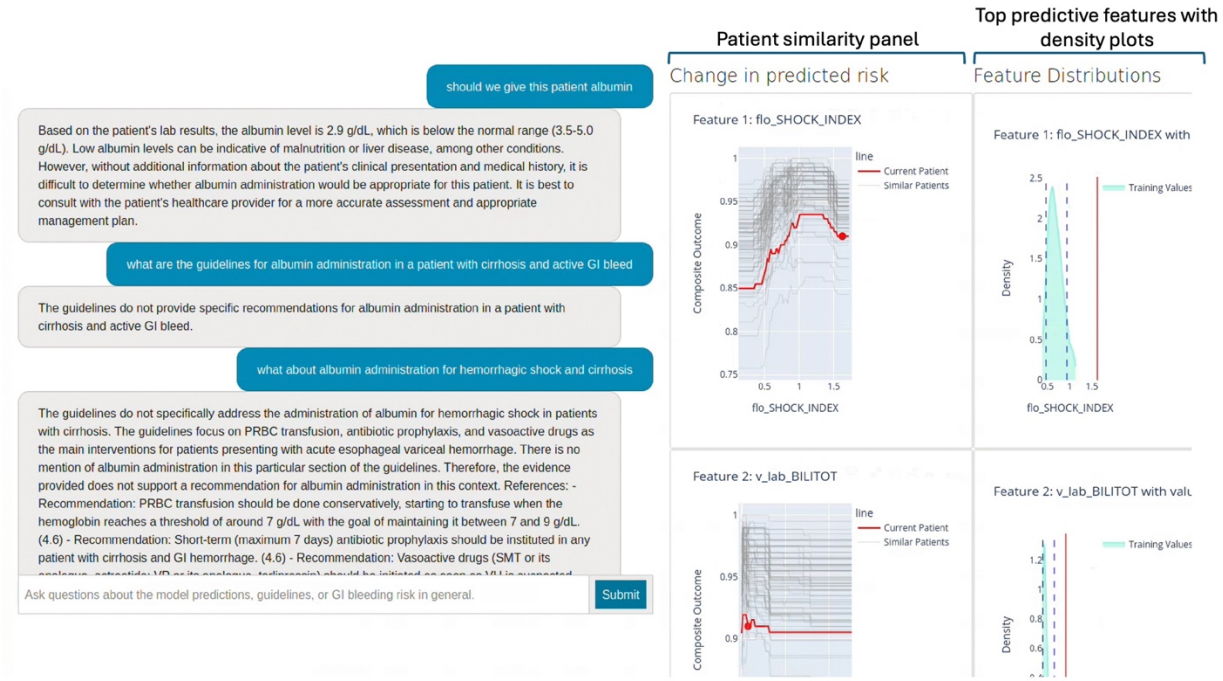
Table of Contents

Supplementary Figure 1. Dashboard Interface. The display includes a predicted risk score, adjustable patient inputs, and outputs based on 100 most similar patients. The central panel shows changes in predicted risk based on top features, while the right panel provides distribution plots comparing the current patient's values to the training dataset. This interface also enables users to test hypothetical changes in input variables and visualize corresponding shifts in model predictions.



Supplementary Figure 2. Screenshot of the GutGPT Interface Used in Simulation-Based Trial.
This figure shows the deployed user interface of GutGPT as used during the study. Clinicians interacted with the system using free-text queries (right-aligned blue bubbles), and the system returned responses (left-aligned white bubbles) integrating guideline-based recommendations and contextual information. On the right side of the interface, model predictions are displayed using a patient similarity panel and visualizations of key predictive features. This layout reflects the real interface tested during simulation and was designed to balance usability, transparency, and interpretability in a clinical context. Adjustable model inputs were accessible in both arms via the interface.

Survey Question	Mapping
Using AI-CDSS will improve my performance in clinical care	UTAUT Performance Expectancy
Using AI-CDSS will increase my productivity in clinical care	UTAUT Performance Expectancy
Using AI-CDSS will enhance my effectiveness in clinical care	UTAUT Performance Expectancy
AI-CDSS are useful in clinical care	UTAUT Performance Expectancy
I find AI-CDSS to be clear and understandable	UTAUT Effort Expectancy
Using AI-CDSS does not require a lot of effort	UTAUT Effort Expectancy
AI-CDSS are easy to use	UTAUT Effort Expectancy
Using AI-CDSS does not require a lot of time	UTAUT Effort Expectancy
I would use AI-CDSS if my co-residents use it	UTAUT Social Influence
I would use AI-CDSS if my attendings use it	UTAUT Social Influence
I would use AI-CDSS if other healthcare professionals (e.g., nurses) use it	UTAUT Social Influence
The healthcare system facilitates the use of AI-CDSS	UTAUT Facilitating Conditions
The EHR makes it easy to use AI-CDSS	UTAUT Facilitating Conditions
There is sufficient integration of AI-CDSS with other clinical tools and systems	UTAUT Facilitating Conditions
There is sufficient training in the use of AI-CDSS	UTAUT Facilitating Conditions
I intend to use AI-CDSS in clinical practice	UTAUT Behavioral Intention
I intend to use AI-CDSS frequently when clinically appropriate	UTAUT Behavioral Intention
I intend to use AI-CDSS in the future	UTAUT Behavioral Intention
I believe AI-CDSS is reliable	Trust
I believe AI-CDSS is trustworthy	Trust
I believe AI-CDSS will work properly	Trust

I feel good about using AI-CDSS in clinical care	Trust
I trust recommendations from AI-CDSS for clinical care	Trust
Develop personalized risk assessments in real time	Perceived Benefit
Serve as an assistant in clinical practice	Perceived Benefit
Identify risk factors to help clinicians decide on treatments	Perceived Benefit
Inform triage of patients based on severity	Perceived Benefit
Help diagnose diseases based on clinical presentation	Perceived Benefit
Anticipate decompensation prior to clinical shift	Perceived Benefit
It misses important social, emotional, and physical aspects of clinical care	Perceived Risk
It is inferior compared to clinicians' real life experience and intuition	Perceived Risk
It fails to be transparent about the decision-making process	Perceived Risk
It reduces clinicians' autonomy and control	Perceived Risk
It raises privacy risks for patients and clinicians	Perceived Risk
Its developers lack accountability when clinical errors occur	Perceived Risk
It increases the risk of clinical errors	Perceived Risk
It promotes discrimination in medical care	Perceived Risk

Supplementary Table 1. Survey Instrument Mapping to UTAUT and Additional Constructs. This table lists all survey questions used in the study, along with their corresponding theoretical constructs. UTAUT constructs include Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions, Behavioral Intention. Additional domains included Trust, Perceived Benefit and Perceived Risk. All items were rated on a 5-point Likert scale (1 = Strongly disagree to 5 = Strongly agree).

Question	**Notes** (Take rough notes, if not possible, that's OK.)
In general, what was your experience like using Gut-GPT?	-
Do you think you would use Gut-GPT to aid your decision making process in a real clinical situation? Please explain why or why not.	-
What did you think about the chat interface itself? What about the dashboard? Was Gut-GPT easy or difficult to use? Please explain.	-
Was there any other feedback you would like to share?	-
Thank you for your feedback. The thoughts that you've shared will be very helpful for our research in developing Gut-GPT. We will stop the recording now.	-

Supplementary Table 2. Semi-Structured Interview Guide.
This table presents the standard set of qualitative interview questions used to elicit participant feedback on GutGPT, including usability, trust, and perceived clinical value. Interviews were conducted immediately following the simulation scenarios. Notes were taken by study staff during each session and recordings were reviewed as needed for clarification.

Themes		Definition	Codes
Sequence		The sequence in which a question is asked within each scenario	Running number
Timestamp			
Guideline-related questions	Medicine (PPI)	Questions about proton pump inhibitors	1 = Presence, 0 = Absence
	Medicine (Vasoactive compound)		1 = Presence, 0 = Absence
	Medicine (aspirin)		1 = Presence, 0 = Absence
	Medicine (antibiotics)		1 = Presence, 0 = Absence
	Medicine (general)	General questions about medicine other than the above)	1 = Presence, 0 = Absence
	Transfusion	Whether need to do blood transfusion	1 = Presence, 0 = Absence
	Scope	Whether need to do scope	1 = Presence, 0 = Absence
	Scope timing	Timing of the scope	1 = Presence, 0 = Absence
	Discharge	Whether to discharge	1 = Presence, 0 = Absence
	Admit	Whether to admit	1 = Presence, 0 = Absence
	Risk (risk of bleeding / needing a hospital intervention)		1 = Presence, 0 = Absence
Other type of questions	Care management questions not posed by guidelines	Care management questions not posed by guidelines. Essentially any clinical questions that don't meet the above categories	1 = Presence, 0 = Absence
	Non-medical related questions		1 = Presence, 0 = Absence
Question Format	Generic vs specific	Generic refers to questions that are not specific to the patient	0 = Generic
			1 = Specific
	Open- (1) vs Closed- (2) ended questions		0 = Closed
			1 = Open
	Rephrased Question	Identify whether a question is rephrased question from earlier	0 = New Question
			1 = Rephrased Question

	Follow Up Questions	Whether a question is a followup from a prior answer that GutGPT gave	0 = New topic 1 = Follow Up
	Complexity	Simple questions are question in a simple sentence. Moderate complexity questions are questions that have conditional clauses . High complexity questions have conditional clauses that are compound	0 = simple 1 = moderate complexity 2 = high complexity
	Purpose: data retrieval (1) vs confirmatory (2)	Defined as questions that begin with "What", "How", or "Which" (such as What are the indications for proton pump inhibitors? What is the dose of ceftriaxone? When should I scope this patient). Confirmatory questions typically include phrases like "Should I", "Is it appropriate", "Would it be correct to" and reference a specific decision or action (such as "Should I discharge this patient").	1 = data retrieval, 2 = Confirmatory, 0 = Uncertain

Supplementary Table 3. Interaction Codebook for GutGPT Query Analysis.
This table presents the structured coding framework used to analyze clinician interactions with the GutGPT interface during simulation sessions. Each query was annotated across multiple dimensions, including clinical content (e.g., guideline topics), structural attributes (e.g., specificity, complexity), and functional intent (e.g., information-seeking vs. decision confirmation).

The coding schema includes:

- Binary indicators (e.g., 1 = presence, 0 = absence) for clinical concepts such as medications or interventions;
- Categorical variables for query characteristics (e.g., generic vs. specific, open vs. closed format);
- Ordinal scales to assess question complexity (e.g., 0 = simple, 1 = moderate, 2 = high); and
- Functional classifications for query purpose (e.g., 1 = data retrieval, 2 = confirmatory, 0 = uncertain).

This rule-based schema was developed a priori based on clinical relevance and refined during pilot sessions. Two independent reviewers applied the codes to each query using detailed definitions and example-based heuristics. Discrepancies were resolved through consensus discussion to ensure reliability. This coding enabled structured analysis of how clinicians interacted with GutGPT and how query styles aligned with clinical reasoning tasks.

Supplementary Table 4. CONSORT-AI Checklist for Reporting Trials Involving Artificial Intelligence Interventions.

Section	Item	CONSORT 2010 Item[a]	CONSORT-AI Item		Addressed on Page No[b]
Title and Abstract					
Title and Abstract	1a	Identification as a randomised trial in the title	CONSORT-AI 1a,b Elaboration	(i) Indicate that the intervention involves artificial intelligence/machine learning in the title and/or abstract and specify the type of model.	1, 2
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)		(ii) State the intended use of the AI intervention within the trial in the title and/or abstract.	1,2
Introduction					
Background and objectives	2a	Scientific background and explanation of rationale	CONSORT-AI 2a (i) Extension	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g. healthcare professionals, patients, public).	2
	2b	Specific objectives or hypotheses			2
Methods					
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio			16
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons			18
Participants	4a	Eligibility criteria for participants	CONSORT-AI 4a (i) Elaboration	State the inclusion and exclusion criteria at the level of participants.	16
			CONSORT-AI 4a (ii) Extension	State the inclusion and exclusion criteria at the level of the input data.	16
	4b	Settings and locations where the data were collected	CONSORT-AI 4b Extension	Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements.	11-15
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	CONSORT-AI 5 (i) Extension	State which version of the AI algorithm was used.	11-15
			CONSORT-AI 5 (ii) Extension	Describe how the input data were acquired and selected for the AI intervention.	10-15

			CONSORT-AI 5 (iii) Extension	Describe how poor quality or unavailable input data were assessed and handled.	10-15
			CONSORT-AI 5 (iv) Extension.	Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users.	10-15
			CONSORT-AI 5 (v) Extension	Specify the output of the AI intervention	10-15
			CONSORT-AI 5 (vi) Extension	Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice.	10-15
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed			18
	6b	Any changes to trial outcomes after the trial commenced, with reasons			N/A No changes to trial outcomes
Sample size	7a	How sample size was determined			19
	7b	When applicable, explanation of any interim analyses and stopping guidelines			N/A
Randomisation					
Sequence generation	8a	Method used to generate the random allocation sequence			16
	8b	Type of randomisation; details of any restriction (such as blocking and block size)			16
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned			16
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions			16
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how			16
	11b	If relevant, description of the similarity of interventions			10-16
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes			19
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses			19
Results					

Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome			3, Figure 3
	13b	For each group, losses and exclusions after randomisation, together with reasons			3, Figure 3
Recruitment	14a	Dates defining the periods of recruitment and follow-up			16
	14b	Why the trial ended or was stopped			N/A
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group			Table 1
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups			3, Table 1
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)			Table 2, 3
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended			N/A
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory			3-4, Supplementary Tables
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	CONSORT-AI 19 Extension	Describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, explain why not.	5-6, Supplementary Table 11
Discussion					
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses			7-8
Generalisability	21	Generalisability (external validity, applicability) of the trial findings			7-8
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence			6-7
Other Information					
Registration	23	Registration number and name of trial registry			16
Protocol	24	Where the full trial protocol can be accessed, if available			16
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	CONSORT-AI 25 Extension.	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.	20

Supplementary Note 1:

Low Risk Scenario: Hematemesis

Patient: 85F
Chief Complaint: Coffee-ground emesis following multiple episodes of diarrhea and vomiting
Vitals: BP 110/70 mmHg, HR 85 bpm, afebrile
Key Labs: Hgb 12.0, INR 1.0, Platelets 220
Clinical Features: Mild epigastric tenderness, hemodynamically stable, normal rectal exam. CT abdomen/pelvis unremarkable.
Clinical Decision: Determined to be low-risk; observed briefly, then discharged home.
Outcome: Outpatient EGD confirmed a healed Mallory-Weiss tear with no need for intervention.

Moderate Risk Scenario: Melena

Patient: 78M
Chief Complaint: Black stools over several days; history of GERD and ischemic heart disease on aspirin
Vitals: BP 110/75 mmHg, HR 95 bpm, afebrile
Key Labs: Hgb 8.1, INR 1.0, Platelets 250
Clinical Features: Epigastric tenderness, no active bleeding signs, stable on observation
Clinical Decision: Admitted for observation due to anemia and GI bleeding history
Outcome: EGD showed severe esophagitis (Class C), no intervention performed. Discharged with PPI.

High Risk Scenario: Hematemesis

Patient: 62M
Chief Complaint: Hematemesis and melena in the setting of cirrhosis
Vitals: BP 105/50 mmHg, HR 120 bpm, afebrile
Key Labs: Hgb 8.2, INR 1.5, Platelets 90
Clinical Features: Signs of portal hypertension, ascites, jaundice, and variceal bleeding suspected
Clinical Decision: Admitted to ICU for urgent intervention
Outcome: EGD revealed high-risk esophageal varices; band ligation performed. Patient stabilized and was stepped down from ICU after 48 hours

		Male		Female	
		GutGPT+ Dashboard n = 43	Dashboard n = 48	GutGPT+ Dashboard n = 9	Dashboard n = 6
BI	Median (95% CI)	0.2 (-0.2, 0.3)	0.0 (-0.3, 0.0)	0.0 (-1.0, 0.3)	0.3 (0.0, 0.7)
	% ≥1 in Likert Scale	25.0% (10.7%, 44.9%)	17.2% (5.8%, 35.8%)	33.3% (15.6%, 55.3%)	28.0% (12.1%, 49.4%)
EE	Median (95% CI)	0.6 (0.1, 1.1)	0.0 (-0.2, 0.0)	0.6 (0.2, 1.5)	0.5 (0.2, 1.0)
	% ≥1 in Likert Scale	42.9% (24.5%, 62.8%)	10.3% (2.2%, 27.4%)	37.5% (18.8%, 59.4%)	24.0% (10.7%, 36.9%)
PE	Median (95% CI)	0.0 (-0.5, 0.0)	0.2 (0.0, 0.5)	0.0 (-0.5, 0.2)	0.2 (0.0, 0.5)
	% ≥1 in Likert Scale	14.3% (4.0%, 32.7%)	6.9% (0.8%, 22.8%)	16.7% (4.7%, 37.4%)	20.0% (6.7%, 40.7%)
FC	Median (95% CI)	0.2 (0.1, 0.6)	0.0 (-0.2, 0.0)	0.0 (-0.4, 0.2)	0.0 (-0.2, 0.2)
	% ≥1 in Likert Scale	17.9% (6.1%, 36.9%)	13.8% (3.9%, 31.7%)	8.3% (1.0%, 27.0%)	4% (0.1%, 20.4%)
SI	Median (95% CI)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (-0.3, 0.0)	0.0 (-0.3, 0.0)
	% ≥1 in Likert Scale	3.6% (0.1%, 18.3%)	6.9% (0.8%, 22.8%)	20.8% (7.1%, 42.2%)	12.0% (2.5%, 31.2%)

Supplementary Table 5: Subgroup Analysis of UTAUT Constructs by Sex and Intervention Arm.
Median changes and the proportion of participants showing a ≥1-point increase on the Likert scale are presented for each UTAUT construct—Behavioral Intention (BI), Effort Expectancy (EE), Performance Expectancy (PE), Facilitating Conditions (FC), and Social Influence (SI)—stratified by sex (male vs. female) and intervention arm (GutGPT+Dashboard vs. Dashboard alone). 95% confidence intervals are reported for all estimates. This table highlights potential heterogeneity in perceived usability and adoption intent by gender subgroup.

		Medical Student		Resident	
		GutGPT+ Dashboard n = 12	Dashboard n = 13	GutGPT+ Dashboard n = 40	Dashboard n = 41
BI	Median (95% CI)	0.3 (-0.2, 0.8)	0.0 (-0.3, 0.3)	0.0 (-0.5, 0.0)	0.3 (0.3, 0.7)
	% ≥1 in Likert Scale	25.0% (5.5%, 57.2%)	15.4% (1.9%, 45.4%)	30.0% (16.6%, 46.5%)	24.4% (12.4%, 40.3%)
EE	Median (95% CI)	1.1 (0.5, 1.6)	0.0 (-0.5, 0.2)	0.5 (0.2, 1.0)	0.2 (0.0, 0.5)
	% ≥1 in Likert Scale	66.7% (34.9%, 90.1%)	15.4% (1.9%, 45.4%)	32.5% (18.6%, 49.1%)	17.1% (7.2%, 32.1%)
PE	Median (95% CI)	0.0 (-0.6, 0.2)	0.0 (0.0, 0.2)	0.0 (-0.2, 0.0)	0.2 (0.0, 0.5)
	% ≥1 in Likert Scale	8.3% (0.2%, 38.5%)	7.7% (0.2%, 36.0%)	17.5% (7.3%, 32.8%)	14.6% (5.6%, 29.2%)
FC	Median (95% CI)	0.2 (0.2, 0.6)	0.0 (-0.2, 0.2)	0.0 (-0.5, 0.2)	0.0 (-0.2, 0.0)
	% ≥1 in Likert Scale	8.3% (0.2%, 38.5%)	7.7% (0.2%, 36.0%)	15.0% (5.7%, 29.8%)	9.8% (2.7%, 23.1%)
SI	Median (95% CI)	0.0 (-0.3, 0.0)	0.0 (0.0, 0.3)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
	% ≥1 in Likert Scale	8.3% (0.2%, 38.5%)	0.0% (0.0%, 27.4%)	12.5% (4.2%, 26.8%)	12.2% (4.1%, 26.2%)

Supplementary Table 6: Subgroup Analysis of UTAUT Constructs by Training Level and Intervention Arm. This table presents median changes and the proportion of participants with ≥1-point increases in Behavioral Intention (BI), Effort Expectancy (EE), Performance Expectancy (PE), Facilitating Conditions (FC), and Social Influence (SI) for medical students and residents, stratified by study arm (GutGPT+Dashboard vs. Dashboard). Confidence intervals are provided for all estimates. These results explore differential perceptions of usability and adoption intent based on clinical training level.

		PGY1		PGY2+	
		GutGPT+ Dashboard n = 10	Dashboard n = 13	GutGPT+ Dashboard n = 30	Dashboard n = 28
BI	Median (95% CI)	0.0 (-1.0, 0.3)	0.7 (0.3, 1.3)	0.0 (-0.5, 0.0)	0.0 (-0.3, 0.0)
	% ≥1 in Likert Scale	30.0% (6.7%, 65.2%)	38.5% (13.9%, 68.4%)	30.0% (14.7%, 49.4%)	17.9% (6.1%, 36.9%)
EE	Median (95% CI)	0.5 (0.2, 1.5)	0.5 (0.5, 1.8)	0.4 (-0.2, 0.8)	0.1 (-0.2, 0.2)
	% ≥1 in Likert Scale	10.5% (0.3%, 44.5%)	15.4% (1.9%, 45.4%)	40.4% (22.7%, 59.4%)	17.9% (6.1%, 36.9%)
PE	Median (95% CI)	0.0 (-0.5, 0.2)	0.5 (0.2, 0.8)	0.0 (-0.4, 0.0)	0.2 (0.0, 0.5)
	% ≥1 in Likert Scale	10.0% (0.3%, 44.5%)	23.1% (5.0%, 53.8%)	20.0% (7.7%, 38.6%)	10.7% (2.3%, 28.2%)
FC	Median (95% CI)	0.0 (-0.5,0.4)	0.0 (0.0, 0.8)	0.1 (-0.2, 0.5)	0.2 (0.0, 0.5)
	% ≥1 in Likert Scale	0.0% (0.0%, 30.8%)	7.7% (0.2%, 36.0%)	20.0% (7.7%, 38.6%)	10.7% (2.3%, 28.2%)
SI	Median (95% CI)	0.0 (0.0, 0.0)	0.0 (-0.3, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
	% ≥1 in Likert Scale	0.0% (0.0%, 30.8%)	15.4% (1.9%, 45.4%)	16.7% (5.6%, 34.7%)	10.7% (2.3%, 28.2%)

Supplementary Table 7: Subgroup Analysis of UTAUT Constructs by Postgraduate Year (PGY) Level.
This table presents changes in UTAUT domains (Behavioral Intention [BI], Effort Expectancy [EE], Performance Expectancy [PE], Facilitating Conditions [FC], Social Influence [SI]) among PGY1 and PGY2+ residents. Data are stratified by intervention arm (GutGPT+Dashboard vs. Dashboard) and include median change with 95% confidence intervals, as well as the proportion of participants with a ≥1-point increase on the Likert scale. These results help assess how training stage influences perceived usability and adoption intent of GenAI tools.

		White		Non-White	
		GutGPT+ Dashboard n = 34	Dashboard n = 26	GutGPT+ Dashboard n = 18	Dashboard n = 28
BI	Median (95% CI)	0.0 (-0.7, 0.0)	0.0 (-0.7, 0.0)	0.0 (-1.0, 0.0)	0.3 (0.3, 0.7)
	% ≥1 in Likert Scale	27.3% (13.3%, 45.5%)	26.1% (10.2%, 48.4%)	29.4% (10.3%, 56.0%)	18.5% (6.3%, 38.1%)
EE	Median (95% CI)	0.5 (0.2, 1.0)	0.2 (0.0, 0.5)	1.0 (0.8, 2.0)	0.2 (0.0, 0.5)
	% ≥1 in Likert Scale	30.3% (15.6%, 48.7%)	17.4% (5.0%, 38.8%)	52.9% (27.8%, 77.0%)	14.8% (4.2%, 33.7%)
PE	Median (95% CI)	0.0 (0.0, 0.2)	0.2 (0.0, 0.5)	0.0 (-1.0, 0.0)	0.2 (0.0, 0.5)
	% ≥1 in Likert Scale	9.1% (1.9%, 24.3%)	13.0% (2.8%, 33.6%)	29.4% (10.3%, 56.0%)	14.8% (4.2%, 33.7%)
FC	Median (95% CI)	0.0 (-0.2, 0.2)	0.0 (-0.2, 0.2)	0.2 (-0.5, 0.5)	0.0 (-0.2, 0.0)
	% ≥1 in Likert Scale	3.0% (0.1%, 15.8%)	13.0% (2.8%, 33.6%)	35.3% (14.2%, 61.75%)	3.7% (0.1%, 19.0%)
SI	Median (95% CI)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (-0.3, 0.0)
	% ≥1 in Likert Scale	9.1% (1.9%, 24.3%)	8.7% (1.1%, 28.0%)	17.6% (3.8%, 43.4%)	11.1% (2.4%, 29.2%)

Supplementary Table 8: Subgroup Analysis of UTAUT Constructs by Race.

This table presents changes in UTAUT domains (Behavioral Intention [BI], Effort Expectancy [EE], Performance Expectancy [PE], Facilitating Conditions [FC], Social Influence [SI]) comparing White and non-White participants. Data are stratified by intervention arm (GutGPT+Dashboard vs. Dashboard) and include median change with 95% confidence intervals, as well as the proportion of participants with a ≥1-point increase on the Likert scale. These results help assess how training stage influences perceived usability and adoption intent of GenAI tools.

		GPT-3.5 n = 21	GPT-4 n = 31
BI	Median (95% CI)	0.3 (0.0, 0.7)	0.0 (-1.0, 0.3)
	% ≥1 in Likert Scale	25.8% (11.9%, 44.6%)	33.3% (14.6%, 57.0%)
EE	Median (95% CI)	0.7 (0.2, 1.2)	0.5 (0.0, 1.0)
	% ≥1 in Likert Scale	33.3% (14.6%, 57.0%)	45.2% (27.3%, 64.0%)
PE	Median (95% CI)	0.0 (-0.5, 1.0)	0.0 (-0.2, 0.0)
	% ≥1 in Likert Scale	12.9% (3.6%, 29.8%	19.0% (5.4%, 41.9%)
FC	Median (95% CI)	0.2 (0.2, 0.8)	0.0 (-0.5, 0.2)
	% ≥1 in Likert Scale	12.9% (3.6%, 29.8%)	14.3% (3.0%, 36.3%)
SI	Median (95% CI)	0.0 (-0.3, 0.0)	0.0 (0.0, 0.0)
	% ≥1 in Likert Scale	12.9% (3.6%, 29.8%)	9.5% (1.2%, 30.4%)

Supplementary Table 9. Post Hoc Subgroup Analysis of UTAUT Outcomes by Large Language Model (LLM) Version: GPT-3.5 vs. GPT-4.

This table summarizes participant responses to the UTAUT domains for users of different underlying GutGPT models. Median changes (with 95% CI) and proportions with ≥1-point Likert increases are presented for GPT-3.5 (n = 21) and GPT-4 (n = 31). These exploratory findings help evaluate whether perceived usability and adoption intention were influenced by the specific LLM version used during simulation sessions.

Query	GutGPT Response
Does this pt need to be admitted?	Based on the guidelines, the patient does not need to be admitted. The guidelines recommend that patients presenting to the emergency department with upper gastrointestinal bleeding (UGIB) who are classified as very low risk (defined as a risk assessment score with ≤1% false negative rate for the outcome of hospital-based intervention or death) can be discharged with outpatient follow-up rather than admitted to the hospital. This recommendation is based on the goal of allowing a subset of patients to be safely discharged from the emergency department, reducing costs without significant risk of poor outcomes that require or could have been prevented with in-hospital management. The Glasgow-Blatchford score (GBS) is a commonly used risk assessment tool, and patients with a GBS of 0-1 have a sensitivity point estimate of 99%, with a lower bound of 95% confidence interval of 97%-98%. Specificities are higher with point estimates ranging from 27% to 40%. Two large multicenter studies reported GBS = 0-1 in 19%-24% of patients presenting with UGIB. Therefore, a GBS of 0-1 meets the requirement for very low risk and allows more patients to be discharged compared to a GBS of 0, which was the previous threshold suggested in the 2012 ACG Guidelines
What is this pt's risk of in-hospital intervention?	The patient's predicted risk for a hospital-based intervention is 5.6%. Of the 100 most similar patients in the training data set, 2 required a hospital based intervention. The American College of Gastroenterology (ACG) guidelines recommend a sensitivity threshold of 99%. Since the predicted risk is below 10.5% (the 99% sensitivity threshold), the patient is considered to be low risk for a hospital-based intervention. Note this algorithm is unable to access any findings you may find on history and physical exam; furthermore, it can only use the diagnoses that are indicated via ICD codes already in the patient's chart. This algorithm also only takes the first recorded vital signs in the chart and does not incorporate any vital signs taken after the first values are measured and recorded.
What should we do with this pt?	Based on the recommendations provided in the guidelines, I would recommend the following: (a) Patients with acute bleeding from gastric varices (GV) should be initially managed in a similar fashion to those bleeding from esophageal varices (EV). This includes using a restrictive transfusion policy, vasoactive drug infusion, and antibiotic prophylaxis. (b) In patients bleeding from gastroesophageal varices type 1 (GOV1), either endoscopic variceal ligation (EVL) or cyanoacrylate glue injection (if available) are the recommended endoscopic treatments. (c) Transjugular intrahepatic portosystemic shunt (TIPS) is the treatment of choice in the control of bleeding from cardiofundal varices, such as GOV2 or isolated gastric varices type 1 (IGV1). (d) Cyanoacrylate glue injection is an alternative treatment option for cases in which TIPS is not technically feasible, but it is not approved for the treatment of gastric varices (GV) in the United States and should only be performed in centers with expertise in this procedure. The evidence supporting these recommendations is as follows: - A meta-analysis of three randomized controlled trials (RCTs) comparing cyanoacrylate injection versus EVL showed that both therapies are equally effective for initial hemostasis, but cyanoacrylate injection is associated with significantly lower rebleeding rates. However, the overall quality of evidence is low due to small sample sizes, and the meta-analysis was dominated by the larger study that included only GOV1 varices. - Endoscopic variceal ligation (EVL) should only be performed on small gastric varices (GV) in which both the mucosal and contralateral wall of the vessel can be suctioned into the ligator. Otherwise, the band may fall off in several days, leading to ulceration and potential rebleeding. - Other emerging endoscopic tools, such as endoscopic ultrasound-guided insertion of coils and cyanoacrylate, may provide greater safety and efficacy than traditional treatments. - TIPS has been shown to be very effective in the treatment of bleeding GV, with a success rate of over 90% for initial hemostasis. TIPS frequently requires additional embolization of spontaneous collaterals feeding the varices. To summarize, based on the available evidence, initial management of bleeding from gastric varices should follow a similar approach to bleeding from esophageal varices. Endoscopic

	treatments such as EVL or cyanoacrylate injection are recommended for bleeding from gastroesophageal varices type 1, while TIPS is the treatment of choice for bleeding from cardiofundal varices. Cyanoacrylate glue injection can be considered as an alternative when TIPS is not feasible, but it should only be performed in centers with expertise in this procedure.
What are the factors in the risk score that led to 11% risk?	The most important features in predicting the risks of gastrointestinal bleeding are the shock index, hemoglobin level, and INR level. The shock index is a measure of how fast the heart is beating in relation to the blood pressure. A value of 0.8 means that the patient's heart is beating faster than average. This could indicate that the patient is experiencing some form of distress or shock. The hemoglobin (HGB) level is a measure of how much oxygen-carrying protein is in the blood. A value of 12 is higher than the average value, indicating that the patient has a higher amount of oxygen-carrying protein in their blood. This could suggest that the patient has good oxygenation and is less likely to experience bleeding. The INR (International Normalized Ratio) level is a measure of how well the blood is clotting. A value of 0.9 is lower than the average value, indicating that the patient's blood is clotting more efficiently. This suggests that the patient is at a lower risk of bleeding. Overall, when these important features are taken into consideration, the model predicts that the patient has a lower risk

Supplementary Table 10. Representative excerpts from GutGPT interactions across simulated UGIB scenarios. This table presents clinician queries and corresponding GutGPT responses selected to illustrate the range of interaction types observed in the study. These examples are drawn from the full interaction dataset (n = 171) and demonstrate how GutGPT generated responses.

Scenario	Query	Expected Answer (ML Dashboard)	Unique Responses	Accuracy
Low Risk	What is the patient's risk?	0.106	1	20/20
	What are the top three most important factors contributing to risk?	Hgb, WBC, Respiratory Rate	20	20/20
Medium Risk	What is the patient's risk?	0.7805	1	20/20
	What are the top three most important factors contributing to risk?	Hgb, Hct, Cl	20	20/20
High Risk	What is the patient's risk?	0.8601	1	20/20
	What are the top three most important factors contributing to risk?	Hgb, Hct, TBili	20	20/20

Supplementary Table 11. Reproducibility of GutGPT Responses to Model Output Queries Across Risk Scenarios

To assess the consistency of GutGPT's interpretations of structured model outputs, two queries were submitted 20 times for each of the three standardized UGIB scenarios. These queries were: (1) “What is the patient’s risk?” and (2) “What are the top three most important factors contributing to risk?” The table reports the number of unique response phrasings and whether each matched the expected output from the ML dashboard. In all cases, the correct risk score and feature-value pairs were returned (100% accuracy).

Supplementary Note 2

Classify whether the prompt provided falls in the following 6 categories (A, B, C, D, E, or F).
Select all that apply. The user will ask a question and your answer must be formatted as "A: X, B: X, C: X, D: X, E: X, F: X" where X is the score between 0 to 100 on how confident you are in your classification that the category applies.
Do not get rid of categories with a score of 0 in your response.
A) Risk prediction for a hospital-based intervention for a potential GI bleeding patient.
B) Variable importance of predicted risk of requiring a hospital-based intervention for a patient.
C) Guidelines for patient management options according to current guidelines.
D) Symptoms, causes, and other general questions of GI bleeding.
E) General questions of this tool and goals.
F) Unsure or not a question related to GI bleeding.
Examples:
Here are 62 examples that you have to consider for classifying the final prompt:
Question: What is the risk for this patient? Answer: A: 100, B: 0, C: 0, D: 0, E: 0, F: 0
Question: Why does the model think the patient is low risk? Answer: A: 0, B: 100, C: 0, D: 0, E: 0, F: 0
Question: What is the evidence that this tool is useful? Answer: A: 0, B: 0, C: 0, D: 0, E: 100, F: 0
Question: Why should I trust this model? Answer: A: 0, B: 0, C: 0, D: 0, E: 100, F: 0
Question: What should I order to assess this patient? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: What are the most common causes of lower GI bleeding? Answer: A: 0, B: 50, C: 0, D: 100, E: 0, F: 0
Question: What do you not know? Answer: A: 0, B: 0, C: 0, D: 0, E: 100, F: 0
Question: What is the risk that this patient needs an endoscopy? Answer: A: 100, B: 50, C: 0, D: 0, E: 0, F: 0
Question: What are high risk features? Answer: A: 0, B: 100, C: 0, D: 80, E: 0, F: 0
Question: What lab values are high risk? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: What medications are high risk? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: What vital signs are high risk? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: Should this patient be admitted, and if so where? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: What are the most common etiologies behind GI bleeding? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: What are the recommendation for endoscopy timing in this patient? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: How urgently does this patient need an endoscopy? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: Does this patient need to stop enoxaparin therapy? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: Does this patient need to stop anticoagulants therapy? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0Q
Question: Does this patient need to stop warfarin therapy? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: Does this patient need to stop dabigatran therapy? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: Does this patient need to stop aspirin therapy? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: What are the current recommendations on anticoagulant therapy for a patient with suspected severe bleeding? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: What would be the best transfusion strategy for our patient? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: What are the features (under the category of vital signs) that are important in the risk prediction model for this patient? Answer: A: 0, B: 100, C: 50, D: 0, E: 0, F: 0
Question: What are the features (under the category of laboratory values) that are important in the risk prediction model for this patient? Answer: A: 0, B: 100, C: 50, D: 0, E: 0, F: 0
Question: What are the features (under the category of current medication) that are important in the risk prediction model for this patient? Answer: A: 0, B: 100, C: 50, D: 0, E: 0, F: 0
Question: What is the likelihood that hemostatic intervention would be needed for this patient? Answer: A: 100, B:0, C: 0, D: 0, E: 0, F: 0
Question: What are common symptoms of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: Is melena common in upper gastrointestinal bleeding? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: Is hematochezia common in upper gastrointestinal bleeding? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: Is hematemesis common in gastrointestinal bleeding? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: What is melena? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: What is hematochezia? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: What is hematemesis? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: What are common causes of upper gastrointestinal bleeding? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0

Question: Is peptic ulcer a common cause of upper gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Are esophagogastric varices a common cause of upper gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Are esophageal tumors a common cause of upper gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Is Mallory-Weiss tear a common cause of upper gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Do NSAIDS increase the risk of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Does ibuprofen increase the risk of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Does aspirin increase the risk of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Do anticoagulants increase the risk of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Does warfarin increase the risk of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Do direct oral anticoagulants increase the risk of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Does dabigratan increase the risk of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Do SSRI increase the risk of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 50, D: 100, E: 0, F: 0
Question: Do aldosterone antagonists increase the risk of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 100, D: 100, E: 0, F: 0
Question: Do calcium channel blockers increase the risk of gastrointestinal bleeding? Answer: A: 0, B: 0, C: 100, D: 100, E: 0, F: 0
Question: What is myocardial infarction? Answer: A: 0, B: 0, C: 0, D: 0, E: 0, F: 100
Question: Does past medical history of myocardial infarction affect transfusion strategies in gastrointestinal bleeding? Answer: A: 0, B: 0, C: 100, D: 0, E: 0, F: 0
Question: What is rheumatoid arthritis? Answer: A: 0, B: 0, C: 0, D: 0, E: 0, F: 100
Question: Does past medical history of rheumatoid arthritis affect transfusion strategies in gastrointestinal bleeding? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: What is meningitis? Answer: A: 0, B: 0, C: 0, D: 0, E: 0, F: 100
Question: What is purpura? Answer: A: 0, B: 0, C: 0, D: 0, E: 0, F: 100
Question: Is aspirin used for primary prophylaxis in ischemic heart disease? Answer: A: 0, B: 0, C: 0, D: 0, E: 0, F: 100
Question: Is coumadin used for atrial fibrillation treatment? Answer: A: 0, B: 0, C: 0, D: 0, E: 0, F: 100
Question: What is Coumadin? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: What is Aspirin? Answer: A: 0, B: 0, C: 0, D: 100, E: 0, F: 0
Question: What can you do? Answer: A: 0, B: 0, C: 0, D: 0, E: 100, F: 0
Question: Does liver health impact the risk for a hospital-based intervention? Answer: A: 0, B: 0, C: 80, D: 80, E: 0, F: 0

Supplementary Note 3: Computational Requirements for GutGPT

Hardware Requirements

- CPU: ≥ 4-core, 2.5 GHz (recommended 8-core for concurrent multi-user simulations)
- RAM: ≥ 32 GB (to support multi-threaded serving)
- GPU: Not necessary, optional
- Disk: ≥ 200 GB SSD (to support model storage, vector databases, and logging) (!) GPU and additional memory may be required if local fine-tuning of LLMs is later implemented
- OS: Ubuntu, Microsoft Windows 10/11, or macOS
- Python: ≥ 3.9

Software & Libraries

- LLM Interface: openai>0.27.6
- Web Frameworks: Dash 2.9.3, Gradio 3.28.3 for UI
- Data Processing: pandas 1.4.4, numpy 1.24.2, scikit-learn 1.2.2
- Visualization: plotly 5.3.1, matplotlib 3.4.3
- Vector Storage: chromadb 0.3.22 (local embedding storage)
- Machine Learning Components: treelite 3.2.0

Model and Data Inputs

- LLM Model: via OpenAI API
- Clinical Prediction Algorithms: XGBoost models compiled via Treelite for low-latency execution
- Input format: JSON-like queries through API interface (fields: clinical variables, risk factors, guideline lookup context)

User Interface and Interoperability

- Browser Support: Chrome, Firefox
- Display Resolution: ≥ 1280×720 pixels
- Authentication: Local token-based session management (no OAuth used in simulation environment)
- API Integration: RESTful endpoints (internal use for component communications

Scalability and Limitations

- The current implementation was optimized for single-user or small-group concurrent simulations (≤ 10 users).
- Horizontal scaling is feasible by deploying multiple Dash/Gradio app instances behind a load balancer.
- Latency bottlenecks mainly stem from LLM API response times (~500–1200 ms per query); pre-caching common queries can improve responsiveness.
- In production settings, GPU acceleration or serverless scaling via OpenAI's managed endpoints is recommended to support clinical load (>100 sessions/day).

Supplementary Movie 1. *Live demonstration of the GutGPT interface.* The video shows a clinician interacting with GutGPT through a natural language chat interface while simultaneously viewing structured model outputs. GutGPT responds to free-text queries by retrieving relevant clinical guidelines and integrating patient-specific model predictions. The right-side panel displays the predicted risk of hospital-based intervention and the top contributing features, supporting transparent and context-sensitive decision-making. This hybrid interface was used during simulation sessions to assess usability, interpretability, and perceived clinical utility of the GenAI tool.