs, and consult discussions. To maintain consistency and minimize treatment contamination, randomization occurred at the session level, with each group assigned via a computer-generated allocation sequence to either the Dashboard or the GutGPT+Dashboard arm.

Due to visible differences between systems, blinding was not feasible. However, proctor behavior was standardized using scripted responses to minimize bias. All sessions followed a consistent structure: (1) pre-simulation surveys and orientation, (2) three clinical scenarios, and (3) post-simulation surveys and semi-structured interviews. Simulation sessions were capped at 90 min.

This study is reported in accordance with the CONSORT-AI guidelines for clinical trials involving artificial intelligence interventions, and a completed CONSORT-AI checklist is included in the Supplementary Table 4)[54].

Simulation scenarios

Each session involved three scenarios simulating UGIB cases across low-, medium-, and high-risk categories. Scenarios were based on real patient cases from our institution, de-identified and modified for educational use by attending gastroenterologists and emergency medicine faculty. Each case included clinical data such as presenting symptoms, vitals, labs, and relevant history. Scenario order was randomized between sessions to control for learning effects and fatigue. For those participants who were randomized to GutGPT, the model did not receive all raw inputs directly; it interpreted the ML dashboard's output, including the risk score and top contributing features.

Participants collaboratively reviewed each case and determined the appropriate clinical disposition (e.g., discharge, ward admission, ICU) based on the information provided and decision support tool assigned to their group. Relevant case details are provided in Supplementary Note 1.

Interventions

GutGPT + Dashboard Arm (Intervention Group). Participants used a web-based interface combining a validated ML risk prediction dashboard with a GenAI system called GutGPT. The dashboard displayed patient-specific risk predictions (0–1 probability of in-hospital intervention or 30-day mortality) based on a previously validated EHR-trained gradient boosted model, along with top contributing features, ICE plots, and a similarity panel showing the 100 most similar patients. Participants could edit patient values and observe real-time prediction updates.

GutGPT was layered on top of this dashboard, enabling participants to type free-text queries and receive natural language explanations or guideline-based recommendations. The GenAI system consisted of a three-LLM architecture: (1) a parser that classified clinician queries, (2) a model that interpreted ML predictions and contributing features, and (3) a guideline-enhanced model that retrieved evidence-based recommendations from North American UGIB guidelines using retrieval-augmented generation. Participants were oriented to this interface before the simulation and could use GutGPT and the dashboard interchangeably.

During the study, OpenAI deprecated GPT-3.5-Turbo, transitioning to GPT-4 in January 2024. Although unplanned, this update reflects real-world conditions where AI systems are routinely

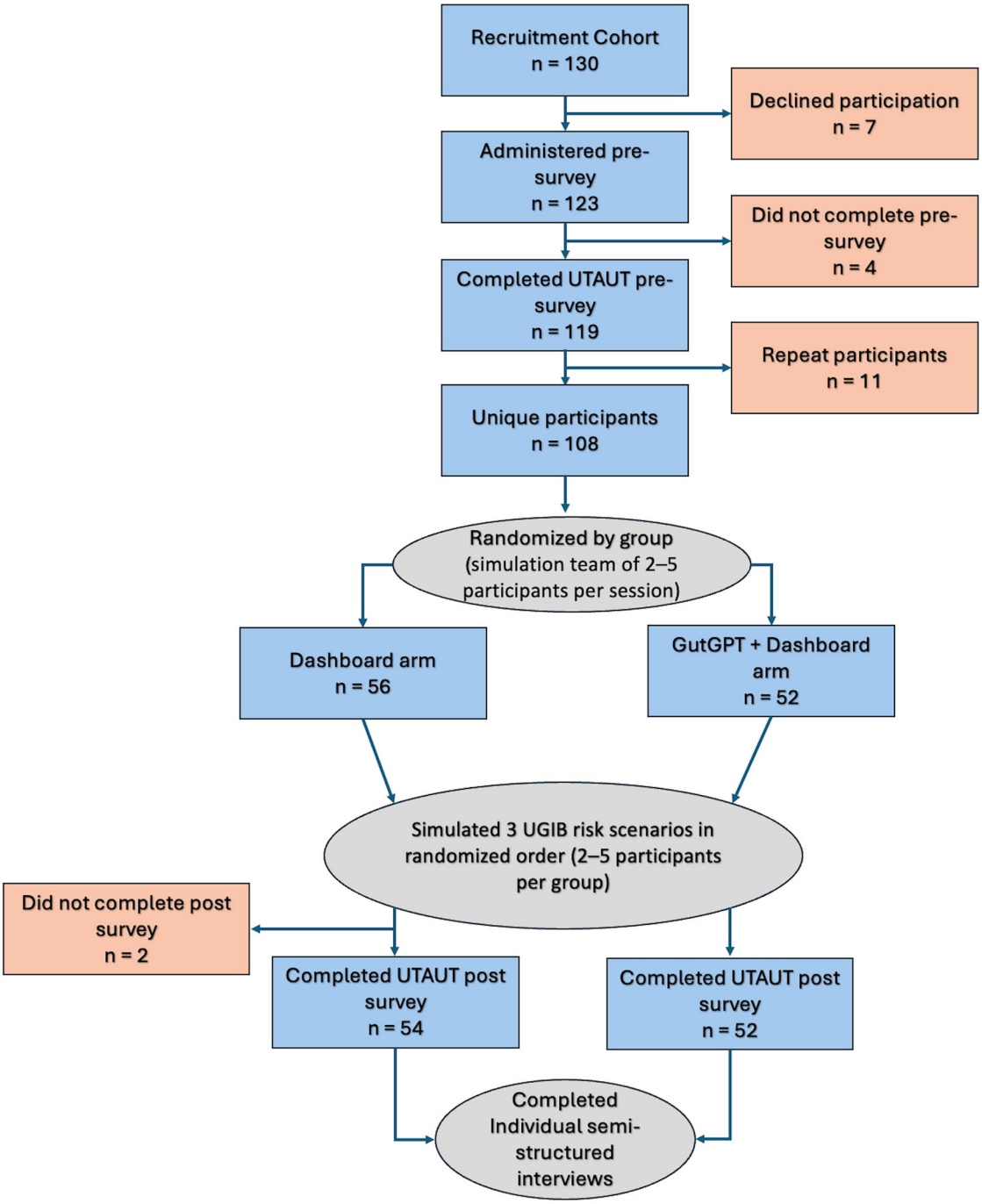

Fig. 3 | CONSORT-style flow diagram of study recruitment, randomization, and data collection. Participants ($n = 130$) were recruited from internal medicine and emergency medicine training programs and invited to participate in a simulation-based study evaluating a generative AI clinical decision support tool (GutGPT). After exclusions due to non-completion of the pre-survey or repeat participation, 108 unique individuals were randomized by group (2–5 participants per session) to either a standard AI dashboard ($n = 56$) or the GutGPT-enhanced dashboard ($n = 52$). Each group completed three UGIB scenarios in randomized order, followed by individual post-surveys and semi-structured interviews. The final analytic sample consisted of participants who completed both the simulation and the UTAUT post-survey ($n = 106$). All 106 participants also completed an individual semi-structured interview. Blue boxes represent participant flow and retention. Orange boxes denote participant exclusions. Gray ovals denote procedural steps (randomization and simulation).

upgraded. Evaluating GutGPT under these evolving conditions enhances the generalizability of our findings, reflecting its relevance in dynamic clinical environments.

Dashboard Arm (Comparator Group). Participants used the same ML dashboard described above, but without access to GutGPT. They interacted with patient cases using structured interface features, such as ICE plots, top features, and similarity-based risk distribution. The layout and input functionality were identical to the intervention group. Free-text queries and natural language responses were not available.

All participants received a standardized training session prior to simulation and were blinded to the assignment of other arms.

Development of survey-based measurement tool

The UTAUT survey[24] was adapted to evaluate clinician acceptance, including PE, EE, SI, FC, and BI, with additional constructs for Trust,

Perceived Risk, and Perceived Benefit. Survey items were reviewed by clinicians and a human factors/information science expert to ensure alignment with study objectives and usability context. All survey questions were required and completed electronically. A mapping of survey items to UTAUT constructs and the additional constructs is provided in the Supplementary Table 1. We selected the original UTAUT model over UTAUT2 because it is more commonly validated in clinical and provider-facing contexts. UTAUT2 introduces constructs like hedonic motivation and price value that are less relevant to professional adoption of clinical tools[55].

A pilot test with 10 clinicians assessed internal consistency using Cronbach's alpha across UTAUT constructs (PE: $\alpha = 0.89$ [0.84–0.92], EE: $\alpha = 0.83$ [0.76–0.88], SI: $\alpha = 0.90$ [0.84–0.92], FC: $\alpha = 0.74$ [0.64–0.82], BI: $\alpha = 0.90$ [0.85–0.93]) and added constructs (Trust: $\alpha = 0.87$ [0.83–0.91], Benefit: $\alpha = 0.88$ [0.83–0.91], Risk: $\alpha = 0.71$ [0.60–0.80])[56].

Quantitative data

The study evaluated the primary outcome of the UTAUT construct, BI, the validated measure of technology acceptance, focusing on its change pre and post simulation[15].

Secondary outcomes included EE (perceived ease of use), PE (perceived usefulness towards clinical goals), SI (perceived encouragement from others), FC (perceived adequacy of resources), trust, perceived risk, and perceived benefit. Clinical decision accuracy, operationalized as the proportion of scenarios in which participants made a guideline-concordant clinical disposition, was also measured.

Qualitative data

Semi-structured interviews. Participants also participated in one-on-one semi-structured interviews, exploring their experiences, perceptions, and feedback on workflow integration of the system they used. Participants were also asked about their willingness to use the system in real-world settings. Interviews were audio-recorded with consent for analysis. Example questions of the interview template are provided in the Supplementary Table 2.

GutGPT interaction log. All participant interactions with GutGPT during simulations were analyzed for content, style, and engagement patterns to capture participant-system interaction dynamics. The codebook utilized for the interactions can be found in the Supplementary Table 3. Two researchers independently coded the logs, ensuring reliability and consistency through iterative comparison and consensus resolution. To improve transparency and illustrate the variety of clinician queries and GutGPT responses, we include representative interaction excerpts in Supplementary Table 10.

Sample size and statistical analysis

A sample size of 104 participants (52 per group) was selected to achieve 80% power to detect a clinically relevant medium effect size (Cohen's $d = 0.5$) with a 5% significance for the outcome BI, the primary measure of acceptance in UTAUT. This effect size was based on prior RCTs analyzing technology acceptance[57,58].

Primary analyses focused on between-arm comparisons of pre- to post-intervention changes. Given the non-normal distribution, two-tailed Mann–Whitney *U* tests were used. For UTAUT constructs, we defined a clinically meaningful threshold as ≥0.5 SD change in median score, using the greater of the two arms' standard deviations. To complement this, we also report the proportion of participants in each arm achieving ≥1-point Likert increases on each UTAUT domain. Confidence intervals for medians and proportions were estimated using bootstrap resampling (1000 iterations). No corrections were made for multiple comparisons.

Decision accuracy was analyzed at the session level, based on the collective decision made by the group for each case. In contrast, usability and adoption outcomes (UTAUT constructs) were collected and analyzed at the individual level via pre- and post-simulation surveys. This dual-level approach allowed us to evaluate both collaborative decision-making performance and individual adoption intent.

Exploratory subgroup analyses were conducted to examine whether changes in UTAUT outcomes or interaction patterns varied by training level (medical student vs. resident), clinical experience, or sex. We were unable to meaningfully compare participants by AI familiarity due to limited representation of those with high baseline familiarity. These subgroup analyses were not pre-specified and were interpreted descriptively to identify potential trends.

No missing survey data were observed, as all questions were mandatory and survey completion was verified in real time by simulation proctors.

Qualitative analysis

For qualitative analysis, thematic analysis of interviews utilized a deductive-inductive approach, guided by pre-determined themes derived from UTAUT and emergent themes identified during coding. As previously reported[59], the process involved creating interview summaries, holistic reviews, and categorization of significant quotes. Since this analysis draws from the same dataset and interview corpus, thematic saturation was achieved, with no new themes emerging after the initial set of interviews.

For the GutGPT interactions, inter-coder reliability was assessed using Cohen's Kappa calculated with Python's *sklearn.metrics* package. Descriptive statistics were computed as mean proportions.

While overall session durations were capped, time-on-task for individual scenarios or GutGPT queries was not consistently captured and was therefore not analyzed.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to sensitive educational history data of participants. However, the minimal dataset necessary to interpret, replicate, and build upon the findings are available from the corresponding authors on request. The underlying code for this study is not publicly available per the IRB as individual level patient data is required for the code, but are available to qualified researchers on request to the corresponding author. All analyses were conducted using Python (v3.11) and the scipy 1.14.1, scikit-learn 1.6.1, pandas 2.2.3, and numpy 1.26.4 packages.

Code availability

The underlying code for this study [and training/validation datasets] is not publicly available but are available to qualified researchers on request to the corresponding author. All analyses were conducted using Python (v3.11) and the scipy 1.14.1, scikit-learn 1.6.1, pandas 2.2.3, and numpy 1.26.4 packages.

Received: 18 February 2025; Accepted: 18 July 2025;
Published online: 18 August 2025

References

1. Arasu, V. A. et al. Comparison of mammography AI algorithms with a clinical risk model for 5-year breast cancer risk prediction: an observational study. *Radiology* **307**, e222733 (2023).
2. Liu, W. et al. Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: a systematic review and meta-analysis. *Eur. Heart J. Qual. Care Clin. Outcomes* **9**, 310–322 (2023).
3. Lu, J. et al. Performance of multilabel machine learning models and risk stratification schemas for predicting stroke and bleeding risk in patients with non-valvular atrial fibrillation. *Comput. Biol. Med* **150**, 106126 (2022).
4. Joshi, M., Mecklai, K., Rozenblum, R. & Samal, L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. *JAMIA Open* **5**, ooac022 (2022).

5. Marwaha, J. S., Landman, A. B., Brat, G. A., Dunn, T. & Gordon, W. J. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. *NPJ Digit. Med.* **5**, 13 (2022).
6. Cutillo, C. M. et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digital. Med.* **3**, 47 (2020).
7. Ghorayeb, A., Darbyshire, J. L., Wronikowska, M. W. & Watkinson, P. J. Design and validation of a new Healthcare Systems Usability Scale (HSUS) for clinical decision support systems: a mixed-methods approach. *BMJ Open* **13**, e065323 (2023).
8. Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F. & Simsekler, M. C. E. The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. *BMC Med. Inform. Decis. Mak.* **25**, 110 (2025).
9. Perivolaris, A. et al. Quality of interaction between clinicians and artificial intelligence systems. A systematic review. *Future Health. J.* **11**, 100172 (2024).
10. Lawton, T. et al. Clinicians risk becoming 'liability sinks' for artificial intelligence. *Future Health. J.* **11**, 100007 (2024).
11. Alanazi, A. Clinicians' views on using artificial intelligence in healthcare: opportunities, challenges, and beyond. *Cureus* **15**, e45255 (2023).
12. Hogg, H. D. J. et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. *J. Med. Internet Res.* **25**, e39742 (2023).
13. Park, J., Zhong, X., Dong, Y., Barwise, A. & Pickering, B. W. Investigating the cognitive capacity constraints of an ICU care team using a systems engineering approach. *BMC Anesthesiol.* **22**, 10 (2022).
14. Sbaffi, L., Walton, J., Blenkinsopp, J. & Walton, G. Information overload in emergency medicine physicians: a multisite case study exploring the causes, impact, and solutions in four North England National Health Service Trusts. *J. Med. Internet Res.* **22**, e19126 (2020).
15. Roshani, M. A. et al. Generatlve large language model-powered conversational AI app for personalized risk assessment: case study in COVID-19. *JMIR AI* **4**, e67363 (2025).
16. Reddy, S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implement Sci.* **19**, 27 (2024).
17. Shung, D. L. et al. Validation of an electronic health record-based machine learning model compared with clinical risk scores for gastrointestinal bleeding. *Gastroenterology* **167**, 1198–1212 (2024).
18. Peery, A. F. et al. Burden and cost of gastrointestinal, liver, and pancreatic diseases in the United States: update 2021. *Gastroenterology* **162**, 621–644 (2022).
19. Stanley, A. J. et al. Outpatient management of patients with low-risk upper-gastrointestinal haemorrhage: multicentre validation and prospective evaluation. *Lancet* **373**, 42–47 (2009).
20. Laine, L., Barkun, A. N., Saltzman, J. R., Martel, M. & Leontiadis, G. I. ACG clinical guideline: upper gastrointestinal and ulcer bleeding. *Am. J. Gastroenterol.* **116**, 899–917 (2021).
21. Dingel, J. et al. Predictors of health care practitioners' intention to use AI-enabled clinical decision support systems: meta-analysis based on the unified theory of acceptance and use of technology. *J. Med. Internet Res.* **26**, e57224 (2024).
22. Kim, S., Lee, K. H., Hwang, H. & Yoo, S. Analysis of the factors influencing healthcare professionals' adoption of mobile electronic medical record (EMR) using the unified theory of acceptance and use of technology (UTAUT) in a tertiary hospital. *BMC Med. Inf. Decis. Mak.* **16**, 12 (2016).
23. Panicker, R. O. & George, A. E. Adoption of automated clinical decision support system: a recent literature review and a case study. *Arch. Med. Health Sci.* **11**, 86–95 (2023).
24. Venkatesh, V., Morris, M. G., Davis, G. B. & Davis, F. D. User acceptance of information technology: toward a unified view. *MIS Q.* **27**, 425–478 (2003).
25. Ilgen, J. S., Sherbino, J. & Cook, D. A. Technology-enhanced simulation in emergency medicine: a systematic review and meta-analysis. *Acad. Emerg. Med.* **20**, 117–127 (2013).
26. Jones, C., Thornton, J. & Wyatt, J. C. Enhancing trust in clinical decision support systems: a framework for developers. *BMJ Health Care Inform.* **28**. https://doi.org/10.1136/bmjhci-2020-100247 (2021).
27. Giuffrè, M. et al. Expert of experts verification and alignment (EVAL) framework for large language models safety in gastroenterology. *NPJ Digital. Med.* **8**, 242 (2025).
28. Abell, B. et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. *Implement. Sci.* **18**, 32 (2023).
29. Shamszare, H. & Choudhury, A. Clinicians' perceptions of artificial intelligence: focus on workload, risk, trust, clinical decision making, and clinical integration. *Healthcare (Basel)* **11**, https://doi.org/10.3390/healthcare11162308 (2023).
30. Nair, M., Svedberg, P., Larsson, I. & Nygren, J. M. A comprehensive overview of barriers and strategies for AI implementation in healthcare: mixed-method design. *PLoS ONE* **19**, e0305949 (2024).
31. Parzen-Johnson, S. et al. Use of the electronic health record to optimize antimicrobial prescribing. *Clin. Ther.* **43**, 1681–1688 (2021).
32. Düvel, J. A. et al. An AI-based clinical decision support system for antibiotic therapy in sepsis (KINBIOTICS): use case analysis. *JMIR Hum. Factors* **12**, e66699 (2025).
33. Buléon, C. et al. Simulation-based summative assessment in healthcare: an overview of key principles for practice. *Adv. Simul. (Lond.)* **7**, 42 (2022).
34. Elendu, C. et al. The impact of simulation-based training in medical education: a review. *Medicine* **103**, e38813 (2024).
35. Liaw, S. Y. et al. Artificial intelligence versus human-controlled doctor in virtual reality simulation for sepsis team training: randomized controlled study. *J. Med. Internet Res.* **25**, e47748 (2023).
36. Feigerlova, E., Hani, H. & Hothersall-Davies, E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Med. Educ.* **25**, 129 (2025).
37. Wutz, M., Hermes, M., Winter, V. & Köberlein-Neu, J. Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: integrative review. *J. Med. Internet Res.* **25**, e46548 (2023).
38. Demsash, A. W., Kalayou, M. H. & Walle, A. D. Health professionals' acceptance of mobile-based clinical guideline application in a resource-limited setting: using a modified UTAUT model. *BMC Med. Educ.* **24**, 689 (2024).
39. Lu, Y., Barkun, A. N. & Martel, M. Adherence to guidelines: a national audit of the management of acute upper gastrointestinal bleeding. The REASON registry. *Can. J. Gastroenterol. Hepatol.* **28**, 495–501 (2014).
40. Huebner, J., Chung, S. L., Kizilcec, R. F., Laine, L. & Shung, D. Tu1975 provider trust and perceived usefulness of machine learning risk stratification tool for acute upper gastrointestinal bleeding using the technology acceptance model: a pilot study. *Gastroenterology* **164**, S-1168–S-1169 (2023).
41. Molnar, C. *Interpretable Machine Learning* (Lulu.com, 2020).
42. Lin, Y. & Jeon, Y. Random forests and adaptive nearest neighbors. *J. Am. Stat. Assoc.* **101**, 578–590 (2006).
43. Charles, J. S. Consistent nonparametric regression. *Ann. Stat.* **5**, 595–620 (1977).
44. Kuenzel, S. R. *Heterogeneous Treatment Effect Estimation Using Machine Learning*, (University of California, Berkeley, 2019).

45. Goldstein, A., Adam, K., Justin, B. & Pitkin, E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **24**, 44–65 (2015).
46. Deshmukh, F. & Merchant, S. S. Explainable machine learning model for predicting GI bleed mortality in the intensive care unit. *Am. J. Gastroenterol.* **115**, 1657–1668 (2020).
47. Shung, D. L. et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. *Gastroenterology* **158**, 160–167 (2020).
48. Chang, S. C. et al. The comparison and interpretation of machine-learning models in post-stroke functional outcome prediction. *Diagnostics (Basel)* **11**, https://doi.org/10.3390/diagnostics11101784 (2021).
49. Elshawi, R., Al-Mallah, M. H. & Sakr, S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med. Inf. Decis. Mak.* **19**, 146 (2019).
50. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
51. Caruana, R. et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1721–1730 (Association for Computing Machinery, 2015).
52. Kaplan, D. E. et al. AASLD Practice Guidance on risk stratification and management of portal hypertension and varices in cirrhosis. *Hepatology* **79**, 1180–1211 (2024).
53. Abraham, N. S. et al. American college of gastroenterology-canadian association of gastroenterology clinical practice guideline: management of anticoagulants and antiplatelets during acute gastrointestinal bleeding and the periendoscopic period. *Am. J. Gastroenterol.* **117**, 542–558 (2022).
54. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).
55. Venkatesh, V., Thong, J. Y. L. & Xu, X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS Q.* **36**, 157–178 (2012).
56. Chan, C. et al. Assessing the usability of gutgpt: a simulation study of an AI clinical decision support system for gastrointestinal bleeding risk. Preprint at https://arxiv.org/abs/2312.10072 (2023).
57. Baumeister, H. et al. Impact of an acceptance facilitating intervention on diabetes patients' acceptance of Internet-based interventions for depression: a randomized controlled trial. *Diabetes Res. Clin. Pr.* **105**, 30–39 (2014).
58. Baumeister, H. et al. Impact of an acceptance facilitating intervention on patients' acceptance of Internet-based pain interventions: a randomized controlled trial. *Clin. J. Pain.* **31**, 528–535 (2015).
59. Rajashekar, N. C. et al. Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* 1–20 (Association for Computing Machinery, 2024).

Acknowledgements

This study was funded by NIDDK T32DK007017 and NIDDK K23DK125718. The funder played no role in study design, data collection, analysis, and interpretation of data, or the writing of this manuscript.

Author contributions

S.C. conceptualized the study, developed the methodology, performed data analyses, drafted the manuscript, and prepared all Figures and Tables except Fig. 1. M.G. conceptualized the study, developed part of the methodology, drafted the manuscript, and prepared Fig. 1. N.R., Y.P., Y.E.S., A.H., J.S., A.W., L.E., T.M. developed part of the methodology. C.C., K.Y., T.S., S.N.S., S.K. developed part of the GutGPT methodology. R.K. developed part of the evaluation methodology. L.L. and J.S. conceptualized the study, provided critical manuscript revisions and developed part of the methodology. D.S. conceptualized the study, developed the methodology, provided critical manuscript revisions, coordinated the effort, and provided supervision. Note that J.S., L.L., D.S. are co-senior authors. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01896-5.

Correspondence and requests for materials should be addressed to Sunny Chung or Dennis L. Shung.