Original Paper

Evaluating Human and AI-Based Triage Performance in the Emergency Department: Real-World Cohort Study

Nada Abedin[1,2], MD; Maximilian Freiherr von Beust, MSc; Nina Kutscher[3], MD; Simon Meier[2,4], MD; André Schiller[2]; Ludwig Hofbauer[1], MD; Hans-Peter Erasmus[1], MD; Sebastian M Haberkorn[3], MD; Alica Kubesch-Grün[1,2], MD; Jürgen Graf[2], MD; Ingo Marzi[4], MD; Stefan Zeuzem[1], MD; Georg Dultz[1,2], MD

[1]Goethe University Frankfurt, University Hospital, Medical Clinic 1, Frankfurt am Main, Hesse, Germany
[2]Department of Acute and Emergency Medicine, Goethe University Frankfurt, University Hospital, Frankfurt am Main, Hesse, Germany
[3]Department of Cardiology, Goethe University Frankfurt, University Hospital, Frankfurt am Main, Hesse, Germany
[4]Department of Trauma Surgery and Orthopedics, Goethe University Frankfurt, University Hospital, Frankfurt am Main, Hesse, Germany

Corresponding Author:

Nada Abedin, MD
Goethe University Frankfurt
University Hospital, Medical Clinic 1
Theodor-Stern-Kai 7
Frankfurt am Main, Hesse 60596
Germany
Email: abedin@med.uni-frankfurt.de

Abstract

Background: Emergency departments (EDs) depend on rapid, reliable triage for patient safety and flow. The Manchester Triage System (MTS) is widely implemented but exhibits known interrater variability. Large language models (LLMs) have emerged as potential decision-support tools for triage and early assessment, yet real-world comparative evaluations remain scarce.

Objective: This study aimed to compare clinician and LLM performance on ED triage accuracy, specialty assignment, diagnostic recommendations, and admission prediction in a real-world cohort.

Methods: In a retrospective cohort study, 305 walk-in ED encounters at a German tertiary-care hospital were analyzed. A nurse performed real-time MTS-assessment and initial specialty assignment at presentation. Each anonymized case was independently evaluated by 4 physicians and 12 LLMs from 4 providers (OpenAI, Google, Anthropic, and Mistral). Triage agreement was assessed against 2 reference standards: a consensus human label (ground-truth 1 [GT1]) and real-time nurse-assigned MTS-triage (ground-truth 2 [GT2]). For human raters, a leave-one-out variant of GT1 (LOO-GT1) assessed incorporation bias. The primary statistic was quadratic-weighted Cohen $\varkappa$ (unweighted $\varkappa$ reported alongside). Secondary outcomes were directional triage errors (over- or undertriage, $|\Delta| \leq 1$), specialty top-1 accuracy, per-case precision, recall, F_1-score of recommended diagnostics, and admission prediction (sensitivity, specificity, positive predictive value [PPV], and negative predictive value [NPV]).

Results: Median age was 45 (IQR 31-62) years; 188 (61.6%) patients were male; admission rate was 20.1% (56/278). Against GT1, human-rater (physicians and nurse) accuracy ranged 0.59-0.83 (quadratic-weighted $\varkappa$ 0.60-0.88; unweighted $\varkappa$ 0.35-0.74), and LLM accuracy 0.35-0.70 (quadratic-weighted $\varkappa$ 0.46-0.72; unweighted $\varkappa$ 0.11-0.53). Under LOO-GT1, human-rater weighted $\varkappa$ fell to 0.48-0.81. Only 1 clinician (Doctor 1, weighted $\varkappa$ 0.81) retained agreement clearly above any LLM, while 3 LLMs (Gemini-3.5-flash=0.72, GPT-5.5=0.69, and GPT-5=0.66) exceeded the LOO-corrected agreement of the remaining 4 human raters. One-off agreement exceeded 86% for all raters and models. Vs GT1, overtriage ranged 6%-56% and undertriage 6%-36%. Vs GT2, corresponding ranges were 11%-62% and 5%-30%. Against the study specialty reference, physician specialty-assignment accuracy ranged 0.88-0.93 and LLM accuracy 0.62-0.80. Against the operational triage-pathway specialty assignment, the corresponding ranges were 0.81-0.82 and 0.57-0.75. Model diagnostic recommendation F_1-scores ranged 0.55-0.69. Admission prediction demonstrated high sensitivity (0.73-0.95) and NPV (0.90-0.96), but modest PPV (0.25-0.41).

Conclusions: In a real-world ED cohort, 1 clinician (Doctor 1, a specialist physician) had the highest urgency agreement point estimate under both GT1 and LOO-GT1. Under LOO-GT1, 3 contemporary LLMs reached agreement that exceeded

that of the other 4 clinicians. For specialty assignment, all clinicians remained clearly ahead of the best LLM. LLMs showed strong sensitivity and high NPV for admission prediction, but modest PPV precludes autonomous use. Because both reference standards are empirical rather than outcome-anchored gold standards, findings represent agreement against expert consensus and routine practice rather than an absolute correctness benchmark, and support a complementary, human-in-the-loop role for LLM-based decision-support in ED triage and early assessment.

J Med Internet Res 2026;28:e91442; doi: 10.2196/91442

Keywords: triage; manchester triage system; emergency department; emergency medicine; large language models; artificial intelligence; clinical decision support; benchmarking; inter-rater reliability; retrospective cohort study

Introduction

Emergency departments (EDs) are under extended pressure from work overload caused by increasing patient volumes, crowding, and staff shortages [1,2]. To maintain patient safety and flow, quick and precise triage and early assessment are key. They determine time to first clinician encounter, diagnostics, and allocation of scarce resources such as monitoring. Among others, the Manchester Triage System (MTS) has become one of the standards for initial urgency assessment, including German EDs [3,4].

The MTS differentiates 5 urgency levels, from immediate to nonurgent, based on reasons for presentation and discriminators. Multiple validation studies and large systematic reviews have evaluated its reliability and validity [4,5]. However, some studies have reported wide interrater variability, relevant rates of under- and overtriage, and limited sensitivity for highest urgency levels, with strikingly lower performance in very young and older adult patients [6-8]. This suggests that while MTS-based triage is structured and reasonably reliable, it still exhibits clinically relevant variability and safety gaps. Additionally, studies show that 60%-70% of walk-in patients have no indication for admission and are discharged from the ED, usually after long waiting times [9].

Meanwhile, AI is increasingly applied across all areas of medicine and emergency medicine. These include risk stratification, deterioration prediction, interpretation of images, operational forecasting, and scribing [10-15]. AI tools prove to be a reliable, quick, and in-time support tool maintaining patient safety. However, it remains hard to benchmark against human raters, and prospective validation remains scarce. While there have been studies in emergency medicine showing the ability of large language models (LLMs) to support in generating recommendations for admission status or LLM-based triage, there are few cohorts that include real patient data [16,17]. Many AI triage studies either use historical labels or single human raters as the reference and seldom report directional triage errors (over- vs undertriage), interrater variability, or multitask performance (triage, specialty routing, diagnostics, and admission) in one dataset [13-15,17-20].

We hypothesized that contemporary LLMs would approach but not equal experienced-clinician performance on MTS-based urgency triage, lag clinicians more substantially on specialty assignment, and provide clinically useful sensitivity and negative predictive value for admission prediction. To test these hypotheses, we systematically compared 4 physicians of varying seniority, a triage nurse, and 12 LLMs from 4 provider families against 2 reference standards in a real-world walk-in ED cohort of 305 encounters.

Methods

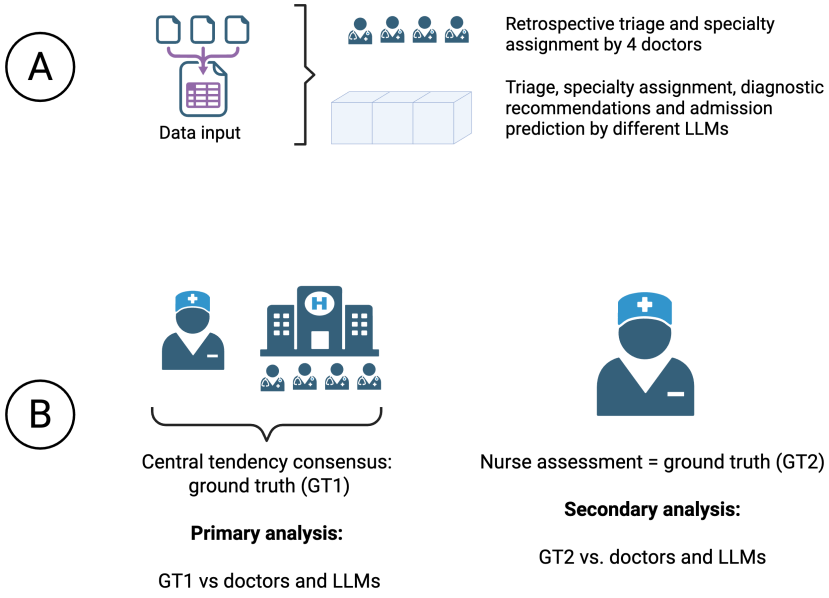
Data Collection

To compare real triage in the ED with LLM-based triage, a total of 305 visits to the ED at University Hospital Frankfurt, Germany, were analyzed in a retrospective analysis. Only self-presenting (walk-in) patients were included in the study. Ambulance arrivals and patients triaged directly into the resuscitation bay were excluded by design, as they enter the ED via a separate pathway and do not undergo nurse-led MTS triage in our institution. This sampling frame is by construction skewed toward lower-acuity presentations and therefore yields a small number of high-acuity (Red, Orange) cases. At the time of presentation, real-time triage was performed by an ED nurse according to the MTS. Initial assessment included assignment of triage urgency category and specialty. Additionally, the reason for presentation and vital signs (heart rate, blood pressure, temperature, and peripheral capillary oxygen saturation [SpO_2]) were documented. Clinical data including the final diagnosis, diagnostics performed, and disposition (discharge vs admission) were taken from patients' electronic files.

Retrospective Human and LLM-Based Assessment

After anonymization, each encounter was independently reassessed in a 2-step approach. First (Figure 1A), 4 physicians with different experience levels (head of department [Doctor 4], attending [Doctor 3], specialist [Doctor 1], and fellow [Doctor 2]), retrospectively assigned triage category and specialty based on the same structured information available at triage. There was exactly 1 physician at each of the described 4 seniority levels. Second (Figure 1B), the same anonymized case information was provided to 12 different LLMs through an API. The LLMs included models from OpenAI (gpt-4.1-mini, gpt-4.1, gpt-5-mini, gpt-5, and gpt-5.5), Google (gemini-2.5-flash, gemini-2.5-pro, gemini-3.5-flash), Anthropic (claude-sonnet-4, claude-sonnet-4.5, claude-sonnet-4.6), and Mistral (mistral-large).

Figure 1. Study design and reference framework. (A) Structured triage data, initially documented by an emergency department nurse, were retrospectively evaluated by 4 physicians and by 12 large language models, generating triage category, specialty assignment, diagnostic recommendations, and admission prediction. (B) Two reference standards were applied. For the primary analysis, ground truth (GT1) was defined as the central-tendency consensus of the 5 human raters (arithmetic mean of ordinal Manchester Triage System scores rounded to the more urgent category on 0.5 ties; see Methods). A leave-one-out variant (LOO-GT1) was used to correct for incorporation bias. For the secondary analysis, real-time nurse-performed Manchester Triage System assessment served as the operational reference standard (GT2). Physician and large language model performance were evaluated against both reference frameworks. Figure was created in BioRender [21]. LLM: large language model.



The retrospective raters (Doctors 1-4 and the 12 LLMs) received only the same structured information that had been available to the triage nurse at presentation (reason for presentation, age, and sex), duration of symptoms, comorbidity flags (diabetes and oral anticoagulation), vital signs (systolic blood pressure [SBP] and diastolic blood pressure [DBP], pulse, body temperature, SpO_2, and respiratory rate), pain score, Glasgow Coma Scale (GCS), and Quick Sequential Organ Failure Assessment (qSOFA), and were blinded to the nurse's MTS triage assignment, to the nurse's specialty assignment, to the assessments of the other retrospective raters, to all subsequent diagnostics performed during the ED encounter, and to the final hospital disposition. Missing values were indicated by a question mark and were not imputed. The retrospective raters therefore differed from the nurse only in that they evaluated the structured data without operational time pressure and without competing patient demands. They did not have access to additional clinical context, follow-up data, or outcome information.

Model runs were performed in sequence between September 10, 2025, and September 12, 2025, with a custom Python (Python Software Foundation) script. During revision, 3 models were added and run from May 19 to May 20, 2026. All models were called through their respective provider's native Python Software Development Kit using provider-supported structured-output mechanisms with the same Pydantic schema. In this manuscript, the term reasoning models refers to the OpenAI checkpoints gpt-5-mini, gpt-5, and gpt-5.5, which expose inference-time reasoning steps; nonreasoning models refers to all other evaluated models. We use "reasoning models" consistently throughout the paper and avoid the looser term "thinking models." Reasoning models do not support a user-settable temperature parameter and were called with the provider's default reasoning configuration. Wherever the temperature parameter was supported, it was set to 0.

All models used the same prompt, including the same output schema and descriptions. Models were accessed from the provider's servers without intermediaries (ie, Anthropic API and not via AWS Bedrock or Google Vertex). Where models supported custom token amounts, output tokens were limited to 5000 tokens and thinking tokens limited to 2500. The prompts were written fully in structured markdown. All model runs were single-pass; each patient case was queried exactly once per model, with no repeated runs.

Input data about the individual cases was shared with the LLMs in markdown format as a user message in key value pairs. Possible keys were study ID, weekday of visit, age, sex or gender, stated reason for visit, duration of symptoms (hours), diabetes, oral anticoagulation, SBP, DBP, pulse, body temperature, SpO2, respiratory rate, Numerical Pain Scale, GCS, and qSOFA. In case of missing values, a question mark was shared with the LLM. Every patient case was processed individually without any context of other patients (Table 1).

Table 1. The data output model was defined as follows (information shared with LLM[a] to define output structure).

Name	Type	Description	Enum values
triage_reasoning	string	Detailed explanation of the reasoning behind the triage decision, including key factors and considerations	_[b]
triage_color	enum	The Manchester Triage color category assigned to the patient	BLUE="blue - non-urgent" GREEN="green - standard" YELLOW="yellow - urgent" ORANGE="orange - very urgent" RED="red - immediate"
max_wait_minutes	Int	Maximum recommended waiting time in minutes before medical assessment based on triage category	_
admission_likelihood	enum	The likelihood that a patient will be admitted to a floor bed in the hospital for further treatment after the emergency room	DISCHARGE = "discharge" LIKELY_DISCHARGE = "likely discharge" POSSIBLE_ADMISSION = "possible admission" ADMISSION = "admission"
chief_complaint	string	The patient's primary symptom or main reason for seeking medical attention	_
discriminators	list (string)	List of specific signs, symptoms, or risk factors that influenced the triage decision according to Manchester Triage criteria	_
recommended_diagnostics	list (enum)	List of recommended diagnostic tests and examinations based on the presenting symptoms and triage assessment	ABSCESS_SWAB="abscess_swab" ASCITES_PARACENTESIS="ascites_paracentesis" BLOOD_GAS_ANALYSIS="blood_gas_analysis" BLOOD_TEST="blood_test" CHIKUNGUYA_SEROLOGY="chikunguya_serology" CLINICAL_EXAMINATION="clinical_examination" CONSULTATION="consultation" COVID_SWAB="covid_swab" CT_SCAN="ct_scan" D_DIMER_TEST="d_dimer_test" DENGUE_SEROLOGY="dengue_serology" DIAGNOSTIC_LAPAROSCOPY="diagnostic_laparoscopy" ECG="ecg" FLUSHING_OF_THE_CATHETER="flushing_of_the_catheter" GYNECOLOGICAL_EXAMINATION="gynecological_examination" INFLUENZA_SWAB="influenza_swab" MALARIA_TEST="malaria_test" PCN_EXCHANGE="pcn_exchange" PUNCTURE="puncture" RSV_SWAB="rsv_swab"

Name	Type	Description	Enum values
			SECOND_ROUND_OF_TROPONIN="second_round_of_troponin" TRANSTHORACIC_ECHOCARDIOGRAPHY="transthoracic_echocardiography" ULTRASOUND="ultrasound"
			URINE_DIPSTICK_TEST="urine_dipstick_test"
			URINE_CULTURE="urine_culture"
			X_RAY="x_ray"
specialty	enum	The medical specialty department to which the patient should be directed	
			CARDIOLOGY="cardiology"
			DENTAL_SURGERY="dental_surgery"
			ENDOCRINOLOGY="endocrinology"
			ENT="ent"
			GASTROENTEROLOGY="gastroenterology"
			GENERAL_SURGERY="general_surgery"
			GYNECOLOGY="gynecology"
			HEMATOLOGY_ONCOLOGY="hematology_oncology"
			INFECTIOLOGY="infectiology"
			NEPHROLOGY="nephrology"
			NEUROLOGY= "neurology"
			PULMONOLOGY="pulmonology"
			RHEUMATOLOGY="rheumatology"
			TRAUMA_SURGERY="trauma_surgery"
			UROLOGY="urology"
			VASCULAR_SURGERY="vascular_surgery"
confidence_rating	enum	Your confidence level in triage assessment and recommendations	
			LOW="low"
			MEDIUM="medium"
			HIGH="high"
Notes	string (nullable)	Additional relevant information or special considerations not captured in other fields	_

[a]LLM: large language model.
[b]Not applicable.

To provide good and stable data, structured output with predefined enum values was strongly preferred. The definition of an output schema with clearly predefined output options can and should be seen as a significant part of the prompting or context engineering technique. The fields "triage_reasoning," "discriminators," and "confidence_rating" were included in the output schema to enforce structured reasoning during generation (improving the stability and interpretability of the primary outputs) and to enable qualitative inspection of model behavior; they were not analyzed quantitatively in this study and are described here for transparency and reproducibility. Quantitative analyses in this paper focus exclusively on triage_color, specialty, recommended_diagnostics, and admission_likelihood. The system prompt is provided in Multimedia Appendix 1. The system prompt provides natural-language descriptions of the output fields, while the actual enum vocabularies and key names are defined and enforced by the Pydantic schema described above (eg, admission_likelihood={DISCHARGE, LIKELY_DISCHARGE, POSSIBLE_ADMISSION, ADMISSION}; specialty {16 enum values}; recommended_diagnostics={26 enum values}). Because all models were called via provider-native structured-output mechanisms with the schema as the authoritative contract, each model response was constrained to the schema regardless of the prompt's narrative phrasing. No retained schema-valid model output required post hoc remapping for parsing. A valid model output was defined as a complete response that passed the enforced structured-output schema for triage, specialty, and admission. An encounter for which no complete schema-valid response was retained was classified as a complete-output failure. A valid response containing an empty recommended-diagnostics list was retained and was not classified as a failure. For the September 2025 runs, provider calls were attempted up to 3 times with a fixed 30-second wait between attempts. For the May 2026 runs, API exceptions were attempted up to 3 times with a fixed 2-second wait, whereas parsing or schema-validation failures were recorded without an additional attempt. Task-specific analyses used complete-case intersections between valid model output and the available reference for that outcome; consequently, smaller denominators

caused by missing operational triage, diagnostic-reference, or disposition data were not counted as additional model-output failures.

Definition of Reference Standards ("Ground Truth") and Incorporation-Bias Correction

As there is no single universally accepted reference of ED triage, 2 complementary reference standards were applied (Figure 1). For the primary analysis, for ground truth 1 (GT1), the 5 ordinal MTS scores (Blue, Green, Yellow, Orange, and Red) were averaged, and the mean rounded to the nearest integer category. When 1 or more human ratings were missing (13 encounters lacked a recorded nurse triage, 3 encounters lacked Doctor 2 rating), the mean was computed over the available raters. In this subset, 0.5 ties were resolved a priori toward the more urgent category as a safety-conservative rule. This approach reflects collective expert judgment and reduces the influence of individual bias. Ground truth 2 (GT2) was defined as the real-time nurse-performed MTS triage recorded at presentation, representing the operational reference used in routine ED practice. This comparator is denoted as "Nurse" in the tables and figures because it represents the hospital's original nurse-triage pathway.

Because the 4 physicians and nurse contribute to GT1, their measured agreement with GT1 is partly circular with each rater's own score being part of the reference standard they're scored against. To correct for this, an additional leave-one-out variant of GT1 (LOO-GT1) was computed. For each evaluated human rater, GT1 was recomputed using the identical algorithm but excluding the respective rater's own score from the mean. Human raters are reported against both GT1 (primary, for direct comparability with the LLMs) and LOO-GT1 (sensitivity, for unbiased human-vs-reference agreement); LLM statistics are identical under both references.

For specialty assignment, the study specialty reference was defined independently of the urgency references GT1 and GT2. It was calculated as the modal specialty assignment among 5 human ratings—the initial specialty assignment recorded in the operational triage pathway and the independent retrospective assignments of 4 physicians. The physicians assigned specialty using the structured triage information and were blinded to the operational specialty assignment, the other raters' assignments, subsequent diagnostic findings, and disposition. The mode was calculated from all available ratings; 299/305 (98.03%) encounters had all 5 specialty ratings, whereas 6/305 (1.97%) had 4 because Doctor 2's rating was missing. In 10/305 (3.28%) encounters with more than 1 modal specialty, the historical deterministic spreadsheet rule ordered the tied normalized specialty labels in descending lexicographical order and selected the first label. No separate clinical adjudicator was used. Specialty performance was evaluated against this 5-rater modal consensus and, separately, against the operational specialty assignment.

Outcomes and Statistical Analysis

Primary outcomes were agreement with ground truth for triage (accuracy and Cohen κ) and specialty (top-1 accuracy). Secondary outcomes included interobserver agreement among clinicians and among models (pairwise κ), directional triage errors (over- and undertriage), and one-off agreement ($|\Delta| \leq 1$), overlap between model-recommended and real diagnostics (mean per-case precision, recall, and F_1-score), and admission decision performance (accuracy, sensitivity, specificity, positive predictive value [PPV], and negative predictive value [NPV]). Model admission likelihood was output on a 4-level ordinal scale (DISCHARGE, LIKELY_DISCHARGE, POSSIBLE_ADMISSION, and ADMISSION). For the primary binary admission analysis, model outputs were collapsed using threshold A (screening); possible admission or admission indicated predicted admission, whereas discharge or likely discharge indicated predicted discharge. This threshold favors sensitivity, consistent with the intended use of admission prediction as an early screening signal in a high-throughput emergency department setting. A prespecified threshold sensitivity analysis used threshold B (strict), under which only admission indicated predicted admission and all other categories indicated predicted discharge. Primary results under threshold A (screening) are reported in Table 2, and the comparison of threshold A (screening) with threshold B (strict) is reported in Table 3.

Table 2. Binary admission prediction performance under the primary threshold (≥ POSSIBLE_ADMISSION → admit). The "naive" baseline always predicts discharge and is shown for reference; its NPV[a] equals the discharge prevalence in the 278-encounter evaluable subsample (0.80). Admission prevalence=20.1% (56/278).

Model	N	Accuracy	Sensitivity	Specificity	PPV[b]	NPV
Naive (always discharge)	278	0.80	0.00	1.00	—[c]	0.80
gpt-5	278	0.73	0.80	0.71	0.41	0.93
gpt-5.5	278	0.68	0.86	0.64	0.38	0.95
gemini-3.5-flash	278	0.67	0.86	0.62	0.36	0.95
gpt-4.1	278	0.65	0.80	0.61	0.34	0.92
claude-sonnet-4.5	278	0.64	0.73	0.62	0.33	0.90
mistral-large	278	0.61	0.88	0.54	0.32	0.94
gpt-5-mini	278	0.59	0.82	0.53	0.30	0.92
gemini-2.5-pro	278	0.58	0.91	0.50	0.31	0.96

Model	N	Accuracy	Sensitivity	Specificity	PPV[b]	NPV
claude-sonnet-4.6	277	0.53	0.89	0.44	0.29	0.94
claude-sonnet-4	269	0.52	0.93	0.42	0.29	0.96
gemini-2.5-flash	278	0.46	0.95	0.33	0.26	0.96
gpt-4.1-mini	277	0.42	0.93	0.29	0.25	0.94

[a]NPV: negative predictive value.
[b]PPV: positive predictive value.
[c]Not applicable.

Table 3. Admission prediction under threshold A (screening) and threshold B (strict). Threshold A was the primary screening threshold: "possible admission" or "admission" indicated predicted admission. Threshold B was the strict threshold: only "admission" indicated predicted admission. The comparison of these thresholds constituted the prespecified threshold sensitivity analysis; columns headed "sensitivity" report sensitivity as a performance metric, distinct from the threshold sensitivity analysis.

Model	N	Threshold A: Accuracy	Sensitivity	Specificity	PPV[a]	NPV[b]	Threshold B: Accuracy	Sensitivity	Specificity	PPV	NPV
gpt-5	278	0.73	0.80	0.71	0.41	0.93	0.81	0.16	0.97	0.60	0.82
gpt-5.5	278	0.68	0.86	0.64	0.38	0.95	0.79	0.18	0.95	0.48	0.82
gemini-3.5-flash	278	0.67	0.86	0.62	0.36	0.95	0.83	0.34	0.95	0.63	0.85
gpt-4.1	278	0.65	0.80	0.61	0.34	0.92	0.81	0.14	0.98	0.67	0.82
claude-sonnet-4.5	278	0.64	0.73	0.62	0.33	0.90	0.79	0.34	0.90	0.46	0.84
mistral-large	278	0.61	0.88	0.54	0.32	0.94	0.80	0.16	0.96	0.50	0.82
gpt-5-mini	278	0.59	0.82	0.53	0.30	0.92	0.80	0.18	0.96	0.53	0.82
gemini-2.5-pro	278	0.58	0.91	0.50	0.31	0.96	0.74	0.70	0.75	0.41	0.91
claude-sonnet-4.6	277	0.53	0.89	0.44	0.29	0.94	0.78	0.57	0.84	0.47	0.89
claude-sonnet-4	269	0.52	0.93	0.42	0.29	0.96	0.80	0.17	0.95	0.47	0.82
gemini-2.5-flash	278	0.46	0.95	0.33	0.26	0.96	0.74	0.46	0.82	0.39	0.86
gpt-4.1-mini	277	0.42	0.93	0.29	0.25	0.94	0.79	0.09	0.97	0.45	0.81

[a]PPV: positive predictive value.
[b]NPV: negative predictive value.

Diagnostic Recommendation Analysis

Diagnostics performed during each ED encounter were extracted from the electronic medical record (laboratory orders, imaging requests, swabs, consultations, and procedures) and mapped during data extraction, prior to any model run, onto the same 26-item enum vocabulary that was provided to the LLMs. Mapping was performed by doctors involved in routine care (NA and GD) using simple equivalence rules (eg, "full blood count," "CRP," "liver enzymes"→BLOOD_TEST; "chest X-ray," "abdominal X-ray"→X_RAY; any consult→CONSULTATION). Items that did not have a corresponding enum value were excluded from both the reference set and any comparison; this affected fewer than 1% of items and is not expected to bias intermodel comparisons because the same exclusion rule was applied to every model's recommendation set. Each encounter was thereby reduced to a set of enum-coded performed diagnostics; the same enum vocabulary was the model output space. Per-encounter values were averaged across the 277 encounters with a nonempty real-diagnostics list (28 encounters with no documented diagnostics were excluded as undefined denominators).

Importantly, the set of diagnostics performed during the encounter is not a true gold standard; some tests may have been ordered defensively rather than necessarily, and some appropriate tests may have been omitted. The F_1-score reported here therefore measures agreement with routine practice rather than agreement with an ideal optimal work-up.

Statistical analyses were performed using Python. *Pandas*, *NumPy*, and scikit-learn were used for computations, figures were generated using matplotlib and biorender. Moreover, 95% CIs for accuracy and quadratic-weighted κ were obtained by case-resampled nonparametric bootstrap with B=1000 replicates (random seed 42). Pairwise differences between selected agents in quadratic-weighted κ vs GT1 were tested by paired bootstrap; in each replicate, the same resampled patient indices were used for both agents, the $\Delta\kappa$ between agents was recomputed, and a 2-sided bootstrap P value was obtained as $2\times \min(P(\Delta\kappa\leq 0), P(\Delta\kappa\geq 0))$ over the 1000 replicates. A Holm-Bonferroni step-down correction was applied to the family of 8 prespecified pairwise tests at family-wise α=.05. Both uncorrected and Holm-adjusted P values are reported in Table 4.

Table 4. Prespecified pairwise tests of differences in quadratic-weighted $\varkappa$ vs GT1. Paired bootstrap with B=1000 patient-resampled replicates (random seed 42); in each replicate both agents are evaluated on the same resampled patient indices. Two-sided *P* values are computed as $2\times\min(P(\Delta\varkappa\leq 0), P(\Delta\varkappa\geq 0))$ over the bootstrap distribution. Holm-Bonferroni step-down correction was applied to the family of 8 pre-specified tests at family-wise α=.05; both uncorrected and Holm-adjusted *P* values are shown.

Rater A	Rater B	N	$\varkappa$ A	$\varkappa$ B	$\Delta\varkappa$ (A–B)	95% CI	*P* value (2-sided)	*P* value (adjusted)
Doctor 1	gemini-3.5-flash	305	0.88	0.72	+0.16	0.10 to 0.23	<.001	<.001
Doctor 1	gpt-5.5	305	0.88	0.69	+0.19	0.13 to 0.26	<.001	<.001
Doctor 1	gpt-5	305	0.88	0.66	+0.22	0.16 to 0.28	<.001	<.001
gemini-3.5-flash	gpt-5.5	305	0.72	0.69	+0.03	–0.02 to 0.09	.20	.39
gemini-3.5-flash	gpt-5	305	0.72	0.66	+0.06	0.00 to 0.12	.04	.18
Doctor 2	gemini-3.5-flash	302	0.76	0.73	+0.03	–0.03 to 0.10	.34	.39
Doctor 2	gpt-5.5	302	0.76	0.70	+0.06	0.00 to 0.13	.07	.20
Nurse	gemini-3.5-flash	292	0.60	0.72	-0.13	–0.22 to –0.03	.008	.04

Ethical Considerations

Ethics Review

The study was approved by the institutional review board of Frankfurt University Hospital (Ethikkommission der Goethe-Universität Frankfurt, decision 2024-2001, date of approval September 13, 2024) and the local data security officer. The study was conducted in accordance with the Declaration of Helsinki and applicable German federal and state data-protection law (Datenschutz-Grundverordnung [DSGVO]).

Informed Consent

The requirement for written informed consent was formally waived by the ethics committee because the study was a retrospective analysis of fully deidentified routine clinical data and did not involve any change in patient care, additional intervention, or contact with the patients or their relatives.

Privacy and Confidentiality and Data Handling

Source data were extracted from the hospital information system by clinicians directly involved in routine care and were deidentified locally prior to any analysis. Each encounter was assigned a study ID with no link back to the source record outside the secured hospital network. API requests were issued from a study workstation to each provider's official API end point; no third-party intermediaries (eg, AWS Bedrock, Google Vertex AI, Azure OpenAI, or other reseller gateways) were used at any point. All 4 providers' standard API tiers (OpenAI, Google AI for Developers, Anthropic, and Mistral Le Plateforme) operate under zero-data-retention and no-training-on-API-data policies for the standard tier we used at the time of analysis; data submitted via these end points is not stored beyond the immediate request-response cycle and is not used to train, fine-tune, or evaluate any provider model.

Compensation

No financial or in-kind compensation was provided to patients, clinicians, or any other participants.

Results

Cohort Characteristics

A total of 305 walk-in encounters at the ED of University Hospital Frankfurt were analyzed; cohort characteristics are summarized in Table 5. Median age was 45 (IQR 31-62; range 1-90) years (15 encounters <18 years, 35 encounters aged 75 years or older). The cohort comprised 188 (61.6%) male and 117 (38.4%) female patients. Among vital signs, SBP was available in 255 cases and had a median of 145 (IQR 130-159) mm Hg. Diastolic blood pressure was available in 254 cases (median 80, IQR 72-90 mm Hg). Body temperature was recorded in 259 cases (median 36.6, IQR 36.4-36.8 °C), pulse in 256 cases (median 85, IQR 72-98 bpm), and oxygen saturation in 257 cases (median 98%, IQR 97%-99%). Respiratory rate was documented for 259 cases (median 14, IQR 13-16 breaths/min). The median qSOFA score was 0 (IQR 0-0; range 0-1).

Table 5. Cohort characteristics. Continuous variables are presented as median and IQR with the number of available observations. Admission proportion refers to the final hospital disposition among 278 evaluable encounters (56 admission, 222 discharge; 27 encounters with edge-case dispositions [transfer, left without being seen, undocumented, left as symptoms subsided, or discharged against medical advice] excluded).

Characteristic	Value (IQR/%)
Encounters, n	305
Sex (male)[a], n (%)	188 (61.6)
Age (y), median (IQR; range)	45 (31-62, 1-90)
Systolic BP[b] (mm Hg), n=255, median (IQR)	145 (130-159)
Diastolic BP (mm Hg), n=254, median (IQR)	80 (72-90)

Characteristic	Value (IQR/%)
Pulse (bpm), n=256, median (IQR)	85 (72-98)
Body temperature (°C), n=259, median (IQR)	36.6 (36.4-36.8)
SpO_2[c] (%), n=257, median (IQR)	98 (97-99)
Respiratory rate (per min), n=259, median (IQR)	14 (13-16)
Pain scale (0-10), n=138, median (IQR)	3 (2-5)
GCS[d], n=63, median (IQR)	15 (15-15)
qSOFA[e], n=232, median (IQR; range)	0 (0-0; 0-1)
Admission proportion, n/N (%)	56/278 (20.14)
Specialty (nurse-assigned), n (%)	
Trauma Surgery	81 (26.6)
Gastroenterology	63 (20.7)
Cardiology	54 (17.7)
Urology	31 (10.2)
Dental Surgery	20 (6.6)
Pulmonology	19 (6.2)
General Surgery	18 (5.9)
Hematology/Oncology	7 (2.3)
Neurology	5 (1.6)
Vascular Surgery	3 (1)
Nephrology	2 (0.7)
Rheumatology	1 (0.3)
Infectious Diseases	1 (0.3)

[a]Sex/gender as documented at presentation (binary: male/female); no other categories were recorded.
[b]BP: blood pressure.
[c]SpO_2: peripheral capillary oxygen saturation.
[d]GCS: Glasgow Coma Scale.
[e]qSOFA: Quick Sequential Organ Failure Assessment.

The overall proportion of hospital admissions was 20.1% (56/278); 27 encounters were transferred, left without being seen, or had no follow-up documentation. Specialties assigned by the nurse included trauma surgery (n=81), gastroenterology (n=63), cardiology (n=54), urology (n=31), dental surgery (n=20), pulmonology (n=19), general surgery (n=18), and others (n=19; Table 5).

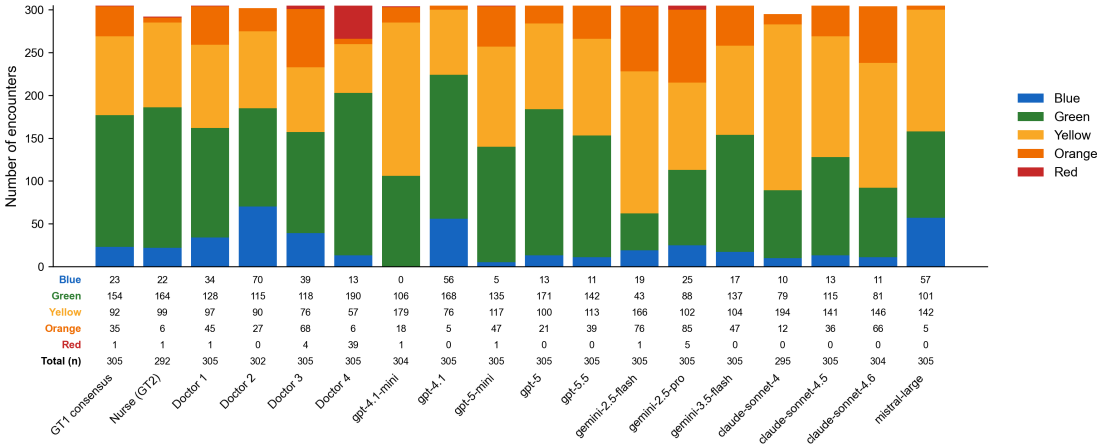
Model-Output Reliability

Across 3660 model-encounter calls (12 models×305 encounters), 3648/3660 (99.67%) produced complete schema-valid outputs, and 12/3660 (0.33%) resulted in complete-output failures. Claude-sonnet-4 produced 295/305 valid outputs (96.72%); 9 failures involved an out-of-enum diagnostic value, and 1 involved strings returned for fields requiring lists. Gpt-4.1-mini and Claude-sonnet-4.6 each produced 304/305 valid outputs (99.67%); the former failure was caused by invalid, unterminated JSON, and the latter by an out-of-enum recommended diagnostic. Each of the other 9 models produced 305/305 (100%) complete outputs. These failures account for the model-specific denominators below 305 in the following analyses. Further denominator reductions reflect the intersection with available operational-triage, diagnostic-reference, and eligible-disposition data, respectively, rather than additional model-output failures.

Triage Category Distributions

The distribution of triage categories (red, orange, yellow, green, and blue) assigned by each rater and model is illustrated in Figure 2. GT1 labels consisted primarily of green (n=154), yellow (n=92), orange (n=35), blue (n=23), and a single red case (n=1). GT2 assigned 164 green, 99 yellow, 6 orange, 22 blue, and 1 red. Physician distributions varied—Doctor 1 assigned 128 green, 97 yellow, 45 orange, 34 blue, and 1 red; Doctor 2 assigned 115 green, 90 yellow, 27 orange, 70 blue, and 0 red; Doctor 3 assigned 118 green, 76 yellow, 68 orange, 39 blue, and 4 red; Doctor 4's distribution was notably different from the other clinicians—39 red ratings (vs 0-4 for Doctor 1-3 and the nurse) and 190 green ratings (vs 115-164 for the other raters), reflecting a markedly different threshold for the red category in particular.

Figure 2. Distribution of triage categories by rater and model. Stacked bar charts showing the number of encounters assigned to each triage category (red, orange, yellow, green, blue) by the ground truth, nurse, physicians, and AI models, illustrating differences in acuity allocation.



Model distributions also varied, with counts detailed in Figure 2, reflecting differences in calibration across provider families and model versions.

Triage Agreement

Triage accuracy and Cohen $\varkappa$ (quadratically weighted as primary, unweighted reported alongside) compared with GT1 are summarized and visualized in Figure 3. The parallel LOO-GT1 statistics for the human raters are reported in Table 6 and visualized in Figure 4. Across 292-305 paired samples, physician accuracies ranged from 0.59 to 0.83 (quadratic-weighted $\varkappa$ 0.70-0.88; unweighted $\varkappa$ 0.35 to 0.74). Nurse triage demonstrated an accuracy of 0.61, quadratic-weighted $\varkappa$ of 0.60 (unweighted $\varkappa$ 0.36; n=292). Among the 12 models, gemini-3.5-flash achieved the highest agreement with GT1 (accuracy 0.70, weighted $\varkappa$ 0.72, unweighted $\varkappa$ 0.53), followed by gpt-5.5 (accuracy 0.66, weighted $\varkappa$ 0.69) and gpt-5 (accuracy 0.66, weighted $\varkappa$ 0.66). The lowest LLM agreement was observed for gemini-2.5-flash (accuracy 0.35, weighted $\varkappa$ 0.46). Exact values for all raters and models are presented in Figure 3.

Figure 3. Agreement with ground-truth 1 triage. Accuracy and quadratic-weighted Cohen $\varkappa$ for the 5 human raters (4 physicians + nurse) and the 12 large language models, plotted side by side and sorted by quadratic-weighted $\varkappa$ descending. Error bars represent 95% confidence intervals from nonparametric case-resampled bootstrap (B=1000, random seed 42).

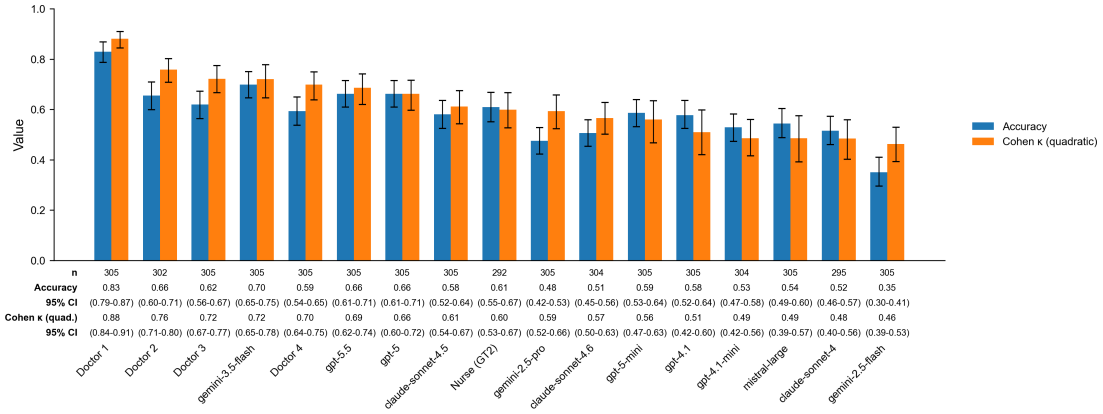
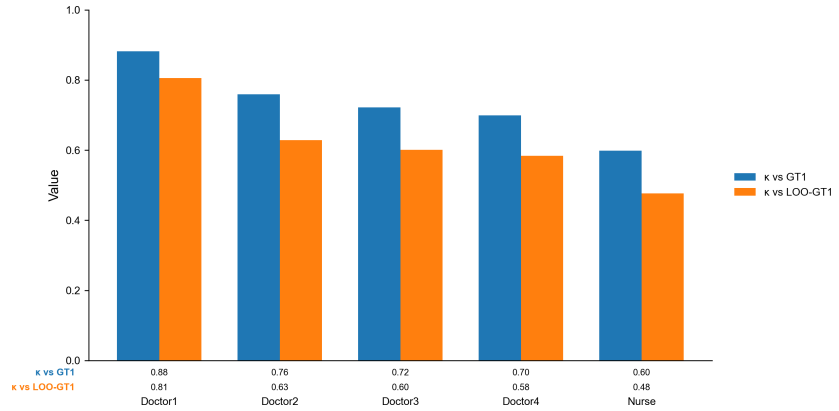


Table 6. LOO-GT1 incorporation-bias correction (human raters). For each human rater, agreement with GT1 (full 5-rater consensus) and with LOO-GT1 (same algorithm but with the rater's own score excluded). Δ quadratic-weighted $\varkappa$ is the inflation attributable to incorporation of the rater's own score in GT1. Δ computed on unrounded values.

Rater	N	Acc vs GT1	$\varkappa$_q vs GT1	$\varkappa$_u vs GT1	Acc vs LOO-GT1	$\varkappa$_q vs LOO-GT1	$\varkappa$_u vs LOO-GT1	Δ$\varkappa$_q
Doctor 1	305	0.83	0.88	0.74	0.73	0.81	0.59	+0.08
Doctor 2	302	0.66	0.76	0.50	0.55	0.63	0.35	+0.13
Doctor 3	305	0.62	0.72	0.45	0.54	0.60	0.34	+0.12
Doctor 4	305	0.59	0.70	0.35	0.49	0.58	0.22	+0.11
Nurse	292	0.61	0.60	0.36	0.51	0.48	0.23	+0.12

Figure 4. Incorporation-bias correction (LOO-GT1). For each of the 5 human raters (doctors 1-4 and the nurse), the quadratically-weighted Cohen κ against the full consensus (ground truth 1, blue) is shown alongside the κ against the leave-one-out variant (LOO-GT1, orange) in which the rater's own score has been excluded from the reference. The drop quantifies the inflation attributable to incorporation of the rater's own score in GT1; the mean Δκ_q is 0.11 (range 0.08-0.13). Large language models are not part of the GT1 reference and therefore require no leave-one-out variant correction; their statistics are unchanged from Figure 3. Sorted by κ vs leave-one-out variant of GT1 descending. Numerical values are reported in Table 6.

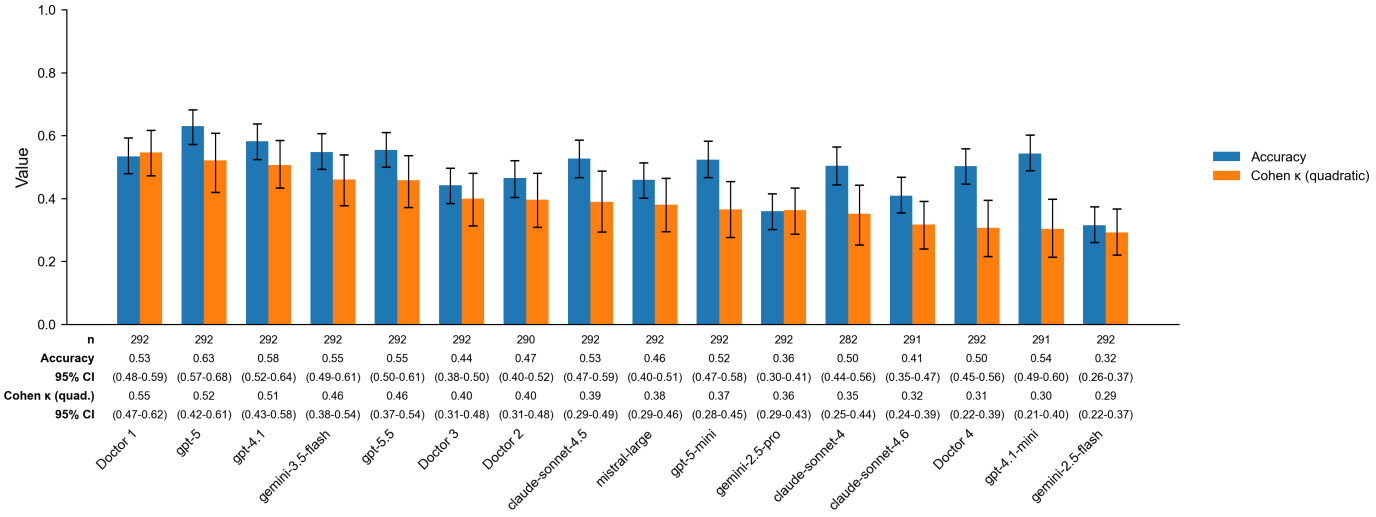


Agreement with the nurse triage as the reference standard (GT2) is provided in Table 7 and visualized in Figure 5. Across the 12 LLMs and 4 physicians, accuracies vs GT2 ranged 0.32-0.63 and quadratic-weighted κ ranged 0.29-0.55 (unweighted κ 0.06-0.35) on 282-292 paired samples. Doctor 1 achieved the highest agreement against GT2 (accuracy 0.53, weighted κ 0.55). Among the 12 LLMs, gpt-5 ranked highest (accuracy 0.63, weighted κ 0.52), followed by gpt-4.1, gemini-3.5-flash, and gpt-5.5 (weighted κ 0.46-0.51). Lower agreement against GT2 was observed for gemini-2.5-flash (accuracy 0.32, weighted κ 0.29) and claude-sonnet-4.6 (accuracy 0.41, weighted κ 0.32). Absolute κ values were generally lower against GT2, and the ordering of raters and models differed between the 2 references. All exact values, denominators, and 95% bootstrap CIs are listed in Table 7.

Table 7. Triage agreement vs nurse triage. Accuracy, Cohen κ, and sample sizes for all physicians and models using the triage nurse's classification as the reference standard.

Rater or model	N	Accuracy (95% CI)	Quadratic-weighted κ (95% CI)	Unweighted κ
Doctor 1	292	0.53 (0.48-0.59)	0.55 (0.47-0.62)	0.28
gpt-5	292	0.63 (0.57-0.68)	0.52 (0.42-0.61)	0.35
gpt-4.1	292	0.58 (0.52-0.64)	0.51 (0.43-0.58)	0.30
gemini-3.5-flash	292	0.55 (0.49-0.61)	0.46 (0.38-0.54)	0.28
gpt-5.5	292	0.55 (0.50-0.61)	0.46 (0.37-0.54)	0.27
Doctor 3	292	0.44 (0.38-0.50)	0.40 (0.31-0.48)	0.19
Doctor 2	290	0.47 (0.40-0.52)	0.40 (0.31-0.48)	0.20
claude-sonnet-4.5	292	0.53 (0.47-0.59)	0.39 (0.29-0.49)	0.25
mistral-large	292	0.46 (0.40-0.51)	0.38 (0.29-0.46)	0.16
gpt-5-mini	292	0.52 (0.47-0.58)	0.37 (0.28-0.45)	0.24
gemini-2.5-pro	292	0.36 (0.30-0.41)	0.36 (0.29-0.43)	0.11
claude-sonnet-4	282	0.50 (0.44-0.56)	0.35 (0.25-0.44)	0.20
claude-sonnet-4.6	291	0.41 (0.35-0.47)	0.32 (0.24-0.39)	0.14
Doctor 4	292	0.50 (0.45-0.56)	0.31 (0.22-0.39)	0.15
gpt-4.1-mini	291	0.54 (0.49-0.60)	0.30 (0.21-0.40)	0.25
gemini-2.5-flash	292	0.32 (0.26-0.37)	0.29 (0.22-0.37)	0.06

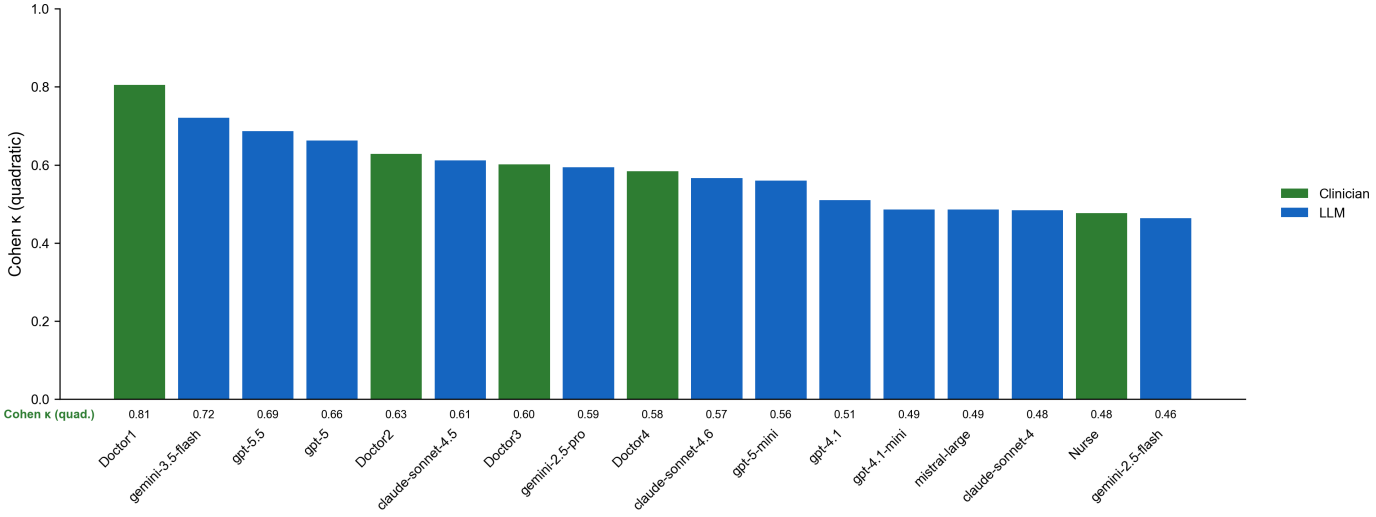
Figure 5. Agreement with ground truth 2. Accuracy and Cohen κ values for physicians and models using the triage nurse's classification as the reference.



Incorporation-Bias-Corrected Agreement (LOO-GT1)

Because the 5 human raters contribute to GT1, their measured agreement is partly circular. Recomputing GT1 per encounter with each evaluated rater's own score excluded (LOO-GT1) reduced quadratically-weighted κ for human raters by a mean of 0.11 (range 0.08-0.13; Table 6, Figure 4). After this correction, human-rater quadratic-weighted κ ranged 0.48-0.81 (accuracy 0.49-0.73). Doctor 1 retained the highest agreement (accuracy 0.73, weighted κ 0.81); the values for Doctor 2 (weighted κ 0.63), Doctor 3 (weighted κ 0.60), Doctor 4 (weighted κ 0.58), and the nurse (weighted κ 0.48) declined into a range overlapping with several LLMs. Moreover, 3 LLMs (gemini-3.5-flash weighted κ 0.72, gpt-5.5 weighted κ 0.69, and gpt-5 weighted κ 0.66) exceeded the LOO-corrected agreement of all human raters except Doctor 1. LLM statistics were, by construction, unchanged from those reported in Figure 3 because LLMs were not part of the GT1 reference and therefore did not require LOO correction. The relative ranking among LLMs was preserved across both references. The complete LOO-corrected ranking across all 5 human raters and 12 LLMs is presented in Figure 6.

Figure 6. Incorporation-bias-corrected ranking of human raters and large language models against ground truth 1 reference. For human raters, quadratically weighted Cohen κ is computed against LOO-GT1 (the leave-one-out variant with the evaluated rater's own score excluded; green); for large language models, κ is unchanged from Figure 3 because large language models are not part of the ground-truth 1 reference (blue). All 17 agents (5 human raters+12 large language models) are sorted by κ descending. Only Doctor 1 (a specialist physician) retains agreement clearly above any large language model after the leave-one-out correction; 3 contemporary large language models (gemini-3.5-flash, gpt-5.5, gpt-5) sit above the other clinicians and the nurse triage; the remaining large language models intersperse with the lower-performing clinicians. LLM: large language model.



Because Doctor 4's triage distribution differed materially from the other 4 raters, particularly for the red category (39 assignments vs 0-4 for all other human raters), we recomputed GT1 using the same central-tendency rule applied to the 4 remaining raters (nurse+Doctor 1-3; GT1_alt). Compared with the original GT1, weighted κ increased for 10 of the 12 LLMs, with a mean increase of 0.02 and individual changes ranging from –0.01 to +0.05. Agreement decreased only for gpt-5 (–0.01) and gpt-4.1 (–0.00). The 3 leading models remained gemini-3.5-flash (κ=0.72), gpt-5.5 (κ=0.70),

and gpt-5 (ϰ=0.66; Figure S1 and Table S1 in Multimedia Appendix 1). The direction of the shift indicates that Doctor 4's distinctive ratings moved the original consensus toward the LLM outputs, so the GT1_alt-based ranking should be regarded as a mildly LLM-favorable estimate; the GT1 provides a more conservative estimate of the LLM performance.

Statistical Comparison of Top Agents

Prespecified pairwise paired-bootstrap tests of quadratic-weighted ϰ vs GT1 are presented in Table 4. After Holm-Bonferroni correction, Doctor 1 showed higher agreement than gemini-3.5-flash (Δϰ +0.16), gpt-5.5 (Δϰ +0.19), and gpt-5 (Δϰ +0.22; all Holm-adjusted $P<.001$). Gemini-3.5-flash also showed higher agreement than the operational triage comparator (Nurse [GT2]; Δϰ +0.13, 95% CI 0.03-0.22; Holm-adjusted $P=.04$). Differences between gemini-3.5-flash and gpt-5.5, between gemini-3.5-flash and gpt-5, and between Doctor 2 and either gemini-3.5-flash or gpt-5.5 did not remain statistically significant after Holm correction. Full pairwise results are reported in Table 4.

Sensitivity Analysis: Completeness of Vital Signs

Sample-level completeness varied across vital signs (SBP available in 255/305 encounters, diastolic in 254, pulse in 256, body temperature in 259, SpO_2 in 257, respiratory rate in 259, and qSOFA in 232); pain scale and GCS were available for the subgroups in which they were documented (n=138 and n=63, respectively). A prespecified sensitivity analysis was therefore conducted in which each model's quadratically-weighted ϰ vs GT1 was recomputed in the subgroup with both SBP and body temperature present (n=250) and compared with the subgroup with at least 1 of these vital signs missing (n=55; Figure S2 and Table S2 in Multimedia Appendix 1). ϰ was higher in the incomplete-vitals subgroup for 8 of 12 models and lower for 4 of 12 models. The largest decreases were observed for gpt-5 (Δϰ –0.13) and mistral-large (Δϰ –0.11), whereas the largest increases were observed for claude-sonnet-4.6 (Δϰ +0.16), gemini-2.5-pro (Δϰ +0.15), and gpt-4.1-mini (Δϰ +0.14), where Δϰ denotes incomplete minus complete. Model rankings differed between subgroups. The 3 leading models in the complete-vitals subgroup were gemini-3.5-flash, gpt-5, and gpt-5.5, whereas the leading models in the incomplete-vitals subgroup were gpt-5.5, gemini-2.5-pro, and claude-sonnet-4.6. These results should be interpreted as exploratory given the small incomplete-vitals subgroup (Figure S2 and Table S2 in Multimedia Appendix 1).

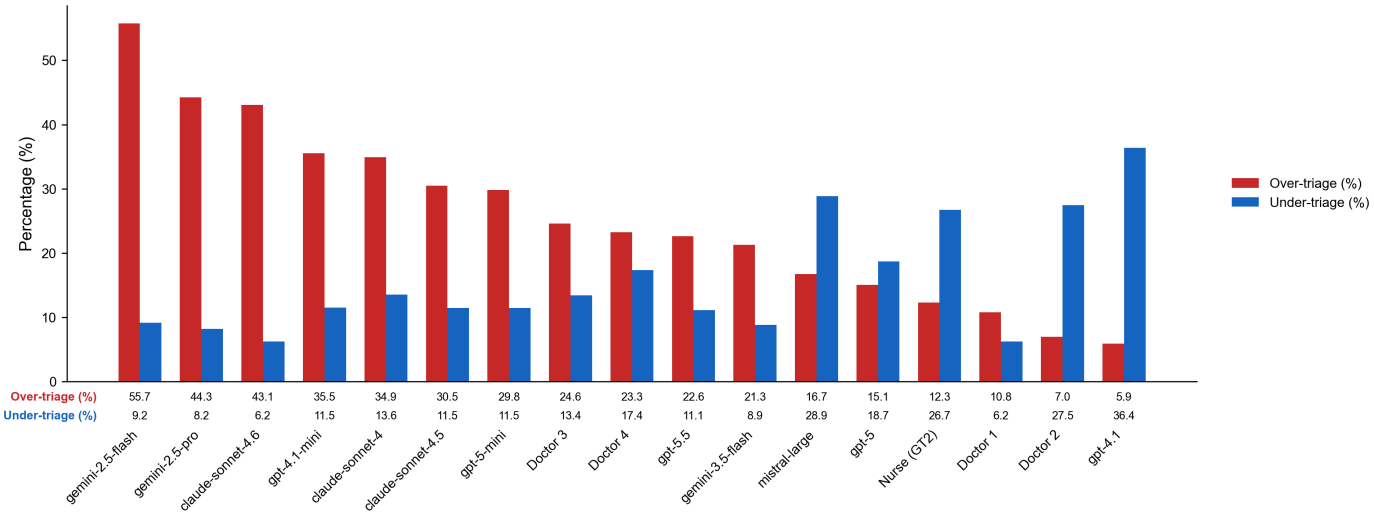
Directional Triage Error

Overtriage, undertriage, one-off agreement ($|\Delta| \leq 1$ category), and mean severity difference values are provided in Table 8 and Figure 7. Overtriage proportions ranged from 5.9% (gpt-4.1) to 55.74% (gemini-2.5-flash). Undertriage proportions ranged from 6.23% (Doctor 1) to 36.39% (gpt-4.1). One-off agreement values ranged from 93.44% to 100% across all raters and models (n=292-305). Mean score differences ranged from –0.37 (gpt-4.1) to +0.52 (gemini-2.5-flash), reflecting the distribution of error relative to the ground-truth ordinal category. Full numeric values for each rater are provided in Table 8.

Table 8. Overtriage (more urgent than GT1), under-triage (less urgent), one-off agreement (difference of ≤1 category), and the mean ordinal difference are reported for all clinicians and models. This table describes the direction and magnitude of deviations from the reference triage classification.

Rater or model	N	Overtriage (%)	Undertriage (%)	One-off agreement (%)	Mean ordinal Δ
gemini-2.5-flash	305	55.74	9.18	93.44	+0.52
gemini-2.5-pro	305	44.26	8.20	94.43	+0.39
claude-sonnet-4.6	304	43.09	6.25	95.72	+0.41
gpt-4.1-mini	304	35.53	11.51	97.70	+0.26
claude-sonnet-4	295	34.92	13.56	97.97	+0.23
claude-sonnet-4.5	305	30.49	11.48	98.03	+0.19
gpt-5-mini	305	29.84	11.48	95.08	+0.22
Doctor 3	305	24.59	13.44	97.05	+0.14
Doctor 4	305	23.28	17.38	96.07	+0.10
gpt-5.5	305	22.62	11.15	98.36	+0.12
gemini-3.5-flash	305	21.31	8.85	97.70	+0.13
mistral-large	305	16.72	28.85	94.10	-0.15
gpt-5	305	15.08	18.69	98.69	-0.04
Nurse (GT2)	292	12.33	26.71	98.29	-0.16
Doctor 1	305	10.82	6.23	100.00	+0.05
Doctor 2	302	6.95	27.48	99.34	-0.21
gpt-4.1	305	5.90	36.39	94.43	-0.37

Figure 7. Directional triage error vs ground truth 1. Rates of over-triage and under-triage for clinicians and models relative to ground truth 1. Values reflect the proportion of encounters assigned to a higher or lower urgency category than the reference.

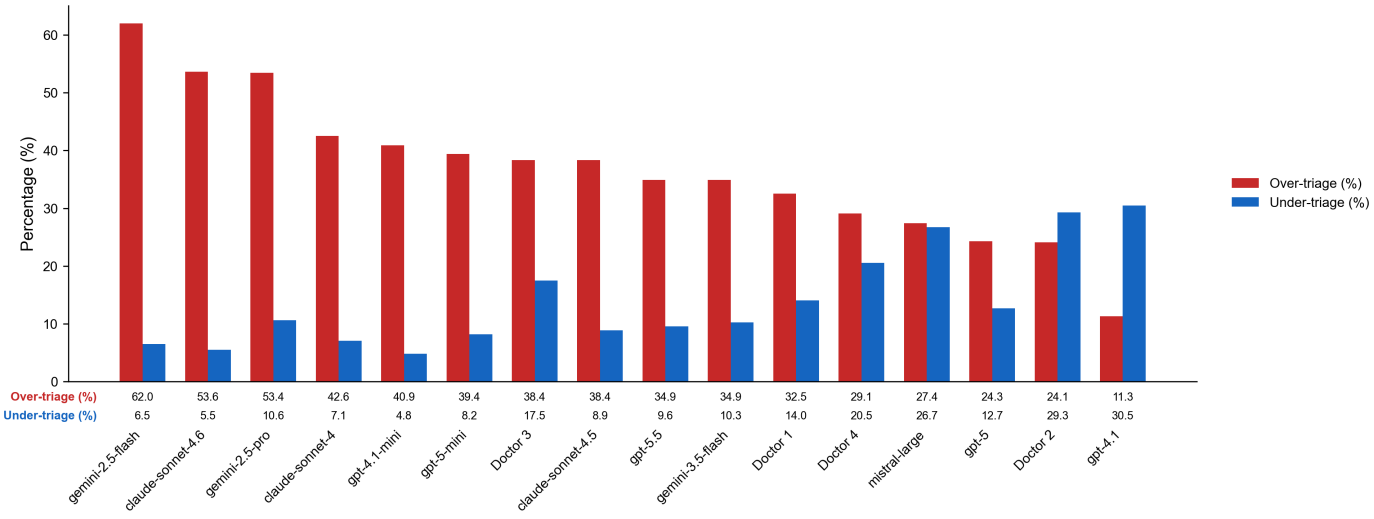


Overtriage, undertriage, one-off agreement, and mean score differences using the nurse reference (GT2) appear in Table 9 and Figure 8. Overtriage proportions ranged 11.3% (gpt-4.1) to 62% (gemini-2.5-flash); undertriage proportions ranged 4.8% (gpt-4.1-mini) to 30.5% (gpt-4.1); and one-off agreement remained ≥86% for all agents (Table 9). Mean ordinal differences ranged –0.21 (gpt-4.1) to +0.69 (gemini-2.5-flash), reflecting the same rater- and model-specific directional signatures observed against GT1.

Table 9. Directional triage error vs GT2. Overtriage, undertriage, one-off agreement, and mean ordinal difference for all physicians and models relative to the nurse triage assignment.

Rater or model	N	Overtriage (%)	Undertriage (%)	One-off agreement (%)	Mean ordinal Δ
gemini-2.5-flash	292	61.99	6.51	86.64	+0.69
claude-sonnet-4.6	291	53.61	5.50	88.32	+0.59
gemini-2.5-pro	292	53.42	10.62	86.64	+0.56
claude-sonnet-4	282	42.55	7.09	96.45	+0.39
gpt-4.1-mini	291	40.89	4.81	94.50	+0.42
gpt-5-mini	292	39.38	8.22	91.78	+0.39
Doctor 3	292	38.36	17.47	88.70	+0.31
claude-sonnet-4.5	292	38.36	8.90	93.84	+0.36
gpt-5.5	292	34.93	9.59	95.21	+0.29
gemini-3.5-flash	292	34.93	10.27	93.84	+0.30
Doctor 1	292	32.53	14.04	96.58	+0.22
Doctor 4	292	29.11	20.55	86.99	+0.26
mistral-large	292	27.40	26.71	96.23	+0.00
gpt-5	292	24.32	12.67	98.29	+0.12
Doctor 2	290	24.14	29.31	93.10	-0.06
gpt-4.1	292	11.30	30.48	97.95	-0.21

Figure 8. Directional triage error vs ground truth 2. Rates of over-triage and under-triage for clinicians and models relative to the nurse's triage assignment.



Clinician Interobserver Agreement

Pairwise Cohen κ values among clinicians are displayed in Table 10. Across 290-305 paired ratings, physician-physician quadratic weighted κ values ranged from 0.45 (Doctor 3 vs Doctor 4) to 0.72 (Doctor 1 vs Doctor 2). Nurse-physician quadratic weighted κ values ranged from 0.31 to 0.55. Exact κ values, sample sizes, and rater pairs are reported in Table 10.

Table 10. Pairwise interobserver agreement among clinicians. Cohen κ values for all rater pairs (nurse and 4 physicians) based on encounters with complete dual ratings. This table reflects interclinician variability in triage assignment under real-world emergency-department conditions.

Rater A	Rater B	N	Quadratic-weighted κ
Doctor 1	Doctor 2	302	0.72
Doctor 1	Doctor 3	305	0.69
Doctor 1	Doctor 4	305	0.64
Doctor 1	Nurse	292	0.55
Doctor 2	Doctor 3	302	0.57
Doctor 2	Doctor 4	302	0.53
Doctor 2	Nurse	290	0.40
Doctor 3	Doctor 4	305	0.45
Doctor 3	Nurse	292	0.40
Doctor 4	Nurse	292	0.31

Intermodel Agreement (Triage)

Pairwise agreement among model outputs is summarized in Figure 9 and Table 11. Figure 9 extends this view to all 18 agents (humans and LLMs) as a symmetric heatmap, allowing visual comparison of inter-LLM agreement against the human-rater submatrix. Across 294-305 paired classifications, quadratic weighted κ coefficients ranged from 0.39 (gpt-4.1 vs gemini-2.5-flash) to 0.82 (gpt-5.5 vs gemini-3.5-flash). Detailed κ values for all provider and model combinations are provided in Table 11 and Figure 9.

Figure 9. Pairwise quadratic-weighted Cohen κ for all raters and models. Symmetric heatmap of pairwise agreement between every pair of the 18 raters and models (ground truth 1, ground truth 2, 4 physicians, 12 large language models). The diagonal is fixed at 1.00 (self-comparison). The same data restricted to the 12 large language models are presented in tabular form in Table 11. Across the 66 large language model-large language model pairs, κ_q ranges 0.39 (gpt-4.1 vs gemini-2.5-flash) to 0.82 (gpt-5.5 vs gemini-3.5-flash).

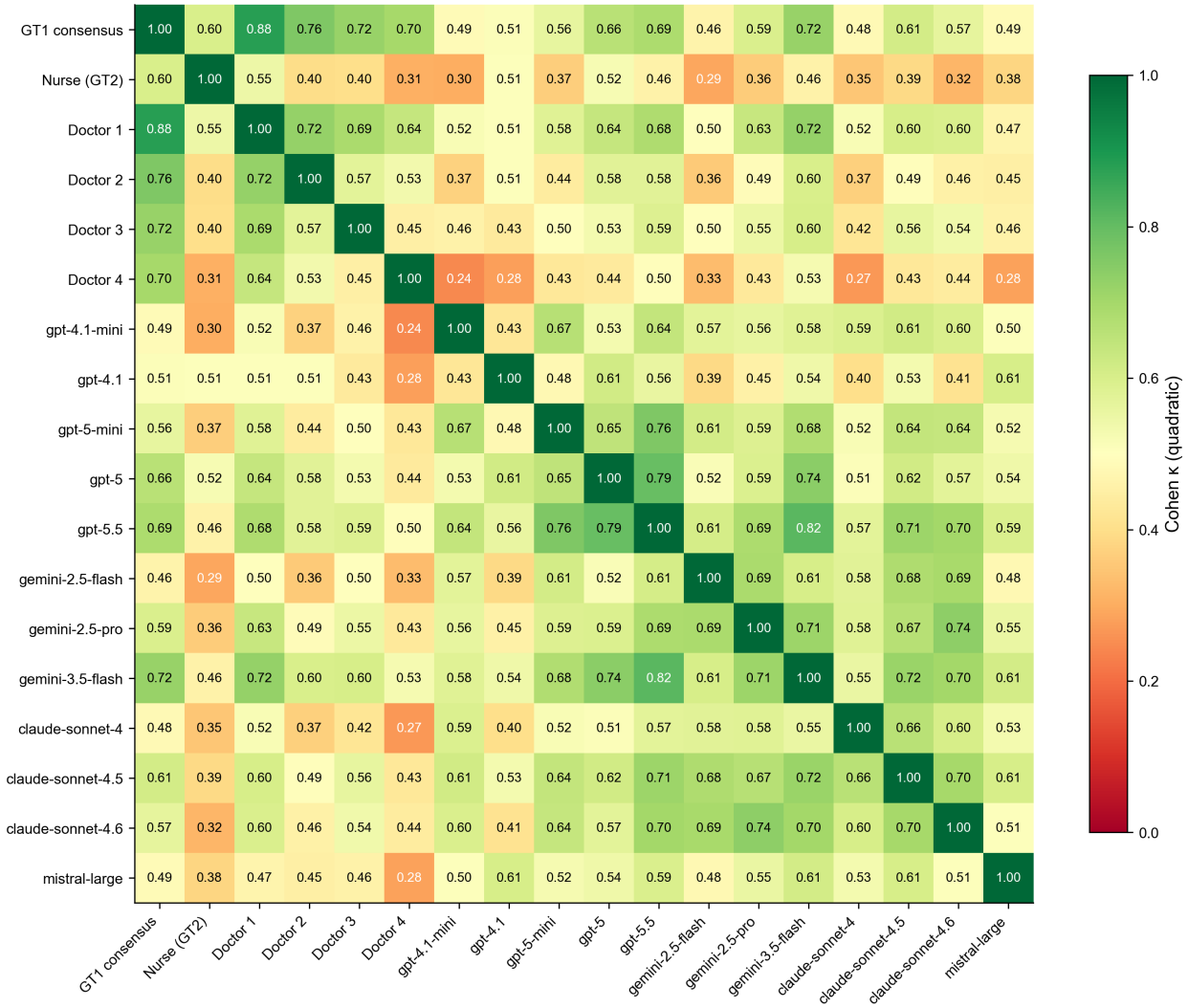


Table 11. Pairwise agreement among AI models. Cohen κ coefficients comparing triage outputs between all AI model pairs. This table describes concordance patterns within and across model families.

	gpt-4.1-mini	gpt-4.1	gpt-5-mini	gpt-5	gpt-5.5	gemini-2.5-flash	gemini-2.5-pro	gemini-3.5-flash	claude-sonnet-4	claude-sonnet-4.5	claude-sonnet-4.6	mistral-large
gpt-4.1-mini	—[a]	0.43	0.67	0.53	0.64	0.57	0.56	0.58	0.59	0.61	0.60	0.50
gpt-4.1		—	0.48	0.61	0.56	0.39	0.45	0.54	0.40	0.53	0.41	0.61
gpt-5-mini			—	0.65	0.76	0.61	0.59	0.68	0.52	0.64	0.64	0.52
gpt-5				—	0.79	0.52	0.59	0.74	0.51	0.62	0.57	0.54
gpt-5.5					—	0.61	0.69	0.82	0.57	0.71	0.70	0.59
gemini-2.5-flash						—	0.69	0.61	0.58	0.68	0.69	0.48
gemini-2.5-pro							—	0.71	0.58	0.67	0.74	0.55
gemini-3.5-flash								—	0.55	0.72	0.70	0.61
claude-sonnet-4									—	0.66	0.60	0.53
claude-sonnet-4.5										—	0.70	0.61
claude-sonnet-4.6											—	0.51
mistral-large												—

[a]Not applicable.

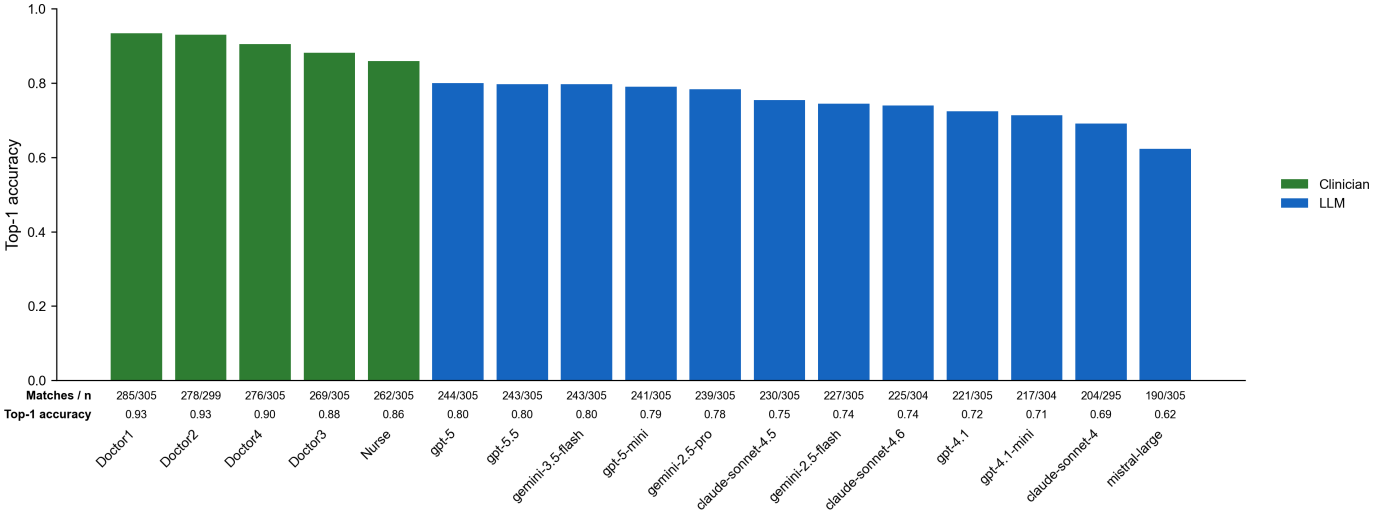
Specialty Assignment

Specialty top-1 accuracy for clinicians and models is presented in Table 12 and Figure 10 (n=295-305 across raters and models; Doctor 2=299, claude-sonnet-4=295, all others 304-305). Physician accuracies ranged from 0.88 (Doctor 3) to 0.93 (Doctor 1). Nurse accuracy was 0.86. Among models, gpt-5, gpt-5.5, and gemini-3.5-flash each achieved the highest specialty assignment accuracy (0.80), followed by gpt-5-mini (0.79) and gemini-2.5-pro (0.78). The lowest specialty accuracy was observed for mistral-large (0.62), with claude-sonnet-4 (0.69), gpt-4.1-mini (0.71), and gpt-4.1 (0.72) at the lower end of the LLM range. Detailed results are summarized in Table 12.

Table 12. Specialty assignment accuracy vs 5-rater modal specialty consensus. Top-1 accuracy for clinicians and models relative to the consensus specialty label, sorted by accuracy descending. Each value reflects the proportion of encounters for which the predicted specialty matched the reference.

Rater or model	N	Top-1 accuracy
Doctor 1	305	0.93
Doctor 2	299	0.93
Doctor 4	305	0.90
Doctor 3	305	0.88
Nurse	305	0.86
gpt-5	305	0.80
gpt-5.5	305	0.80
gemini-3.5-flash	305	0.80
gpt-5-mini	305	0.79
gemini-2.5-pro	305	0.78
claude-sonnet-4.5	305	0.75
gemini-2.5-flash	305	0.74
claude-sonnet-4.6	304	0.74
gpt-4.1	305	0.72
gpt-4.1-mini	304	0.71
claude-sonnet-4	295	0.69
mistral-large	305	0.62

Figure 10. Specialty assignment accuracy vs the 5-rater modal specialty consensus. Bar chart showing top-1 specialty accuracy for clinicians and AI models relative to the consensus specialty label. LLM: large language model.



Specialty top-1 accuracy vs GT2, the operational triage-pathway specialty assignment, is reported in Table 13, using all available paired labels. Physician accuracies remained in the range of 0.81-0.82 (Doctor 1 0.81, Doctor 2 0.82, Doctor 3 0.81, and Doctor 4 0.81). Across the 12 LLMs, accuracies ranged 0.57-0.75—gpt-5.5 (0.75) and gemini-3.5-flash (0.75) achieved the highest specialty accuracy against GT2, followed by gemini-2.5-pro (0.73), gpt-5 (0.73), and gpt-5-mini (0.73) while mistral-large was lowest (0.57). Detailed values and denominators are in Table 13.

Table 13. Specialty assignment accuracy vs the operational specialty assignment.

Rater or model	N	Top-1 accuracy
Doctor 2	299	0.82
Doctor 4	305	0.81
Doctor 1	305	0.81
Doctor 3	305	0.81
gpt-5.5	305	0.75
gemini-3.5-flash	305	0.75
gemini-2.5-pro	305	0.73
gpt-5-mini	305	0.73
gpt-5	305	0.73
claude-sonnet-4.5	305	0.69
gemini-2.5-flash	305	0.68
claude-sonnet-4.6	304	0.66
gpt-4.1-mini	304	0.65
gpt-4.1	305	0.65
claude-sonnet-4	295	0.64
mistral-large	305	0.57

Because each human rater contributed to the 5-rater modal specialty consensus, a leave-one-out sensitivity analysis was performed for the human raters. After excluding each evaluated rater from their respective reference, physician specialty accuracy ranged from 0.86 to 0.90—Doctor 1: 0.90, Doctor 2: 0.90, Doctor 3: 0.86, and Doctor 4: 0.88. The operational triage-pathway assignment had a leave-one-out accuracy of 0.83. The reduction relative to the full consensus was 0.02-0.03.

Diagnostic Recommendation Overlap vs Real Diagnostics

Diagnostics performed during the ED encounter (mapped to the same 26-item enum vocabulary provided to the LLMs, refer to Methods section) were compared with each model's recommended diagnostics on a per-encounter basis (n=277 encounters with nonempty real-diagnostics list; Table 14). Mean per-case F_1-scores ranged from 0.55 to 0.69. The highest F_1-score was observed for gemini-3.5-flash (0.69), followed by gpt-4.1 (0.66), and by gemini-2.5-flash (0.65). Precision and recall values for all models are detailed in Table 14.

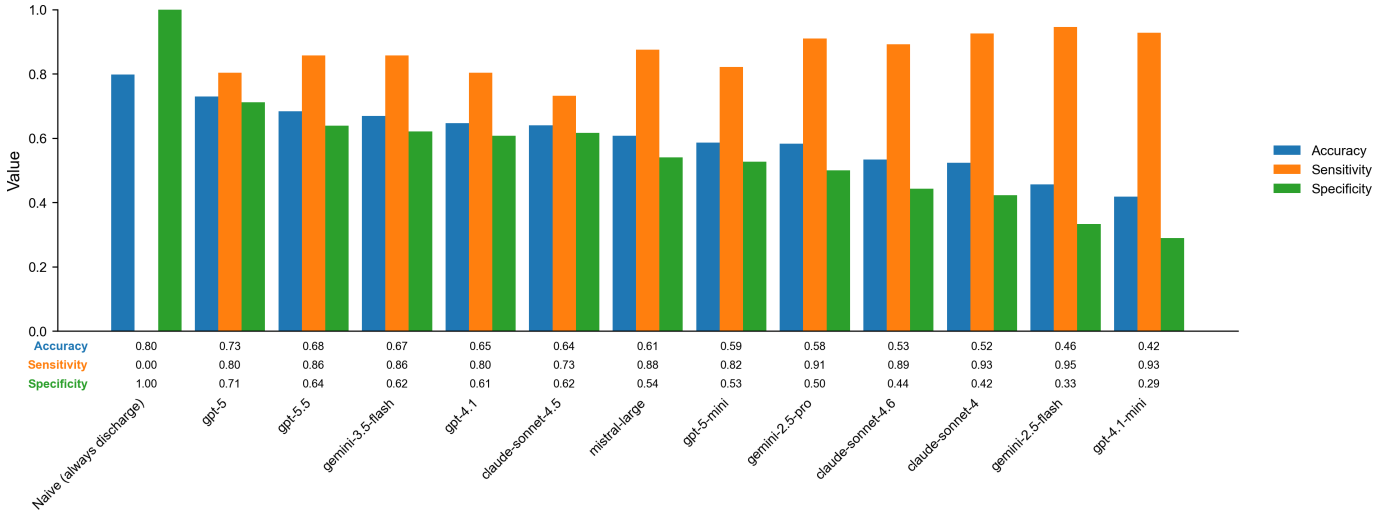
Table 14. Diagnostic recommendation performance vs real diagnostics. Mean per-encounter precision, recall, and F_1-scores for each model's diagnostic recommendations compared with the diagnostics performed in the emergency department. Precision represents the proportion of model recommended diagnostic items that were performed in routine care; recall reflects the proportion of performed diagnostic items included among the model's recommendations.

Model	N	Mean precision	Mean recall	Mean F_1-score
gemini-3.5-flash	277	0.69	0.78	0.69
gpt-4.1	277	0.70	0.71	0.66
gemini-2.5-flash	277	0.64	0.76	0.65
gpt-5.5	277	0.61	0.79	0.64
gpt-4.1-mini	276	0.67	0.68	0.64
claude-sonnet-4	268	0.64	0.69	0.63
gpt-5	277	0.58	0.81	0.62
claude-sonnet-4.5	277	0.60	0.73	0.62
gemini-2.5-pro	277	0.58	0.75	0.62
mistral-large	277	0.66	0.58	0.59
gpt-5-mini	277	0.50	0.80	0.58
claude-sonnet-4.6	276	0.52	0.67	0.55

Admission Decision vs Real Outcome

Admission prediction metrics for each model appear in Table 2 and Figure 11. Model accuracy ranged from 0.42 (gpt-4.1-mini) to 0.73 (gpt-5). GPT-5 achieved the highest admission prediction accuracy (0.73), as well as high sensitivity (0.80) and negative predictive value (0.93). Sensitivity across models ranged from 0.73 to 0.95, specificity from 0.29 to 0.71, PPV from 0.25 to 0.41, and NPV from 0.90 to 0.96. For reference, a naive baseline that predicts discharge for every encounter would yield an NPV equal to the discharge prevalence in the evaluable subsample (0.80; Table 2, "Naive (always discharge)" row). The model NPV improvement over this baseline is therefore approximately 10-16 percentage points, rather than the absolute 0.90-0.96 figure considered in isolation. Complete performance numbers under the primary threshold (≥POSSIBLE_ADMISSION→admit) are presented in Table 2.

Figure 11. Admission prediction performance vs real outcome. Accuracy, sensitivity, and specificity of each admission-prediction model compared with actual admission and discharge outcomes.



Discussion

Principal Findings

This study has 4 principal findings. First, in this real-world walk-in ED cohort, an experienced clinician outperformed all 12 contemporary LLMs on urgency triage, under both GT1 and under an incorporation-bias-corrected reference (LOO-GT1). Under LOO-GT1, 3 contemporary LLMs (gemini-3.5-flash, gpt-5.5, and gpt-5) reached agreement that exceeded that of the nurse and 3 of the 4 physicians.

Second, clinicians remained clearly ahead of all LLMs for specialty assignment (0.88-0.93 vs ≤0.80). Third, all evaluated LLMs showed high sensitivity (0.73-0.95) and NPV (0.90-0.96) for admission prediction at the screening threshold, but modest PPV (0.25-0.41). Fourth, directional triage errors (over- vs undertriage) showed rater- and model-specific signatures that were stable across both reference standards, suggesting systematic calibration patterns rather than random error.

By using both a consensus human reference (GT1) and the operationally relevant nurse triage (GT2) as reference standards, the analysis offers a dual perspective on performance, including experienced clinicians and the real frontline practice. However, consensus ground truth is not identical to strict MTS application, particularly in pain-triggered pathways. Clinicians commonly contextualize MTS trigger categories (eg, orange for chest pain or dyspnea, or yellow for moderate pain or visual analog scale 4-7) through additional history and clinical impression, which can shift classifications toward lower urgency and thereby resemble undertriage relative to a rule-based reference. In total, the results highlight several consistent patterns across triage, specialty assignment, diagnostic recommendations, and admission prediction that could be guiding medical practice in emergency care in the future.

All findings are framed against empirical reference standards (a consensus human label and the operational nurse triage) rather than outcome-anchored truth. The LOO-GT1 correction addresses incorporation bias but does not transform either reference into an absolute clinical gold standard.

Triage Agreement and Interrater Variability

Under GT1 reference, Doctor 1-3 had weighted κ point estimates at or marginally above that of the leading LLM, whereas Doctor 4 had a lower estimate than gemini-3.5-flash. Against GT2, only Doctor 1 retained the highest agreement point estimate above every LLM, while gpt-5, gpt-4.1, gemini-3.5-flash, and gpt-5.5 exceeded the agreement of Doctors 2-4. The highest performing physician achieved substantial concordance (weighted $\kappa \approx 0.88$, unweighted $\kappa \approx 0.74$ vs GT1; $\kappa \approx 0.55$ vs GT2). Yet the data also reveal a wide internal range among clinicians themselves; pairwise physician-physician weighted κ values spanned 0.45-0.72 (Doctor 3 vs Doctor 4 to Doctor 1 vs Doctor 2), and nurse-physician κ values remained in the 0.31-0.55 range. This degree of variability reflects findings in prior MTS

research, where interrater reliability is often moderate even with protocolized frameworks and structured discriminators.

Because GT1 incorporates the human rating being evaluated and thus GT1-based comparison being partly circular, LOO-GT1 was calculated. Under this incorporation-bias-corrected reference, only Doctor 1 retained agreement clearly above any LLM (weighted $\varkappa$ 0.81). The agreement values for the other physicians and nurse dropped into a range that overlapped with several of the LLMs evaluated. The accurate framing is task- and rater-specific; for urgency triage, Doctor 1 had a higher LOO-GT1 point estimate than all evaluated LLMs; the best contemporary LLMs reached agreement at the upper end of the interclinician range observed in the same cohort, equaling or modestly exceeding the other clinicians and the nurse triage under the incorporation-bias-corrected reference (LOO-GT1), while for specialty assignment all clinicians remained clearly ahead of the best LLM. The performance of the best AI systems with $\varkappa$ values around 0.60-0.72 falls within the spectrum of human variation rather than outside it, as published before by de Souza et al [22]. This supports a complementary human-in-the-loop rather than autonomous deployment. It is important to emphasize that neither GT1 nor GT2 is a true gold standard for ED triage. Both references should therefore be interpreted as empirical reference standards rather than ground truth in the strict outcome-based sense. The consistency of the relative agent ranking across both references is therefore informative; it demonstrates that the differences observed between clinicians and LLMs are not an artifact of either reference's particular construction. Per-class F_1-scores stratified by triage category are presented in Figure S3 in Multimedia Appendix 1.

In the paired-bootstrap analysis against full GT1, Doctor 1 differed significantly from the 3 leading LLMs, and gemini-3.5-flash differed significantly from the operational triage comparator after Holm-Bonferroni correction. These comparisons must be interpreted cautiously because full GT1 incorporates the ratings of the human raters being evaluated. Table 4 therefore does not constitute an inferential test of clinician-LLM differences under LOO-GT1. Under LOO-GT1, Doctor 1 had the highest point estimate ($\varkappa$=0.81), compared with 0.72 for the leading LLM, but this comparison remains descriptive unless the paired-bootstrap analysis is repeated using the rater-appropriate LOO reference.

Directional Triage Error

One-off agreement exceeded 86% across all clinicians and models in every analysis. This means that most misclassifications occur at the border between adjacent MTS categories. From a clinical perspective, such category shifts often result in similar diagnostic and monitoring steps [23]. The critical dimension is therefore not the presence of disagreement, but its directionality. While overtriage has been reported to be associated with low-value care, Loftus et al [24] have shown a higher morbidity and mortality associated with undertriage [25]. Directional triage error strongly differs between raters and models; certain clinicians consistently assigned lower acuity than the reference (systematic undertriage), while others assigned higher acuity (systematic overtriage). The same was true of models. These consistent signatures suggest underlying behavioral patterns or bias rather than random error. Human undertriage is often associated with experience level, cognitive load, or atypical presentations, whereas model under- or overtriage likely reflects thresholding and training-data distributions [23]. Undertriage is a patient-safety threat; overtriage, while safer, aggravates crowding, resource strain, and staff workload. Thus, the stability of these signatures across both reference standards highlights the need for explicit calibration and governance frameworks to align both human and AI decision-makers toward consistent acuity signaling. A validated AI model could help address this, avoid bias, yet give reliable output in a high-acuity setting [26].

Specialty Assignment

Regarding specialty-assignment, humans showed very high contextual accuracy, consistently above 0.86, whereas the best LLM achieved 0.80. Specialty assignment is a task rooted not only in symptom description but also in pattern recognition, institutional knowledge, and interdisciplinary nuances, but also institutional definitions. Across both reference standards, clinicians remained ahead of all 12 LLMs (physician accuracy ranged from 0.88-0.93 vs best LLM 0.80). Against the operational triage-pathway specialty assignment, physician accuracy ranged from 0.81-0.82 vs best LLM 0.75. The size of this gap quantifies how much of the specialty-routing decision remains data-bound at the present time. Across triage and specialty assignment, the highest-ranking models were the most recent, reasoning-capable checkpoints (gemini-3.5-flash, gpt-5.5, and gpt-5), and the lowest-ranking were older or smaller variants (gpt-4.1-mini, gemini-2.5-flash, and mistral-large). This is an observational association. Model generation, parameter count, training-data scale, and reasoning-capability features co-vary across publicly released checkpoints and cannot be disentangled in this cohort. We therefore describe the pattern without inferring that any single architectural property (inference-time reasoning, longer context, or multistep generation) is causally responsible.

Diagnostic Recommendation

The diagnostic-recommendation analysis further supports this conclusion with an important reflection on the fact that the comparator (diagnostics actually performed during routine care) is not an outcome-validated gold standard. LLMs tend to deliver high-recall, low-precision suggestions capturing most diagnostics performed in the real-world setting, but also adding tests that were not ultimately ordered. This "coverage-first" behavior observed across all evaluated LLMs is descriptively consistent with either an intrinsic preference for higher recall in trained models or with the retrospective study design (which prevents the model from asking follow-up questions). The current data cannot distinguish these mechanisms, and the description should be read as an observed behavioral pattern against routine practice rather than against an optimal evidence-based work-up. Both higher and lower model F_1-score should be interpreted in this

context. High recall is clinically reassuring, but excess recommendations need to be curtailed to avoid unnecessary testing, patient burden, and cost escalation. Most importantly, a safe recommendation of diagnostics to be performed during waiting time can reduce waiting time and thereby increase patient safety and reduce physician workload [12,17,18,27]. The clinical utility of model-recommended diagnostics requires prospective validation.

Admission Prediction

Admission-prediction is a second task where LLMs may be useful. The admission rate of 20.1% is in line with what has been published before for walk-in patients and is hence representative of a regular German ED. The low admission prevalence has 2 consequences that should be borne in mind when interpreting the headline metrics. First, sensitivity and NPV can appear high even for weakly informative classifiers; second, PPV is inherently capped by the small admission base rate, so even a well-calibrated model in this setting will not achieve PPV close to 1. High sensitivity and high NPV are clinically relevant and valuable, particularly early in the ED encounter. The headline LLM numbers (sensitivity 0.73-0.95 and NPV 0.90-0.96) must be read against the appropriate baseline; a naive "always-discharge" classifier would already achieve NPV=0.80 simply from the imbalance between discharges and admissions. The model improvement over this baseline is approximately 10-16 percentage points of NPV. However, the modest PPV (~0.25-0.41; median 0.32, IQR 0.29-0.35) across all models indicates that they are not yet suitable for autonomous admission decisions. Instead, they can function as a prioritization tool for additional review, flagging patients who "may require admission" without preemptively committing resources.

The introduction of nurse-reference analyses provides additional operational insight. Because many EDs rely on nurse-led MTS triage with physician oversight only later, alignment with nurse distributions can determine how well an AI model integrates into workflow. Here, model performance was slightly lower against the nurse than against GT1, but the relative performance hierarchy remained unchanged. The relative ranking of agents is broadly preserved across both references (with the further refinements introduced under LOO-GT1 for human raters), which is reassuring evidence that the observed differences across agents reflect behavior consistent under multiple reasonable references rather than being driven by the construction of any single reference. From a deployment perspective, this stability enables targeted model calibration.

Human-AI Complementarity

An important conceptual aspect of these findings is that the MTS was designed for human use, not for AI systems. MTS relies on narrative flowcharts, implicit clinical reasoning, and subjective discriminators that align with human cognition. Applying this framework directly to AI may hence constrain model performance. Future systems may benefit from alternative representations of acuity that are better suited to machine reasoning, or from interactive AI systems that dynamically ask targeted follow-up questions. This represents a key motivation for prospective evaluation, where AI systems can adaptively gather missing information rather than relying solely on static triage notes.

Across all tasks, the data collectively support human-AI complementarity. AI models, even in their current form, offer consistency, full-case memory, and scalability; humans contribute contextual judgment, clinical nuance, and adaptive reasoning. Variability among human raters highlights where AI systems might stabilize decision-making or offer real-time quality control; differences between AI models and clinicians identify where interpretability, calibration, and local fine-tuning are needed. In this sense, AI should not be viewed as a replacement for clinical expertise but as a structural and operational support; a prescreening tool, a second-check layer, a signal-stability mechanism, and a risk-awareness tool [13]. Previous studies have shown physician engagement and excitement for AI support in the ED setting [28].

Limitations

This study has several limitations. First, it is a single-center dataset, and case mix may differ from other EDs, influencing triage distributions and $\varkappa$ values. Generalizability to other settings is therefore not guaranteed. The case mix at a tertiary-care center is enriched for higher-acuity, more complex presentations than at community or rural EDs and may underrepresent presentations seen elsewhere; MTS application is operator-dependent, and local training, departmental conventions, and standard-operating-procedure adaptations vary across institutions, which means the GT2 distribution and specialty assignment labels observed here may not be representative of nurse-triage distributions in other centers. The German health care context (eg, insurance-based access, separate ambulance triage pathway, and walk-in-only ED population studied here) differs systemically from other countries. Headline statements about clinician-vs-LLM performance and about model ranking should therefore be interpreted as calibrated within this institutional context, and would need revalidation in any new setting.

Second, due to the retrospective design of the study, reasons for presentation were solely taken from triage notes, and no additional questions for further assessment could be asked, and ground truth was defined through expert adjudication rather than outcome-based measures. GT1 incorporated the 5 human ratings being evaluated. We quantified the magnitude of this bias and report a leave-one-out variant. Residual bias may remain because human raters trained in the same institution share systematic priors that the leave-one-out calculation cannot eliminate.

Third, the triage nurse assigned MTS in real time at presentation, with the operational pressures of an ED (competing patient demands, partial documentation at the moment of decision), but also with direct clinical context unavailable to the retrospective raters (patient appearance, demeanor, breathing pattern, pain behavior, and overall clinical gestalt), which can complement the MTS algorithm with experiential gut-feeling judgment. The 4 retrospective

physicians and the 12 LLMs evaluated the same encounters from the fully extracted structured case file with identical information content, but without that direct clinical impression, and blinded to the nurse's triage assignment, to the other retrospective raters' assessments, to all diagnostics performed, and to the final disposition. The asymmetries therefore go in both directions—retrospective raters had no time pressure but no clinical gestalt; the nurse had clinical gestalt but operational pressure and incomplete structured data. Lower nurse agreement values should not be interpreted as evidence that operational MTS triage is inherently less reliable than retrospective expert review under structured data alone.

Fourth, the cohort comprises self-presenting walk-in patients only and excludes ambulance arrivals and resuscitation-bay activations, which in our institution follow a separate triage pathway. The resulting GT1 distribution contains only 35 orange and a single red encounter, mirroring the documented walk-in population case-mix in German EDs. Statements in this paper therefore characterize model and clinician performance on the lower-to-moderate acuity ED workflow, which represents 60%-70% of total ED volume but excludes the highest-acuity presentations. Additionally, the cohort's age distribution (median 45, IQR 31-62 y) concentrates the data on the working-age population in which MTS is best-validated. The small number of encounters involving patients younger than 18 years (n=15, 4.9%) and patients aged ≥75 years (n=35, 11.5%), groups in which MTS reliability is known to be reduced, was insufficient for powered subgroup analyses.

Fifth, diagnostic categorization was simplified, and models were evaluated as fixed snapshots. The diagnostics performed during each encounter were used as the comparator for model recommendations, but they are not a true gold standard. Practice patterns vary by department and clinician, where some tests may be ordered defensively without clinical necessity, while some clinically appropriate tests may be omitted under time pressure or operational constraints. Per-encounter precision, recall, F_1-score therefore measure alignment with routine practice in our institution, not alignment with an ideal evidence-based work-up.

Sixth, vital-sign completeness varied across encounters (SBP available in 255/305 and temperature in 259/305). Missing values were indicated to the models with a question mark and were not imputed. Residual confounding between vital-sign completeness and case complexity cannot be excluded with a retrospective design. Moreover, the subgroup with incomplete SBP or temperature measurements was small, and the observed differences and ranking changes should not be interpreted as causal effects of missing vital-sign data.

Seventh, the system prompt narrative descriptions (Multimedia Appendix 1) used human-readable labels (eg, "Admission Likelihood (Low/Mid/High)") that do not lexically match the enforced *Pydantic* schema keys and enum values. While the provider-native structured-output mechanism constrained every response to the schema regardless of prompt phrasing, alignment between prompt narrative and output schema is best practice.

Eighth, the original 9-model panel was evaluated in September 2025, and the 3 additional models were added during revision and queried in May 2026, an 8-month gap. Of the 12 models, 11 were called via either explicitly dated API snapshots or stable IDs that providers document as nonrolling, with the one exception, mistral-large-latest, for which a dated pin was available but not used. No provider-returned version fingerprint was logged at call time. Direct cross-batch model comparisons, in particular any ranking involving mistral-large-latest, should be interpreted with this caveat in mind. Comparisons against the fixed human reference standards are unaffected by this limitation.

Ninth, the system prompt provided to the models included epidemiological priors on the expected distribution of Manchester Triage System categories and on admission rates by category (Multimedia Appendix 1), whereas the retrospective physicians received only the structured case information and relied on their clinical training and institutional familiarity rather than an equivalent written guide. The 2 conditions were therefore not fully information-symmetric with respect to contextual priors, which is an intrinsic feature of prompt design in comparative benchmarking rather than a departure from the study protocol. These priors did not correspond closely to the present cohort, in which green and blue together accounted for 58% of GT1 labels and the observed admission prevalence was 20.1%, and the models overpredicted both acuity and admission relative to the stated priors. Any effect on the reported human-model comparison is therefore unlikely to favor the models, but the asymmetry should be borne in mind when interpreting model calibration.

Finally, the dataset does not include time-to-provider, resource usage, or clinical outcomes that would help contextualize the operational impact of triage or diagnostic variation.

Future Directions

Future work should incorporate direct patient input in a prospective evaluation, dynamic recalibration, and integration into real-time ED decision pathways, as planned in the currently enrolling study "AI in Triage in the Emergency Department" (DRKS-ID DRKS00038034). Studies that combine LLM-based reasoning with structured vital-sign inputs and temporal modeling may improve accuracy and contextual sensitivity. With these foundations, AI-supported triage and assessment may help reduce variability, enhance diagnostic completeness, and improve patient flow in the high-pressure environment of the ED [13]. Quantitative analysis of the qualitative output fields, particularly the relationship between the model-reported confidence_rating and prediction correctness or directional triage error, was not within the scope of the present analysis but represents a low-cost extension of the existing dataset that will be part of the prospective study.

Conclusion

In a real-world, multitask, multirater ED triage cohort, an experienced physician outperformed contemporary LLMs for both urgency triage and specialty assignment. Under the incorporation-bias-corrected reference (LOO-GT1), 3 contemporary LLMs (gemini-3.5-flash, gpt-5.5, and gpt-5) reached agreement that exceeded that of the 4 other human raters. The best LLMs achieved quadratically weighted κ within the spectrum of interphysician variability and demonstrated clinically meaningful sensitivity and NPV for admission prediction. Modest PPV and persistent directional triage-error patterns argue against autonomous LLM-driven triage but support a complementary human-in-the-loop role. It can serve as a prescreening signal, a second-check layer, a consistency monitor, and a diagnostics-completeness aid, particularly in operationally pressured ED settings.

Generative AI Use

Generative AI tools were used for data visualization and manuscript formatting. All scientific content, analyses, and interpretations were performed by the authors.

Funding

This study received no external funding.

Data Availability

The analysis code and machine-readable source data are provided in Multimedia Appendix 2 and regenerate every reported table and figure with one documented exception. The matrix contains only the variables required to reproduce the reported analyses and excludes direct identifiers, encounter dates, source-system identifiers, clinical history, diagnoses, raw prompts and responses, and personnel names. To keep the deposited matrix de-identified, age, sex and the vital-sign values were replaced by constant placeholders and the vital-sign missingness pattern was not retained; Table 1, Table S2, and Figure S2 in Multimedia Appendix 1 therefore cannot be regenerated from the public file and are provided as computed output. The underlying clinical source data are not publicly available because of privacy and ethical restrictions.

Authors' Contributions

Conceptualization: NA, GD
Data curation: NA, MFvB, NK, SM, LH, H-PE, GD
Formal analysis: NA, MFvB, NK, SM, LH, H-PE, GD
Methodology: NA, GD
Writing – original draft: NA, MFvB, GD
Writing – review & editing: NK, SM, AS, LH, H-PE, SMH, AK-G, JG, IM, SZ
All authors have read and agreed to the published version of the manuscript.
The author MFvB is an Independent Researcher.

Conflicts of Interest

NA: Consultancy and/or speaker's bureau: AbbVie, Gilead, Ipsen, Novartis, all unrelated to presented work. Founder of AIxMedical UG, which did not fund this research or hold direct interest in the methods discussed. SZ: Speaker's bureau and/or consultancy: Abbvie, BioMarin, Boehringer Ingelheim, Gilead, GSK, Ipsen, Madrigal, MSD, NovoNordisk and SoBi, all unrelated to presented work. All other authors declare no conflicts of interest.

Multimedia Appendix 1

System prompt, supplementary tables, and figures. Sensitivity analyses excluding Doctor 4 (Table S1, Figure S1) and stratified by completeness of vital signs (Table S2, Figure S2), and per-class F_1-scores vs GT1 (Figure S3).
[DOCX File (Microsoft Word File), 367 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Analysis code and source data. Python script, derived encounter-level rating matrix, and machine-readable source data for all tables and figures.
[ZIP File (ZIP archive File), 2316 KB-Multimedia Appendix 2]

References

1. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: a systematic review of causes, consequences and solutions. PLoS One. 2018;13(8):e0203316. [doi: 10.1371/journal.pone.0203316] [Medline: 30161242]
2. Wise J. Risk of death more than doubles with long emergency waits, analysis shows. BMJ. 2025;388:r119. [doi: 10.1136/bmj.r119]
3. Christ M, Grossmann F, Winter D, Bingisser R, Platz E. Modern triage in the emergency department. Dtsch Arztebl Int. Dec 2010;107(50):892-898. [doi: 10.3238/arztebl.2010.0892] [Medline: 21246025]

4. Zachariasse JM, van der Hagen V, Seiger N, Mackway-Jones K, van Veen M, Moll HA. Performance of triage systems in emergency care: a systematic review and meta-analysis. BMJ Open. May 28, 2019;9(5):e026471. [doi: 10.1136/bmjopen-2018-026471] [Medline: 31142524]
5. Parenti N, Reggiani MLB, Iannone P, Percudani D, Dowding D. A systematic review on the validity and reliability of an emergency department triage scale, the Manchester Triage System. Int J Nurs Stud. Jul 2014;51(7):1062-1069. [doi: 10.1016/j.ijnurstu.2014.01.013] [Medline: 24613653]
6. Jesus A de, Batista REA, Campanharo CRV, Lopes M, Okuno MFP. Evaluation of the Manchester Triage System quality indicator: service time. Rev Gaucha Enferm. 2021;42:e20200371. [doi: 10.1590/1983-1447.2021.20200371] [Medline: 34878015]
7. Li B, Zhang Z, Li K, Deng Y. The effectiveness of a modified Manchester Triage System for geriatric patients: a retrospective quantitative study. Nurs Open. Sep 2024;11(9):e70024. [doi: 10.1002/nop2.70024] [Medline: 39231303]
8. Zachariasse JM, Seiger N, Rood PPM, et al. Validity of the Manchester Triage System in emergency care: a prospective observational study. PLoS One. 2017;12(2):e0170811. [doi: 10.1371/journal.pone.0170811] [Medline: 28151987]
9. Gries A, Schrimpf AM, Dercks N von. Hospital emergency departments—utilization and resource deployment in the hospital as a function of the type of referral (article in German). Dtsch Arztebl Int. Sep 23, 2022;119(38):640-646. [doi: 10.3238/arztebl.m2022.0276] [Medline: 35912425]
10. Niset A, El Hadwe S, Englebert A, Barrit S. AI in emergency medicine: building literacy or castles in the air. Am J Emerg Med. Jan 2025;87:145-146. [doi: 10.1016/j.ajem.2024.08.020] [Medline: 39198071]
11. Chen IY, Zink A. AI tools in human hands: measuring real-world impact in the emergency department. NEJM AI. Feb 27, 2025;2(3):AIe2500083. [doi: 10.1056/AIe2500083] [Medline: 41048339]
12. Gaber F, Shaik M, Allega F, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. NPJ Digit Med. May 9, 2025;8(1):263. [doi: 10.1038/s41746-025-01684-1] [Medline: 40346344]
13. Taylor RA, Chmura C, Hinson J, et al. Impact of artificial intelligence–based triage decision support on emergency department care. NEJM AI. Feb 27, 2025;2(3):AIoa2400296. [doi: 10.1056/AIoa2400296]
14. El Arab RA, Al Moosa OA. The role of AI in emergency department triage: an integrative systematic review. Intensive Crit Care Nurs. Aug 2025;89:104058. [doi: 10.1016/j.iccn.2025.104058] [Medline: 40306071]
15. Da'Costa A, Teke J, Origbo JE, Osonuga A, Egbon E, Olawade DB. AI-driven triage in emergency departments: a review of benefits, challenges, and future directions. Int J Med Inform. May 2025;197:105838. [doi: 10.1016/j.ijmedinf.2025.105838] [Medline: 39965433]
16. Masanneck L, Schmidt L, Seifert A, et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. J Med Internet Res. Jun 14, 2024;26:e53297. [doi: 10.2196/53297] [Medline: 38875696]
17. Arslan B, Nuhoglu C, Satici MO, Altinbilek E. Evaluating LLM-based generative AI tools in emergency triage: a comparative study of ChatGPT Plus, Copilot Pro, and triage nurses. Am J Emerg Med. Mar 2025;89:174-181. [doi: 10.1016/j.ajem.2024.12.024] [Medline: 39731895]
18. Haim GB, Saban M, Barash Y, et al. Evaluating large language model-assisted emergency triage: a comparison of acuity assessments by GPT-4 and medical experts. J Clin Nurs. Oct 2026;35(10):4249-4255. [doi: 10.1111/jocn.17490] [Medline: 39610042]
19. Brodeur PG, Buckley TA, Kanjee Z, et al. Performance of a large language model on the reasoning tasks of a physician. Science. Apr 30, 2026;392(6797):524-527. [doi: 10.1126/science.adz4433] [Medline: 42060751]
20. Van Der Haas Y, Roskamp W, Chang-Willems LEM, et al. Evaluating an AI decision support system for the emergency department: retrospective study. JMIR AI. Jan 26, 2026;5:e80448. [doi: 10.2196/80448] [Medline: 41587768]
21. BioRender. URL: https://www.biorender.com/ [Accessed 2026-11-10]
22. Souza C de, Chianca TCM, Cordeiro Júnior W, Rausch M do C, Nascimento GFL. Reliability analysis of the Manchester Triage System: inter-observer and intra-observer agreement. Rev Lat Am Enfermagem. Jul 16, 2018;26:e3005. [doi: 10.1590/1518-8345.2205.3005] [Medline: 30020337]
23. Jung S, Yi Y. Incidence of overtriage and undertriage and associated factors: a cross-sectional study using a secondary data analysis. J Adv Nurs. Apr 2024;80(4):1405-1416. [doi: 10.1111/jan.15895] [Medline: 37828736]
24. Loftus TJ, Ruppert MM, Shickel B, et al. Overtriage, undertriage, and value of care after major surgery: an automated, explainable deep learning-enabled classification system. J Am Coll Surg. Feb 1, 2023;236(2):279-291. [doi: 10.1097/XCS.0000000000000471] [Medline: 36648256]
25. Asplin BR. Undertriage, overtriage, or no triage? In search of the unnecessary emergency department visit. Ann Emerg Med. Sep 2001;38(3):282-285. [doi: 10.1067/mem.2001.117842] [Medline: 11524648]
26. Cao B, Huang S, Tang W. AI triage or manual triage? Exploring medical staffs' preference for AI triage in China. Patient Educ Couns. Feb 2024;119:108076. [doi: 10.1016/j.pec.2023.108076] [Medline: 38029576]

27. Amiot F, Potier B. Artificial intelligence (AI) and emergency medicine: balancing opportunities and challenges. JMIR Med Inform. Aug 13, 2025;13:e70903. [doi: 10.2196/70903] [Medline: 40802997]
28. Townsend BA, Plant KL, Hodge VJ, Ashaolu O, Calinescu R. Medical practitioner perspectives on AI in emergency triage. Front Digit Health. 2023;5:1297073. [doi: 10.3389/fdgth.2023.1297073] [Medline: 38125759]

Abbreviations

DBP: diastolic blood pressure
DSGVO: Datenschutz-Grundverordnung (General Data Protection Regulation)
ED: emergency department
GCS: Glasgow Coma Scale
GT1: ground truth 1
GT2: ground truth 2
LLM: large language model
LOO: leave-one-out variant
MTS: Manchester Triage System
NPV: negative predictive value
PPV: positive predictive value
qSOFA: Quick Sequential Organ Failure Assessment
SBP: systolic blood pressure
SpO_2: peripheral capillary oxygen saturation

Edited by Andrew Coristine; peer-reviewed by Krithik Vishwanath, Lex Shao, Ling Lin; submitted 14.Jan.2026; initial comments to author 24.Apr.2026; final revised version received 17.Aug.2026; accepted 21.Aug.2026; published 24.Nov.2026

Please cite as:
Abedin N, Freiherr von Beust M, Kutscher N, Meier S, Schiller A, Hofbauer L, Erasmus HP, Haberkorn SM, Kubesch-Grün A, Graf J, Marzi I, Zeuzem S, Dultz G
Evaluating Human and AI-Based Triage Performance in the Emergency Department: Real-World Cohort Study
J Med Internet Res 2026;28:e91442
URL: https://www.jmir.org/2026/1/e91442
doi: 10.2196/91442

© Nada Abedin, Maximilian Freiherr von Beust, Nina Kutscher, Simon Meier, André Schiller, Ludwig Hofbauer, Hans-Peter Erasmus, Sebastian M Haberkorn, Alica Kubesch-Grün, Jürgen Graf, Ingo Marzi, Stefan Zeuzem, Georg Dultz. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.