

Diagnostic Accuracy of Four Large Language Models for the Cipto Triage Method: A Comparative Evaluation for Emergency Department Triage in Resource-Limited Settings

Ahmad Ridwan Fauzi¹, Adelia Chitra Sazkia², and Muhammad Ardiansyah³

Department of Information Systems, Telkom University, Jakarta, Indonesia

Corresponding author: Ahmad Ridwan Fauzi (e-mail: ahmadridwanfauzi@telkomuniversity.ac.id), **Author(s)**
Email: Adelia Chitra Sazkia (e-mail: adeliachitrasazkia@telkomuniversity.ac.id, Muhammad Ardiansyah (e-mail: muhhammadardiansyahma@telkomuniversity.ac.id)

Abstract Triage misclassification, and undertriage in particular, remains a persistent patient-safety problem in resource-limited emergency departments (EDs), yet the performance of large language models (LLMs) on simplified three-tier systems such as the Cipto Triage Method (CTM) outside English-language settings is largely unexamined. This study evaluated four low-cost LLMs (DeepSeek V4 Pro, DeepSeek V4 Flash, Tencent Hy3, and MiMo V2.5) for CTM classification within a framework coupling diagnostic accuracy with undertriage safety and cost analysis for resource-limited deployment. The models were tested on a stratified random sample of 1,200 clinical records from the MIETIC corpus (v1.0.0; 5,528 unique cases after deduplication), relabeled from the Emergency Severity Index to CTM by four board-certified emergency physicians (400 cases per category), with physician consensus as the reference standard. Outcomes were overall accuracy and Cohen kappa, per-class sensitivity, specificity, and predictive values with 95% Wilson-score confidence intervals, undertriage detection, and API cost. Accuracy ranged from 66.5% to 68.1% (kappa 0.50–0.52), with DeepSeek V4 Flash highest (68.1%; 95% CI 65.4–70.7%) at the lowest cost (\$0.21 total); no pairwise difference in overall accuracy was statistically significant after adjustment. Resuscitation sensitivity reached only 58.5–64.8%, and undertriage detection varied from 11.9% to 56.6%, with model-generated risk flags uncalibrated across models. Resuscitation sensitivity fell to 39.7–43.4% among patients without red-flag vital signs, and replicate runs changed 7–30% of classifications, indicating material run-to-run noise. Although accuracy exceeded the 47.5% reported for ChatGPT on the Canadian Triage and Acuity Scale, that comparison is indicative only, as benchmarks differ in triage system, case-mix, and protocol. These findings do not support autonomous LLM triage; they motivate prospective, multicenter validation of LLMs strictly as secondary, clinician-overridden decision support for CTM triage in Indonesian EDs, where the most realistic near-term role is low-cost undertriage alerting.

Keywords large language models; emergency department triage; clinical decision support; Cipto Triage Method; diagnostic accuracy

1. Introduction

Triage matches patients to the urgency of clinical response they require, yet despite structured protocols, misclassification remains common [1], [2]. Undertriage, in which a critically ill patient is assigned to a lower-acuity category, carries tangible risks: a relative risk of 3.1 (95% CI, 2.11–4.54) for poor outcomes in the CTM context [3] and an odds ratio of 1.24 (95% CI, 1.10–1.40) in major trauma [4]. Low- and middle-income countries face amplified challenges: high patient volumes, staffing shortages, and constrained resources [5]. AI adoption remains uneven [6], and in Indonesia, most ED patients arrive without pre-hospital transport, leaving clinicians to assess acuity unaided

[7]. The Cipto Triage Method (CTM), a validated three-tier instrument widely used in Indonesian hospitals [3], [5], [8], remains vulnerable to human error and inconsistency under high throughput. To guide the design of a clinical decision support system for ED triage, we conducted semi-formal interviews with two board-certified emergency medicine specialists at a national referral hospital in Indonesia to elicit the triage workflow and requirements for a triage recommendation system. At the front desk, nurses gather the chief complaint, vital signs, and a brief examination, then decide three things: the best-fitting CTM category, the clinical and resource risks assumed, and the most appropriate specialist or pathway. The

specialists stressed that these judgments unfold within minutes under heavy cognitive load, where errors can prove life-threatening, and that the main value of a recommendation system would be a secondary opinion at the point of doubt, with the nurse retaining the final decision. The interviews involved no patient data and are reported as an informal needs assessment.

A web-based clinical decision support system was designed for the ED front desk (Fig. 1). It incorporates automatic speech recognition (ASR) for data entry and a hybrid LLM core built for fluctuating connectivity: data flow to a cloud-based LLM when connectivity is available and to a lightweight local LLM when it drops, delivering offline classification at reduced but still clinically useful accuracy. For nurse verification and override, it displays a CTM category, a brief rationale, a hospital acceptance risk assessment, and a specialist pathway recommendation. Large language models (LLMs) have gained traction for clinical reasoning and diagnostic support [9], [10], [11], [12], [13], [14]; reviews of intelligent triage decision support identify machine learning and natural language processing as dominant approaches, noting that most such systems rely on

[18] probed ChatGPT against the Korean Triage and Acuity Scale and the Emergency Severity Index; Arslan et al. [19] benchmarked AI platforms against triage nurses; and Franc et al. [20] showed ChatGPT reaching only 47.5% on the Canadian Triage and Acuity Scale.

The gap is specific: prior evaluations cover five-level frameworks in English-language settings, whereas simplified three-tier systems such as CTM, standard across many resource-limited settings, remain untested with current-generation models, and undertriage safety and deployment cost have not been quantified in that context. This study therefore evaluated the diagnostic accuracy, undertriage safety, and API cost of four low-cost models for CTM classification to support the cloud-based component of the proposed hybrid system. The scope is limited to evaluating only the cloud-based LLM of the architecture and not the other pathways like ASR and local-based LLM. The novelty lies particularly in the CTM-specific evaluation framework. Three contributions follow: (i) the first diagnostic-accuracy evaluation of LLMs on a

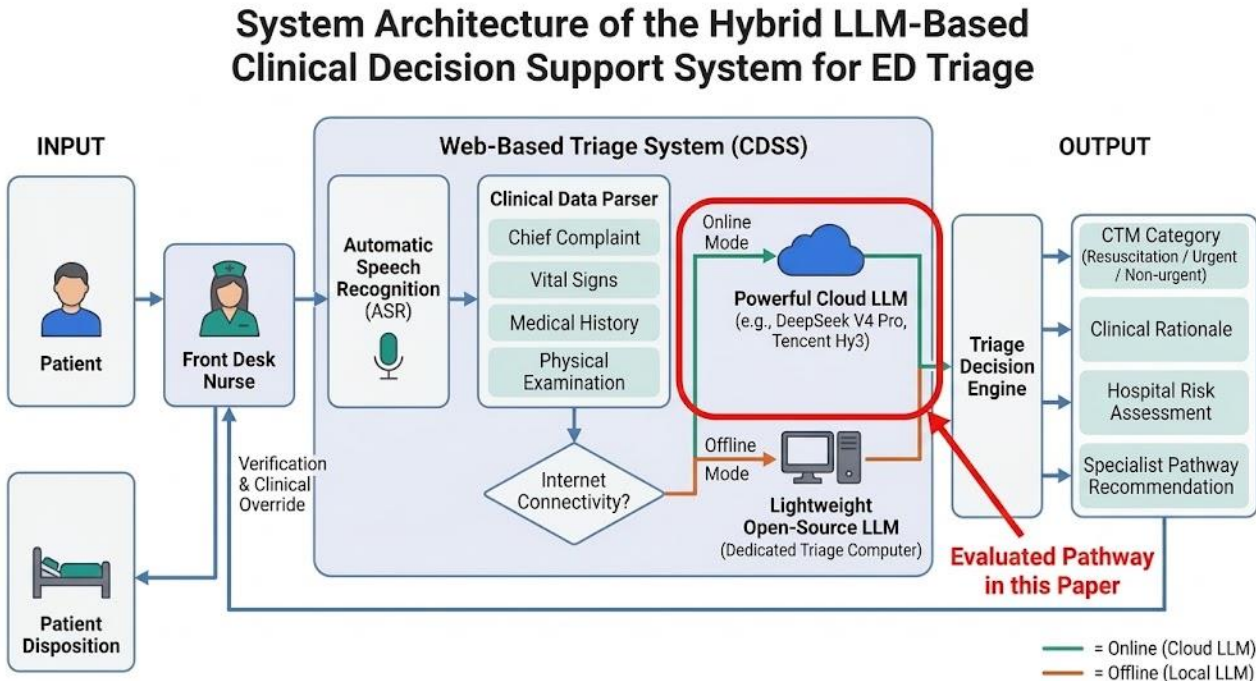


Fig. 1. Proposed architecture of the hybrid large language model-based clinical decision support system for emergency department triage, illustrating patient data acquisition, automatic speech recognition, cloud-based and local LLM pathways, CTM classification, and clinician verification. The cloud-based LLM pathway is the only component evaluated in this study.

structured rather than unstructured clinical text [15]. In ED triage, Williams et al. [16] found 89% pairwise accuracy for GPT-4 in separating higher- from lower-acuity presentations; Kim et al. [17] and Colakca et al.

simplified three-tier CTM; (ii) an explicit safety analysis quantifying undertriage, the most clinically dangerous misclassification form, via agreement between model-generated risk flags and reference-confirmed

undertriage; and (iii) a cost and deployment analysis for resource-limited settings, where per-query cost governs adoption as much as diagnostic performance.

Four research questions guided the work: (1) classification accuracy, (2) detection of undertriage, (3) inter-model performance differences, and (4) API cost-efficiency for resource-constrained hospitals. The four models occupy the lowest-cost tier able to process Indonesian clinical text; recent open-weight models approach frontier performance at a fraction of the per-query cost, and for Indonesian hospitals, where annual per-capita health expenditure stays below \$200, testing these options equips administrators to weigh clinical utility against economic sustainability.

including the critical and borderline presentations CTM is designed to classify, and confirmed them as realistic material for triage labeling. Each row was then labeled directly in CTM form, blind: physicians saw only the translated clinical text alongside the standardized CTM criteria, and the original ESI tags were never displayed. The CTM criteria applied were the validated definitions and decision rules of Habib et al. [3]: the three-tier definitions (Resuscitation: immediate intervention; Urgent: treatment within 30 minutes; Non-urgent: within 60 minutes), a syndromic assessment approach (chief complaint; comorbidity or injury mechanism; mental status, peripheral pulse, respiratory rate, capillary refill, and temperature), and two decision rules: when in

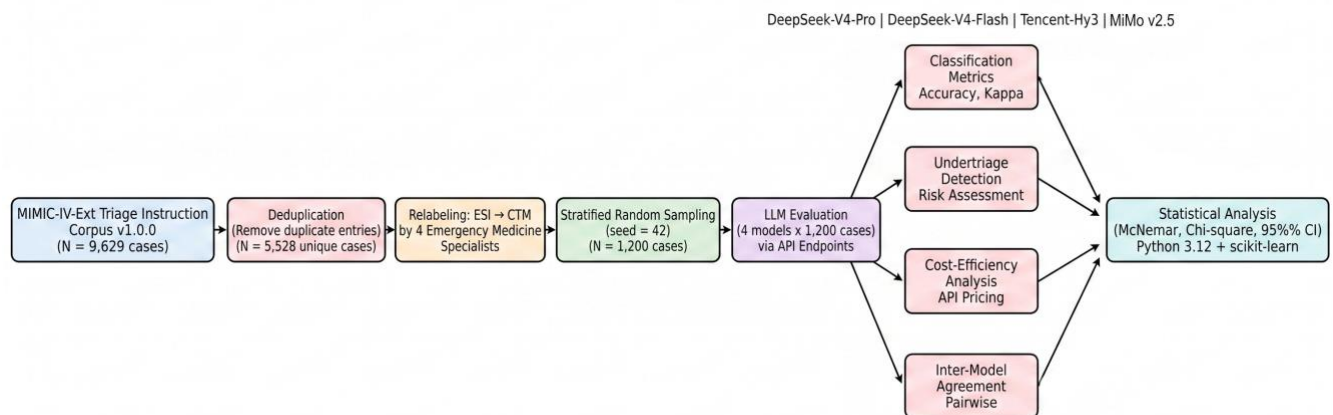


Fig. 2. Flowchart of the diagnostic accuracy evaluation, illustrating MIETIC dataset preprocessing, removal of duplicate records, English-to-Indonesian translation, physician consensus labeling, stratified sampling of 1,200 cases, evaluation of four LLMs, and statistical performance analysis.

II. Method

A. Study Design and Time Period

The study adopted a diagnostic accuracy design, adhering to the STARD 2015 reporting checklist [21] and the TRIPOD+AI statement [22]. Each model's triage output was assessed retrospectively against a specialist-derived reference standard from an accessible clinical dataset, as explained in Fig. 2.

B. Study Setting

Data were drawn from the MIETIC Triage Instruction Corpus v1.0.0, a publicly available derivative of MIMIC-IV holding de-identified records from the ED of a US tertiary care hospital; after removing duplicates from the original 9,629 cases, 5,528 unique records remained. To adapt the corpus for an Indonesian triage system, all 5,528 records were translated from English to Indonesian using the MiMo V2.5 model, preserving the standardized presentation format (demographics, vital signs, chief complaint, and past medical history); the four board-certified emergency physicians reviewed the translations across the full acuity range,

doubt assign the higher category, and regard undertriage as more dangerous than overtriage. Per-physician workloads ranged from 1,125 to 1,420 records. At checkpoint meetings every 600 labels per physician, the panel verified conformity with the standard criteria, and disputed cases were finalized by consensus before labeling resumed. The consensus-labeled corpus constituted the reference standard.

C. Population and Sample Size

Before sampling, the relabeled corpus contained 2,006 Resuscitation (36.3%), 536 Urgent (9.7%), and 2,986 Non-urgent (54.0%) cases. Proportional allocation (Eq. (1))[23] would have allocated 435, 116, and 649 cases; instead, a disproportionate allocation of 400 cases per CTM category was drawn (19.9%, 74.6%, and 13.4% of the three strata) with a fixed seed, each stratum being large enough to realize the allocation. Equal allocation does not mirror real-world prevalence but guarantees sufficient power for class-level estimates, particularly for Resuscitation. Because the design is unbalanced relative to practice, predictive values and the distribution of errors are governed partly by the

analysis case-mix: with one third of cases per stratum, the rarer Resuscitation and Urgent strata are over-represented, magnifying class-specific errors in aggregate error counts. A post-hoc analysis showed 80% power to detect a 5.4-percentage-point sensitivity gap between models (odds ratio ≈ 1.22) at $\alpha = 0.05$.

$$n_h = \frac{N_h}{N} \cdot n \quad (1)$$

where n_h is the number of cases allocated to stratum h , N_h the stratum size, N the total population (5,528), and n the total sample size (1,200); the present study instead allocated 400 cases per CTM category.

D. Intervention

All four models were queried through the OpenRouter chat-completions API on 14 July 2026, using fixed versioned identifiers: deepseek/deepseek-v4-pro, deepseek/deepseek-v4-flash, tencent/hy3, and xiaomi/mimo-v2.5 [24]–[28]. No parameters or weights were updated during data collection, and no “latest”-style aliases were used; the selection spanned architectures, parameter scales, and cost tiers. All queries used a temperature of 0.1 and no other decoding constraints; each case was queried once per model in the primary evaluation, and classification stability was assessed with replicate runs on a subsample (Data Analysis). Each model received the same Indonesian-language record (truncated to 2,500 characters) containing chief complaint, vital signs, past medical history, and physical examination findings. Models returned a structured JSON response with the CTM category, a brief clinical rationale, and a binary undertriage-risk flag, enabling automated parsing and cross-model comparison.

E. Outcome Measures

Primary outcomes were overall classification accuracy and Cohen kappa against the reference standard. Secondary outcomes included per-class sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) with 95% Wilson-score confidence intervals; undertriage detection and overtriage rates; undertriage-alert reliability; subgroup performance; pairwise inter-model comparisons; and total API cost for 1,200 queries. The undertriage-risk flag is a model-generated self-report: each model stated, in the same JSON response, whether the case risked being assigned a lower category than warranted. The flag derives from the case presentation and the CTM criteria alone, and the model never has access to the reference standard, so it cannot know whether its own assignment was correct.

F. Data Analysis

All analyses were run in Python 3.12 (scikit-learn, pandas, scipy, statsmodels). Confusion matrices per model yielded per-class sensitivity, specificity, PPV, and NPV (Eqs. (2)–(6)) with 95% Wilson-score confidence intervals (Eq. (12)); Cohen kappa (Eq. (7))

measured agreement with the reference standard. Reference-label reliability was gauged with Fleiss kappa (Eq. (11)) on a 553-case subset (10% of the corpus) independently re-labeled in an automated second-opinion pass by a separate language model (DeepSeek V4.1 Flash), blind to the reference labels (Supplementary Material S1). Pairwise comparisons used McNemar's test with continuity correction (Eq. (13)); the Bonferroni-adjusted threshold (Eq. (14)) for six comparisons was $\alpha = 0.0083$. Paired-difference 95% confidence intervals, Cohen's h effect sizes, and adjusted p values for the pairwise comparisons are reported in Table 2; subgroup performance across age, sex, chief complaint, and red-flag vital signs (systolic blood pressure <90 mmHg, respiratory rate >30 /min, SpO2 $<88\%$, or heart rate >120 /min) in Table 6; undertriage-alert reliability (recall, precision, specificity, F1, and false-alert rate) in Table 5; and repeatability across replicate runs in Section III-D. Prevalence-adjusted predictive values were derived from per-class sensitivity and specificity using Bayes' theorem under real-world CTM prevalence (Eqs. (19) and (20); Table 7). All 4,800 queries (1,200 per model) were completed with no missing or unparseable responses; response latency was not measured because only the classification pathway was exercised. From each confusion matrix, overall accuracy and the per-class diagnostic indices were computed with Eqs. (2)–(6)[29]:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Se = \frac{TP}{TP+FN} \quad (3)$$

$$Sp = \frac{TN}{TN+FP} \quad (4)$$

$$PPV = \frac{TP}{TP+FP} \quad (5)$$

$$NPV = \frac{TN}{TN+FN} \quad (6)$$

where TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives per CTM category against the reference standard. Agreement with the reference standard was quantified with Cohen kappa (Eq. (7)) [30], computed from the observed and chance-expected agreement proportions in Eqs. (8) and (9):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (7)$$

$$p_o = \frac{1}{n} \sum_{i=1}^k n_{ii} \quad (8)$$

$$p_e = \sum_{i=1}^k p_{i.} \cdot p_{.i} \quad (9)$$

where n_{ii} is the diagonal (concordant) count for category i , $p_{i.}$ and $p_{.i}$ the row and column marginal proportions, and k the number of categories ($k = 3$). The asymptotic standard error (Eq. (10))[30] supported

Algorithm 1. End-to-end evaluation procedure for large language model-based Cipto Triage Method classification, physician-consensus labeling, stratified sampling, standardized API inference, performance metric computation, undertriage detection, statistical comparisons, prevalence adjustment, and API cost-efficiency analysis.

- (1) **Input:** CTM-deduplicated and translated corpus (5,528 cases) x ; models = DeepSeek V4 Pro, DeepSeek V4 Flash, Tencent Hy3, MiMo V2.5 l ; evaluation sample = 1,200 cases (400 per CTM category).
- (2) **Output:** evaluation sample = 1,200 cases (400 per CTM category) y ; per-model accuracy, kappa, class-level metrics, undertriage detection rate, and API cost.
- (3) Initialization
- (4) **FOR EACH** row of a medical case $x \in \{1, 2, \dots, 5528\}$ **AND** each 4 LLMs $l \in \{1, 2, 3, 4\}$
- (5) Pass the case to OpenRouter OpenAI API to determine triage from the input
- (6) **END FOR**
- (7) Randomly pick 1200 cases from dataset x with seed $s = 42$
- (8) **FOR EACH** row of an evaluated case $y \in \{1, 2, \dots, 1200\}$
- (9) Four physicians determine triage label
- (10) **END FOR**
- (11) Compute overall accuracy, Cohen kappa, and per-class sensitivity, specificity, PPV, and NPV with Wilson intervals (Eqs. (2)–(12)) from evaluated dataset y .
- (12) Compute undertriage detection and overtriage rates (Eqs. (15) and (16)); run pairwise McNemar tests with Bonferroni adjustment (Eqs. (13) and (14)) from evaluated dataset y .
- (13) Compute prevalence-adjusted predictive values (Eqs. (19) and (20)) and API cost with the cost-efficiency ratio (Eqs. (17) and (18)) from evaluated dataset y .

the reported intervals, and reference-label reliability used Fleiss kappa (Eq. (11)) [31]:

$$SE(\kappa) = \sqrt{\frac{p_o(1-p_o)}{n(1-p_e)^2}} \quad (10)$$

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (11)$$

$\bar{P}$ is the mean agreement across rater–subject pairs and $\bar{P}_e$ the mean chance agreement. All proportions were reported with 95% Wilson confidence intervals (Eq. (12)) [32]. Pairwise comparisons used McNemar's test with continuity correction (Eq. (13)) [33], with the Bonferroni-adjusted threshold (Eq. (14)) [34] applied across the six comparisons:

$$CI_{95\%} = \frac{\hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \quad (12)$$

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \quad (13)$$

$$\alpha_{adj} = \frac{\alpha}{m} \quad (14)$$

$\hat{p}$ is the observed proportion, n the number of observations, $z = 1.96$ the standard normal quantile, b and c the discordant counts, and $m = 6$ the number of pairwise comparisons. Undertriage detection and overtriage rates (Eqs. (15) and (16)) [35]:

$$UDR = \frac{n_{UT, detected}}{n_{UT}} \quad (15)$$

$$OTR = \frac{n_{OT}}{n_{total}} \quad (16)$$

$n_{UT, detected}$ is the number of undertriaged cases flagged by the model, n_{UT} the total number of undertriaged cases, n_{OT} the number of overtriaged cases, and $n_{total} = 1,200$, the total number of cases. The flag is a self-report of perceived risk rather than an internal error-detection mechanism, and its calibration differs across models (13.6–47.8% of cases flagged). Algorithm 1 summarizes the complete evaluation procedure, from data preprocessing to statistical analysis; the full prompt template, output constraints, language handling, and failure-handling procedure are provided in Supplementary Material S1.

III. Result

A. Accuracy

Agreement between the reference labels and the independent automated re-labeling on the 553-case subset was fair (Fleiss kappa = 0.33; 95% CI: 0.25–0.40), indicating that CTM boundary decisions retain a component of label uncertainty. A sensitivity analysis probed its influence on accuracy. Of the 1,200 evaluated cases, 132 fell within the reliability subset: 83 with concordant labels, 49 discordant. Accuracy was high on concordant cases (75.9–85.5%) but fell sharply on discordant cases (2.0–14.3%), where models instead

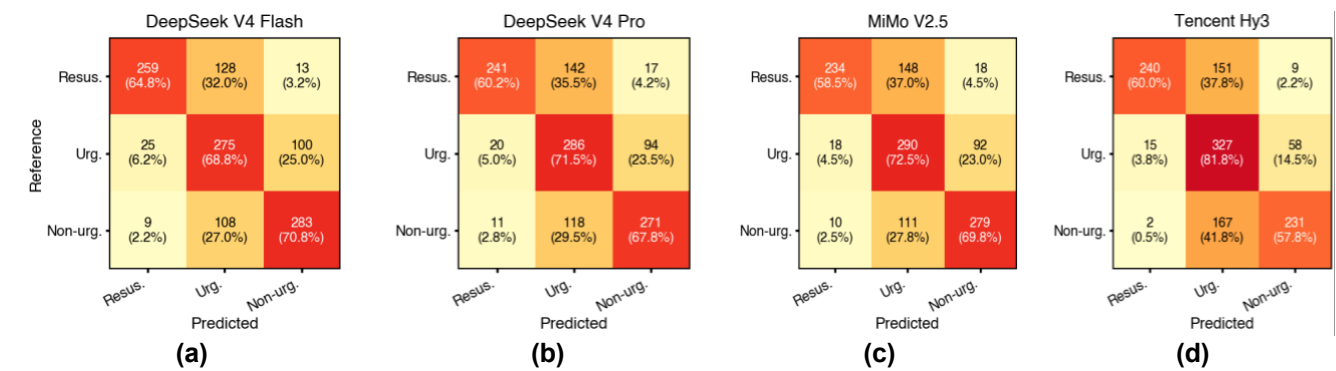


Fig. 3. Row-normalized confusion matrices comparing (a) DeepSeek V4 Flash, (b) DeepSeek V4 Pro, (c) MiMo V2.5, and (d) Tencent Hy3 for three-class Cipto Triage Method classification on 1,200 clinical cases. Rows represent physician-consensus reference categories, columns represent model predictions, and cell values indicate classification percentages.

Table 1. Overall classification accuracy, 95% confidence intervals, Cohen's kappa agreement coefficients, and numbers of correctly classified cases for four large language models evaluated on 1,200 CTM-labeled emergency department records.

Model	Accuracy	Cohen kappa	Correct Cases
DeepSeek V4 Flash	68.1% (65.4–70.7)	0.521	817
MiMo V2.5	66.9% (64.2–69.5)	0.504	803
DeepSeek V4 Pro	66.5% (63.8–69.1)	0.498	798
Tencent Hy3	66.5% (63.8–69.1)	0.498	798

aligned with the independent re-labeling in 71.4–91.8% of cases. Apparent errors therefore concentrated where the two labeling processes disagreed, so measured accuracy is bounded by reference-label uncertainty at CTM boundaries; with only 49 discordant cases, these estimates are exploratory. Table 1 summarizes overall accuracy: all four models clustered within a narrow band of 66.5% to 68.1%, with DeepSeek V4 Flash on top (68.1%), MiMo V2.5 next (66.9%), and DeepSeek V4 Pro and Tencent Hy3 tied at 66.5%; Cohen kappa values fell between 0.50 and 0.52. No pairwise difference in overall accuracy reached significance after adjustment (all $p_{adj} = 1.0$; Table 2); because failure to reject the null hypothesis does not demonstrate equivalence, no equivalence claim is made. Fig. 3 shows the row-normalized confusion matrices. Misclassification was local rather than arbitrary: 93–97% of all errors were adjacent-category shifts, and this pattern held across models. True Resuscitation cases were correctly classified in 58.5–64.8% of cases, with residual errors overwhelmingly assigned to Urgent (32.0–37.8%) rather than directly to Non-urgent (2.2–4.5%). True Urgent cases were classified correctly in 68.8–81.8% of cases (lowest for DeepSeek V4 Flash), with the remainder split mainly toward Non-urgent (14.5–25.0%). The model correctly placed true non-urgent cases in 57.8–70.8% of

cases, and its errors concentrated on urgent cases (27.0–41.8%). Table 2 reports the pairwise comparisons. No difference in overall accuracy survived Bonferroni adjustment (all $p_{adj} = 1.0$; the largest raw difference was 1.6 percentage points, 95% CI –0.7 to 3.9). The safety-relevant metrics were less tightly clustered: DeepSeek V4 Flash exceeded MiMo V2.5 on Resuscitation sensitivity by 6.3 percentage points (95% CI 2.6–9.9; $p_{adj} = 0.009$ and Tencent Hy3 recorded the lowest undertriage rate, 2.9 and 3.3 points below DeepSeek V4 Pro and MiMo V2.5 (both $p_{adj} < 0.002$). Effect sizes remained small throughout ($|h| \leq 0.13$ so where differences were detectable, they were modest.

B. Performance

Undertriage frequency ranged from 218 to 258 cases per model (18.2–21.5% of all cases; Table 3). MiMo V2.5 caught the most (56.6%), including 81.1% of the most dangerous pathway (Resuscitation misclassified as Urgent), whereas DeepSeek V4 Pro flagged only 11.9% (30 of 253); overtriage ranged from 11.6% (MiMo V2.5) to 15.3% (Tencent Hy3).

$R \rightarrow U$ = Resuscitation $\rightarrow$ Urgent; $R \rightarrow N$ = Resuscitation $\rightarrow$ Non-urgent; $U \rightarrow N$ = Urgent $\rightarrow$ Non-urgent. The flag was generated from the case presentation alone, so Table 3 quantifies how often it

Table 2. Pairwise statistical comparisons of four large language models based on overall classification accuracy, Resuscitation sensitivity, and undertriage rate, including percentage-point differences, 95% confidence intervals, and adjusted significance values.

Pair	Δ Accuracy, pp (95% CI)	p_{adj}	Δ Resus. sens., pp (95% CI)	p_{adj}	Δ Undertriage, pp (95% CI)	p_{adj}	$\max h $
Flash – Pro	1.6 (–0.7 to 3.9)	1.00	4.5 (1.1 to 7.9)	0.085	–1.0 (–2.8 to 0.8)	1.00	0.09
Flash – MiMo	1.2 (–1.1 to 3.5)	1.00	6.3 (2.6 to 9.9)	0.009	–1.4 (–3.2 to 0.4)	0.855	0.13
Flash – Hy3	1.6 (–0.8 to 3.9)	1.00	4.8 (1.4 to 8.2)	0.061	1.9 (0.2 to 3.7)	0.241	0.10
Pro – MiMo	–0.4 (–2.8 to 2.0)	1.00	1.8 (–2.2 to 5.7)	1.00	–0.4 (–2.3 to 1.5)	1.00	0.04
Pro – Hy3	0.0 (–2.1 to 2.1)	1.00	0.3 (–2.7 to 3.2)	1.00	2.9 (1.4 to 4.4)	0.001	0.07

Table 3. Undertriage detection performance and reliability of model-generated undertriage-risk alerts across four large language models, including (a) classification error pathways, true positives, false positives, false negatives, true negatives, (b) recall, precision, specificity, F1-score, and false-alert rate.

(a)									
Model	Undertriaged		Flagged		Detected	Rate	R→U	R→N	U→N
MiMo V2.5	258		573		146	56.6%	120/148	5/18	21/92
Tencent Hy3	218		275		74	33.9%	71/151	0/9	3/58
DeepSeek V4 Flash	241		260		45	18.7%	42/128	1/13	2/100
DeepSeek V4 Pro	253		163		30	11.9%	28/142	1/17	1/94
(b)									
Model	TP	FP	FN	TN	Recall, % (95% CI)	Precision, % (95% CI)	Specificity, % (95% CI)	F1	False-ale % (95% CI)
MiMo V2.5	146	427	112	515	56.6 (50.5–62.5)	25.5 (22.1–29.2)	54.7 (51.6–57.8)	0.351	45.3 (42.2–48.5)
Tencent Hy3	74	201	144	781	33.9 (28.0–40.5)	26.9 (22.0–32.4)	79.5 (76.5–82.3)	0.300	20.5 (18.1–23.1)
DeepSeek V4 Flash	45	215	196	744	18.7 (14.3–24.1)	17.3 (13.2–22.4)	77.6 (74.5–80.5)	0.180	22.4 (19.9–25.2)
DeepSeek V4 Pro	30	133	223	814	11.9 (8.4–16.4)	18.4 (13.2–25.1)	86.0 (83.3–88.3)	0.144	14.0 (12.0–16.4)

coincided with reference-confirmed undertriage. Flagging rates differed markedly across models (13.6% for DeepSeek V4 Pro, 21.7% for DeepSeek V4 Flash, 22.9% for Tencent Hy3, and 47.8% for MiMo V2.5), so the flag is model-specific and uncalibrated: a higher flagging propensity raises the detection rate without establishing better clinical judgment. Treating each model's flag as a classifier of reference-confirmed undertriage sharpens this picture (Table 3). Recall

ranged from 11.9% to 56.6% and precision from 17.3% to 26.9%, so most alerts did not correspond to confirmed undertriage: at the highest flagging rate (MiMo V2.5, 573 of 1,200 cases flagged), 74.5% of alerts were false, and the false-alert rate against non-undertriaged cases was 45.3%. Specificity ranged from 54.7% (MiMo V2.5) to 86.0% (DeepSeek V4 Pro), and F1 scores from 0.14 to 0.35. The flag is therefore best understood as a broad safety-net prompt rather than a calibrated detector.

Table 4. Comparative API cost and classification efficiency of four large language models for 1,200 emergency department triage cases, reporting overall accuracy, total API expenditure, cost per case, and accuracy-to-cost ratio in US dollars.

Model	Accuracy	Cost (\$)	Cost/Case	Accuracy/Cost
Tencent Hy3	66.5%	3.95	0.00329	0.168
DeepSeek V4 Flash	68.1%	0.21	0.00017	3.289
MiMo V2.5	66.9%	0.66	0.00055	1.015
DeepSeek V4 Pro	66.5%	3.00	0.00250	0.222

Table 5. Class-specific diagnostic performance of four large language models across the Resuscitation, Urgent, and Non-urgent CTM categories, presenting sensitivity, specificity, positive predictive value, and negative predictive value with 95% Wilson-score confidence intervals.

Model	Category	Sensitivity	Specificity	PPV	NPV
DeepSeek V4 Flash	Resuscitation	64.8 (59.9–69.3)	95.8 (94.1–96.9)	88.4 (84.2–91.6)	84.5 (82.0–86.7)
	Urgent	68.8 (64.0–73.1)	70.5 (67.2–73.6)	53.8 (49.5–58.1)	81.9 (78.8–84.6)
	Non-urgent	70.8 (66.1–75.0)	85.9 (83.3–88.1)	71.5 (66.8–75.7)	85.5 (82.8–87.7)
DeepSeek V4 Pro	Resuscitation	60.2 (55.4–64.9)	96.1 (94.6–97.3)	88.6 (84.3–91.9)	82.9 (80.3–85.2)
	Urgent	71.5 (66.9–75.7)	67.5 (64.2–70.7)	52.4 (48.2–56.5)	82.6 (79.5–85.3)
	Non-urgent	67.8 (63.0–72.1)	86.1 (83.6–88.3)	70.9 (66.2–75.3)	84.2 (81.6–86.6)
MiMo V2.5	Resuscitation	58.5 (53.6–63.2)	96.5 (95.0–97.6)	89.3 (85.0–92.5)	82.3 (79.7–84.6)
	Urgent	72.5 (67.9–76.6)	67.6 (64.3–70.8)	52.8 (48.6–57.0)	83.1 (80.0–85.8)
	Non-urgent	69.8 (65.1–74.0)	86.3 (83.7–88.5)	71.7 (67.1–76.0)	85.1 (82.5–87.4)
Tencent Hy3	Resuscitation	60.0 (55.1–64.7)	97.9 (96.6–98.7)	93.4 (89.7–95.8)	83.0 (80.5–85.3)
	Urgent	81.8 (77.7–85.2)	60.3 (56.8–63.6)	50.7 (46.8–54.5)	86.9 (83.8–89.4)
	Non-urgent	57.8 (52.9–62.5)	91.6 (89.5–93.4)	77.5 (72.4–81.9)	81.3 (78.6–83.7)

API cost per model was computed with Eq. (17), and the accuracy-to-cost ratio reported in Table 4 with Eq. (18)[25]–[28]:

$$C = \frac{N_{in} c_{in} + N_{out} c_{out}}{1000} \quad (17)$$

$$R = \frac{Acc}{C} \quad (18)$$

N_{in} and N_{out} are input and output token counts across all 1,200 queries, c_{in} and c_{out} the per-1,000-token prices in US dollars, and Acc the overall accuracy (%). Prevalence-adjusted predictive values were computed from per-class sensitivity and specificity via Bayes' theorem [29]:

$$PPV_{adj,c} = \frac{Se_c \pi_c}{Se_c \pi_c + (1 - Sp_c)(1 - \pi_c)} \quad (19)$$

$$NPV_{adj,c} = \frac{Sp_c(1 - \pi_c)}{Sp_c(1 - \pi_c) + (1 - Se_c)\pi_c} \quad (20)$$

where π_c denotes the real-world prevalence of CTM category c , Se_c and Sp_c the per-class sensitivity and

specificity from Table 5, and $PPV_{adj,c}$ and $NPV_{adj,c}$ the predictive values expected under that prevalence. Total API spending spanned from \$0.21 (DeepSeek V4 Flash) to \$3.95 (Tencent Hy3; Table 4). DeepSeek V4 Flash had the best accuracy-to-cost ratio, pairing the highest accuracy with the lowest total cost.

Because the balanced 1:1:1 test design does not reflect the natural case-mix of emergency departments, the predictive values in Table 5 are sample-dependent and should not be read as deployment estimates. Table 7 reports prevalence-adjusted positive and negative predictive values computed from the per-class sensitivity and specificity of Table 3 using Bayes' theorem under the real-world CTM prevalence reported for an Indonesian referral ED (Resuscitation 14.9%, Urgent 63.8%, Non-urgent 21.3%) [3]. Under this prevalence, Resuscitation PPV falls from the observed 88–93% to approximately 73–83%, Urgent PPV rises to 78–80%, Non-urgent PPV falls to 57–65%, and Urgent NPV falls

Table 6. Subgroup analysis of triage classification accuracy, Resuscitation sensitivity, and undertriage rates across patient age, sex, vital-sign abnormalities, and chief complaint categories for four large language models. Subgroup sample sizes and performance ranges are reported.

Subgroup	N (nR)	Flash	Pro	MiMo	Hy3	SeR, % (range)	UT, % (range)
Age <65 y	827 (222)	68.0	66.6	66.6	65.3	56.3–60.8	18.0–20.7
Age ≥65 y	290 (140)	74.8	70.3	72.8	74.8	65.7–77.1	15.5–21.7
Female	531 (147)	70.8	67.4	67.0	67.0	53.7–64.6	16.0–21.5
Male	465 (156)	70.1	68.8	70.3	69.7	63.5–67.9	18.3–21.7
Red-flag vitals present	109 (87)	78.9	76.1	73.4	81.7	77.0–86.2	11.0–21.1
Red-flag vitals absent	923 (189)	65.1	64.1	65.9	63.7	39.7–43.4	17.9–21.8
Trauma/injury	103 (41)	68.9	76.7	68.9	72.8	63.4–73.2	13.6–20.4
Respiratory	186 (75)	66.1	62.4	66.1	66.1	54.7–66.7	22.6–25.8
Chest/cardiac	71 (52)	74.6	71.8	67.6	71.8	78.8–82.7	14.1–19.7
Abdominal/GI	178 (33)	60.1	60.7	64.6	52.8	27.3–39.4	14.0–19.1
Fever/infection	85 (20)	71.8	62.4	71.8	74.1	55.0–60.0	20.0–34.1
Neurological	89 (42)	65.2	62.9	56.2	65.2	59.5–73.8	23.6–34.8
Other	405 (94)	71.1	68.9	69.1	68.6	44.7–50.0	18.5–21.2

from 82–87% to 56–65%. Under the study corpus’s own class distribution (36.3%, 9.7%, 54.0%), Urgent PPV would fall to approximately 18–20%. These shifts confirm that the balanced design inflates apparent performance for rare classes; Table 5 should be read as class-level discrimination rather than deployment predictive values, and even the model ranking depends on the target case-mix.

C. Subgroup Performance

Performance tracked clinical presentation (Table 6). Accuracy was higher for patients aged ≥65 years (70.3–74.8%) than for younger patients (65.3–68.0%) and similar between sexes. The sharpest gradient followed vital signs: with red-flag values present, accuracy was 73.4–81.7% and Resuscitation sensitivity 77.0–86.2%; without them, accuracy fell to 63.7–65.9% and sensitivity

to 39.7–43.4%, so most missed critical cases occurred among patients whose recorded vitals did not signal critical illness. By chief complaint, accuracy was lowest for abdominal/GI (52.8–64.6%) and neurological presentations (56.2–65.2%), and Resuscitation sensitivity was lowest for abdominal/GI (27.3–39.4%) and highest for chest/cardiac presentations (78.8–82.7%). These contrasts are exploratory, but they consistently identify vitally quiet and non-specific presentations as the highest-risk settings.

D. Repeatability

On the 240-case repeatability subsample (80 per category), complete agreement across the original run and two replicates ranged from 70.4% (MiMo V2.5) to 92.9% (Tencent Hy3). Agreement with the original July run (77.5–95.8%) was slightly lower than between the

Table 7. Prevalence-adjusted positive and negative predictive values for Resuscitation, Urgent, and Non-urgent CTM categories across four large language models, calculated using the reported Indonesian referral emergency department category prevalence of 14.9%, 63.8%, and 21.3%, respectively.

Model	Resus. PPV	Urgent PPV	Non-urg. PPV	Resus. NPV	Urgent NPV	Non-urg. NPV
DeepSeek V4 Flash	73.0	80.4	57.6	94.0	56.2	91.6
DeepSeek V4 Pro	73.0	79.5	56.9	93.3	57.3	90.8
MiMo V2.5	74.5	79.8	58.0	93.0	58.2	91.3
Tencent Hy3	83.3	78.4	65.1	93.3	65.3	88.9

two October replicates (84.6–94.6%) for three of four models, consistent with minor serving drift over three months. Category switching affected 7.1–29.6% of cases; 89–100% of switches were adjacent-category, and direct Resuscitation-to-Non-urgent switches occurred in at most 0.6% of case pairs. Undertriage detection also varied across runs (9.6–20.4% of undertriaged cases for DeepSeek V4 Pro and 46.3–64.3% for MiMo V2.5). Single-run estimates therefore carry material run-to-run noise, and small between-model differences should be read cautiously; replicate runs or agreement-based consensus are advisable before any safety-critical use.

IV. Discussion

A. Interpretation of Findings

For every model, a substantial share of truly critical cases was missed: Resuscitation sensitivity ranged from 58.5% to 64.8% (141–166 of 400 cases per model), and 89.2–94.4% of those misses were assigned Urgent rather than Non-urgent (Table 3: 128–151 versus 9–18 transitions). At the 14.9% Resuscitation prevalence of an Indonesian referral ED [3], a 100-patient day includes roughly 15 critical patients, of whom 5–6 would be underassigned by any evaluated model without clinician oversight. Pathway-wise, Resuscitation-to-Urgent errors (32.0–37.8% of critical cases) mainly delay time-critical intervention, whereas the rarer Resuscitation-to-Non-urgent errors (2.2–4.5%) bypass emergency resources entirely and pose the greatest hazard. No formal sensitivity threshold exists for CTM, yet against the ≥90% gate adopted for piloting (Section IV-A), every model fell 25 to 31 points short, a gap invisible in aggregate accuracy because the balanced sample over-represents Resuscitation relative to real case-mix.

The four models tracked each other closely despite differences in architecture and price: accuracy fell within a two-point spread (66.5–68.1%) with no significant pairwise differences, and kappa differed by at most 0.03. DeepSeek V4 Flash combined the highest accuracy with the lowest cost (\$0.21), whereas Tencent Hy3, the priciest option (\$3.95), matched the weakest performers; for budget-constrained systems, better triage does not require higher spending. Those four models spanning different architectures, parameter scales, and a more than tenfold price range landed within two points of accuracy, suggesting a performance ceiling set by the task rather than the models. Three coarse categories, a compact prompt, Indonesian-rendered clinical text, and substantial label ambiguity at CTM boundaries (Section III-A) all compress achievable discrimination, and the shared error pattern, with adjacent-category shifts dominating and the lowest sensitivity for vitally quiet presentations

(Section III-C), points to ambiguity in the inputs and reference labels rather than model-specific weakness. Separating these factors, prompt depth, translation fidelity, and boundary ambiguity, will require bilingual re-evaluation and adjudicated boundary subsets.

B. Implications for the Proposed Hybrid Triage System

These results speak to the cloud-based component of the hybrid system outlined in the Introduction, which places a cloud LLM at the front desk to support CTM classification, risk assessment, and specialist pathway recommendation. The models achieved moderate overall accuracy (66.5–68.1%), with strong Resuscitation PPV (88.4–93.4%) and Non-urgent specificity (85.9–91.6%). Only the classification module was evaluated; the ASR, offline fallback, routing, risk, and latency components remain unvalidated design concepts. The Urgent category showed the widest swings (sensitivity 68.8–81.8%, PPV 50.7–53.8%), where a structured rationale at ambiguous category boundaries could add the most value.

This workflow aligns with the human-in-the-loop paradigm: the model offers decision support while the clinician keeps final say [36], [37]. Wang et al. [38] recently showed LLM-based triage guidance is feasible in the ED. Current LLMs are not accurate enough for autonomous triage (ceiling 68.1%), yet a front-desk helper that flags undertriage risk and supplies a structured rationale could cut diagnostic error: MiMo V2.5 caught 56.6% of undertriaged cases, including 81.1% of the most dangerous pathway, suggesting that even a middling model can serve as a safety net. The LLM would operate strictly as secondary decision support: the nurse remains the decision-maker, and every output is a recommendation with a rationale the nurse can accept or override, with overrides recorded for audit. Two functions could carry immediate value: a structured second opinion at ambiguous CTM boundaries and an undertriage alert whenever the model flags undertriage risk or predicts a lower category than the nurse's provisional assignment, prompting reassessment rather than any automatic category change. Before any clinical pilot, three requirements should be met: pre-specified prospective safety thresholds, including Resuscitation sensitivity of at least 90% and detection of at least 75% of reference-confirmed undertriage in shadow mode; a response-time budget compatible with front-desk workflow, with end-to-end latency under 5 seconds per case; and prospective multicenter shadow deployment on authentic Indonesian records with predefined go/no-go gates on Resuscitation sensitivity, undertriage detection, and calibration at CTM boundaries.

Resource-limited emergency departments face practical constraints beyond cost: intermittent connectivity argues for the offline fallback already

designed but not yet validated; cloud queries must be limited to de-identified fields and reviewed against hospital and national data-protection requirements; and response latency, unmeasured here, must fit the seconds-scale demands of front-desk triage. Integration with existing hospital information systems will determine whether the tool removes or adds a step. Human-in-the-loop oversight remains unavoidable: unflagged cases must still receive full nurse scrutiny, and alerting at the highest observed flag rate (MiMo V2.5, 47.8% of cases; precision 25.5%) risks alarm fatigue.

C. Interpretability and Clinical Trust

The models also produced brief reasoning strings, but their quality was not assessed; interpretability benefits are therefore not claimed, and clinician-side

assessment of the rationales remains future work. The tension between accuracy and interpretability in clinical AI remains central [37]. LLMs produce free-text explanations clinicians can read, unlike traditional black-box models, but can also generate plausible yet clinically wrong reasoning, known as hallucination [39] that risks patient safety when a confident but incorrect recommendation delays care. Clinician acceptance depends on perceived usefulness and workflow integration; clinicians remain cautious about replacing human judgment with algorithmic output [40].

D. Comparison to Previous Studies

Table 8 compares this study with prior work, but with caution. Virtually all prior investigations used five-level frameworks (ESI, CTAS, or MTS) in English-language settings [16]–[19], [44]–[47]. Franc et al. [20] reported

Table 8. Systematic comparison of the closest prior triage evaluations and the present study across triage frameworks, datasets, models, accuracy, agreement, and undertriage outcomes.

Study	Country, triage system	Dataset and sample	Model(s)	Accuracy	Under triage / mis triage
Williams et al. [16]	USA, ESI (5-level)	10,000 paired presentations sampled from 251,401 adult ED visits; real records	GPT-4	89% (pairwise acuity separation)	comparable to physician classification (500-pair subsample)
Franc et al. [20]	Canada, CTAS (5-level)	61 patient vignettes; 10,980 repeated simulated triages	ChatGPT	47.5%	undertriage 13.7%; overtriage 38.7%
Sarbay et al. [41]	Turkey, ESI (5-level)	50 constructed case scenarios	ChatGPT	overall sensitivity 57.1%	undertriage 18%; overtriage 22%
Moon et al. [42]	South Korea, KTAS (5-level)	1,267 records from two EDs; real records	triage nurses versus expert panel	(benchmark study; no LLM)	mistriage 14.7%; 70.4% of errors were undertriage
Nedos et al. [43]	Greece, ESI (5-level)	39,375 real ED patients	7 LLMs	up to 61.7% (up to 81% in subgroups)	over/ undertriage tendencies reported
This study	Indonesia-context CTM (3-level); Indonesian translation of US records	1,200 sampled records (400 per CTM category)	4 low-cost LLMs	66.5–68.1%	undertriage 18.2–21.5%; detection 11.9–56.6%

47.5% accuracy with a 13.7% undertriage rate for ChatGPT on the Canadian Triage and Acuity Scale, with substantial variability across prompt formulations. The models evaluated here produced higher accuracy (66.5–68.1%), but this is indicative only, not evidence of superiority: the evaluations differ in triage system (CTM versus CTAS), category definitions, populations, class prevalence, prompting, and evaluation procedure. Sarbay et al. [41] reported a Cohen kappa of 0.341 for ChatGPT on Turkish ESI triage, below the 0.50–0.52 range observed here; kappa values are system-specific. Colakca et al. [18] achieved 76.6% accuracy with ChatGPT on Turkish ESI across 745 consecutive patients, possibly because three coarse CTM categories are harder than five finer-grained ESI levels, where adjacent-category errors are less penalised, and because of differences in case-mix and protocol. Moon et al. [42] reported a 14.7% mistriage rate for nurse triage with the Korean Triage and Acuity Scale, 70.4% of which was undertriage. These differences position the present results within the performance regime of current-generation models rather than superior triage ability; the simpler CTM structure reduces the clinician's cognitive burden, which may partially offset the lower per-class accuracy in decision-support use.

Across the five closest studies, datasets ranged from 50 scenarios to 39,375 patients, and reported agreement spanned weighted kappa 0.18 to 0.83 with undertriage of 13.7–18%; the present results (kappa 0.50–0.52; undertriage 18.2–21.5%; detection 11.9–56.6%) fall within these brackets. Only Franc et al. assessed repeatability, and only across prompt formulations rather than repeat runs; no prior study evaluated a three-tier system or reported run-to-run stability alongside classification accuracy, so comparisons here are limited to the direction of findings rather than their magnitudes. These results also dovetail with broader machine learning triage research: Raita et al. [48] and Goto et al. [49] showed that models trained on structured ED data predict triage-related outcomes with AUCs above 0.85, though they depend on hand-crafted features. Porto [50] noted how few studies have tested LLMs in non-English, resource-limited settings, and Névél et al. [51] identified key challenges for non-English clinical NLP: limited annotated corpora, language-specific preprocessing, and English-centric models, barriers especially acute for Indonesian clinical text. Nedos et al. [43] probed multiple LLMs on 39,375 ED patients in Greece, finding model-dependent agreement with physician-assigned ESI (weighted kappa 0.18–0.47; accuracy up to 61.7%), and Nourbakhsh [52] proposed an ED triage LLM benchmark for future CTM studies.

E. Strengths and Limitations

To the authors' knowledge, this is the first evaluation of language model performance on CTM triage: many studies tested LLMs on five-level frameworks in English, but none on simplified three-level systems in non-English languages, standard across Southeast Asia, Africa, and Latin America, where healthcare resources are scarcest. At 1,200 cases, the sample far exceeds most AI-triage studies (61 records in Franc et al. [20], 50 scenarios in Sarbay et al. [41], and 745 cases in Colakca et al. [18]), powering class-level analysis, especially for the rare but critical Resuscitation category.

Reporting cost-efficiency across price tiers gives practical guidance for administrators at under-resourced facilities, a dimension rarely addressed despite cost being decisive in AI adoption. The economic evaluation covers API expenditure only; deployment would add infrastructure, integration, maintenance, training, monitoring, and governance costs. The undertriage analysis addresses a patient-safety outcome that comparable studies seldom examine: most prior work reported only accuracy and kappa without probing misclassification patterns, whereas quantifying caught undertriaged cases by pathway offers a clinically richer picture. The study also followed recently published methodological standards for AI-based clinical decision support in emergency medicine [53].

Each limitation affects validity or deployment and requires a corresponding validation strategy. First, the evaluation relied on a single US-derived corpus (MIETIC): case-mix, documentation habits, vital-sign conventions, and diagnostic resources differ from Indonesian practice [7], [35], so external validity cannot be assumed. Although Indonesian emergency physicians translated and relabeled all records blind to the original acuity tags, they reviewed the translation only for clinical realism, so subtle information loss cannot be ruled out. The decisive next step is prospective validation on authentic Indonesian emergency department data. Second, the retrospective design is exposed to documentation and information bias, as records reflect routine care rather than a standardized examination; prospective collection with protocolized capture is required to reduce this dependence on charting quality [39], [54]. Third, the reference standard rested on expert consensus rather than patient outcomes, and agreement with the independent automated re-labeling was only fair (kappa = 0.33), so CTM boundary labels carry non-trivial uncertainty; the Section III-A sensitivity analysis localizes this uncertainty to a small boundary stratum, which bounds measured accuracy. Multi-reader reliability studies with human adjudication of disputed cases should accompany future evaluations. Fourth, the balanced 1:1:1 sampling makes predictive values

and error counts sample-dependent; prevalence-adjusted analyses mitigate but cannot replace evaluation on a naturally distributed cohort, which prospective studies should prioritize. Fifth, only four models were tested, each under a single prompt formulation with two to three examples per category, and most cases were scored once: replicate runs changed roughly a quarter of classifications (Section III-D), so single-run results carry run-to-run noise, and prompt design can swing performance further [20], [55], so results likely represent a floor rather than a ceiling, and prompt optimization and broader model coverage are future work. Finally, only the cloud pathway was exercised: speech recognition, the offline fallback, specialist routing, hospital-risk assessment, and latency (Fig. 1, dashed components) remain unvalidated design concepts, and system-level claims await end-to-end prospective testing.

F. Clinical Implications

The 66.5–68.1% accuracy ceiling rules out autonomous triage under CTM: even the best model missed 43.4% of undertriaged cases, incompatible with unsupervised use. Lower-cost models matched pricier alternatives, pointing to AI as a low-cost safety layer that augments rather than replaces human judgment; realizing that potential requires investment in data infrastructure, regulatory frameworks, and workforce readiness [56], and in helping clinicians understand what these tools can and cannot do [37].

G. Research Implications

Several research directions follow. First, the compact CTM summary in the prompt may partly explain the moderate accuracy: prompt formulation can swing LLM performance by more than 20 percentage points (Franc et al. [20] and Yao et al. [55]), so future work should test detailed criteria, few-shot examples, and structured outputs against the S1 baseline. Second, prospective multicenter validation in Indonesian EDs is essential: at 14.9% Resuscitation prevalence [3], Resuscitation PPV falls to 73–83% (Table 7) rather than the 88–93% observed here, and predictive values must be re-estimated on local cohorts. Third, hybrid approaches pairing LLM recommendations with rule-based triggers for vital-sign abnormalities (systolic blood pressure <90 mmHg, respiratory rate >30/min) could combine nuanced reasoning with hard safety thresholds for undertriage detection; Almeshairi et al. [57] showed LLMs can outperform emergency medicine trainees, but interpretability gaps still erode clinical trust [37], which will be the field's central challenge. Fourth, workflow integration must address implementation barriers: electronic health record integration, front-desk latency (target: under 5 seconds), transparent reasoning that lets nurses verify and override, and institutional governance spanning clinicians, informatics researchers, and administrators.

V. Conclusion

This study evaluated the diagnostic accuracy, undertriage safety, and API cost of four low-cost language models for CTM triage classification, as the cloud component of a proposed hybrid triage system for resource-limited emergency departments. All four delivered comparable accuracy (66.5–68.1%; kappa 0.50–0.52) with no significant pairwise differences, and lower-cost models matched pricier alternatives. Safety-critical performance was less reassuring: Resuscitation sensitivity reached only 58.5–64.8% and fell further when red-flag vital signs were absent (39.7–43.4%), and undertriage detection remained weak (11.9–56.6%). Replicate runs changed roughly a quarter of classifications, underlining the need for repeated inference in safety-critical use. These results position LLMs as adjunctive, clinician-overridden decision support rather than autonomous triage, and the fair reference-label reliability at CTM boundaries (kappa = 0.33) indicates that part of the measured error reflects label uncertainty itself. The evidence derives from retrospective testing on an artificially balanced sample of US-derived records and does not establish clinical effectiveness, safety, or deployability in Indonesian EDs. Future work should prioritize prospective, multicenter validation on authentic Indonesian data; end-to-end evaluation of the complete hybrid architecture; prompt optimization; and clinical workflow integration assessment before any implementation.

Acknowledgment

Their clinical expertise and dedication to triage accuracy were essential to the integrity of this study. The authors also acknowledge the PhysioNet repository for making the MIETIC Triage Instruction Corpus available for secondary analysis.

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The MIETIC Triage Instruction Corpus v1.0.0 is publicly available at PhysioNet: <https://physionet.org/content/mietic/1.0.0/>. The CTM-relabeled dataset (5,528 cases with physician consensus labels) used in this study is available from the corresponding author upon reasonable request. The complete prompt templates, output constraints, and evaluation procedure are disclosed in Supplementary Material S1. Analysis code is available from the corresponding author upon reasonable request.

Author Contribution

Ahmad Ridwan Fauzi conceptualized and designed the study, conducted data curation and formal analysis, performed investigation and methodology development, developed software and validation protocols, created visualizations, and wrote the original draft with review and editing. Adelia Chitra Sazkia contributed to conceptualization, data curation, investigation, and provided critical review and editing of the manuscript. Muhammad Ardiansyah contributed to methodology, provided supervision, and participated in writing review and editing. All authors reviewed and approved the final version of the manuscript and agreed to be responsible for all aspects of the work ensuring integrity and accuracy.

Declarations

Ethics Approval and Consent to Participate

This study involved a retrospective computational analysis of the publicly available, de-identified MIETIC Triage Instruction Corpus (version 1.0.0), which is derived from MIMIC-IV. No patients were recruited, and no intervention was performed. Because the analysis used only de-identified secondary data released under the PhysioNet credentialed data use agreement, the study qualified for exemption from full ethics review under the institutional research regulations of Telkom University, and no ethics approval number was therefore issued. Four board-certified emergency physicians took part in the clinical relabeling of the corpus to establish the consensus reference labels; their participation involved no patient contact and no access to identifiable information. The corpus was accessed in accordance with the PhysioNet data use agreement for MIMIC-IV, and the study was conducted in accordance with the Declaration of Helsinki (1964 and subsequent amendments).

Consent for Publication

Not applicable. The study did not include identifiable patient information, individual photographs, or any other personally identifiable material requiring publication consent.

Code Availability

The analyses were conducted in Python 3.12 with scikit-learn, pandas, SciPy, and statsmodels. The scripts that reproduce the model inference, the diagnostic accuracy calculations, the statistical comparisons, the prevalence adjustment, and the cost analysis are available from the corresponding author upon reasonable request.

Competing Interests

The authors declare that they have no known financial or non-financial competing interests that could have

influenced the design, conduct, analysis, interpretation, or reporting of this study.

Declaration of Generative AI and AI-Assisted Technologies

Generative artificial intelligence was the primary technology investigated in this study. Four large language models (DeepSeek V4 Pro, DeepSeek V4 Flash, Tencent Hy3, and MiMo V2.5) were evaluated for automated classification of emergency department records according to the Cipto Triage Method. The models were accessed through the OpenRouter API using standardized Indonesian-language inputs and predefined instructions, and MiMo V2.5 was additionally used to translate clinical records from English into Indonesian during dataset preparation. The translated records were reviewed by emergency medicine specialists, and physician-consensus labels served as the reference standard. Model outputs were treated strictly as experimental results. AI-generated classifications, rationales, and undertriage-risk flags were not used to make clinical decisions or to replace professional medical judgment. Generative AI tools were also used during manuscript preparation only for language editing; every such suggestion was reviewed, verified, and accepted or rejected by the authors. All authors retain full responsibility for the accuracy, originality, integrity, and scientific validity of the final manuscript.

Data Privacy and Confidentiality

The study used de-identified clinical records obtained from a publicly available research corpus, and no directly identifiable patient information was included in the experimental inputs; no re-identification was attempted. Clinical text was transmitted to the OpenRouter API over encrypted connections in de-identified form only, without names, dates, medical record numbers, or geographic identifiers. The authors did not authorize any secondary use of the transmitted records by the API service provider, retained no data beyond those required to reproduce the reported results, and complied with the access conditions of the original dataset and with institutional data-protection requirements.

Clinical Safety and Human Oversight

The evaluated models were investigated solely for research purposes. Their outputs were not used for autonomous triage or for the management of real patients, and no clinical decision in this study depended on model output. The proposed clinical decision support system is intended to assist health care professionals, with qualified clinical personnel retaining responsibility for the final triage decision. Any future implementation requires prospective clinical validation, a formal patient-

safety assessment, and institutional and regulatory approval.

Declaration of Originality

The authors declare that the manuscript presents original research, has not been published previously, and is not under consideration for publication elsewhere. All sources of information have been appropriately acknowledged, and all authors have approved the submitted version.

References

- [1] G. Fitzgerald, G. Jelinek, D. Scott, and M. Gerdzt, "Emergency department triage revisited," *Emerg. Med. J.*, vol. 27, no. 2, pp. 86–92, 2010, doi: <https://doi.org/10.1136/emj.2009.077081>.
- [2] J. M. Zachariasse, V. van der Hagen, N. Seiger, K. Mackway-Jones, M. van Veen, and H. A. Moll, "Performance of triage systems in emergency care: A systematic review and meta-analysis," *BMJ Open*, vol. 9, p. e026471, 2019, doi: <https://doi.org/10.1136/bmjopen-2018-026471>.
- [3] H. Habib, S. Sulistio, I. A. Albar, et al., "Validation of the Cipto Triage Method: A single-centre study from Indonesia," *Open Access Emerg. Med.*, vol. 12, pp. 137–143, 2020, doi: <https://doi.org/10.2147/OAEM.S246598>.
- [4] B. Haas et al., "Survival of the fittest: The hidden cost of undertriage of major trauma," *J. Am. Coll. Surg.*, vol. 211, no. 6, pp. 804–811, 2010, doi: <https://doi.org/10.1016/j.jamcollsurg.2010.08.014>.
- [5] M. A. Robertson and E. M. Molyneux, "Triage in the developing world—can it be done?," *Arch. Dis. Child.*, vol. 85, pp. 208–213, 2001, doi: <https://doi.org/10.1136/ad.85.3.208>.
- [6] N. Schwalbe and B. Wahl, "Artificial intelligence and the future of global health," *Lancet*, vol. 395, no. 10236, pp. 1579–1586, May 2020, doi: [https://doi.org/10.1016/S0140-6736\(20\)30226-9](https://doi.org/10.1016/S0140-6736(20)30226-9).
- [7] M. Mulyadi, B. Dedi, W.-L. Hou, I.-C. Huang, and B.-O. Lee, "Nurses' experiences of emergency department triage during the COVID-19 pandemic in Indonesia," *J. Nurs. Scholarsh.*, vol. 54, no. 1, pp. 15–23, 2022, doi: <https://doi.org/10.1111/jnu.12709>.
- [8] H. Habib, R. M. Mulyana, S. Sulistio, and I. A. Albar, "Hospital modern triage and its application in Indonesia," *Medika*, Jakarta, Indonesia, 2016. [Online]. Available: <https://www.researchgate.net/publication/311715654>
- [9] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023, doi: <https://doi.org/10.1038/s41586-023-06291-2>.
- [10] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, et al., "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digit. Health*, vol. 2, p. e0000198, 2023, doi: <https://doi.org/10.1371/journal.pdig.0000198>.
- [11] J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, et al., "The future landscape of large language models in medicine," *Commun. Med.*, vol. 3, p. 141, 2023, doi: <https://doi.org/10.1038/s43856-023-00370-1>.
- [12] A. Rao, M. Pang, J. Kim, M. Kamineni, et al., "Assessing the utility of ChatGPT throughout the entire clinical workflow: Development and usability study," *J. Med. Internet Res.*, vol. 25, p. e48659, 2023, doi: <https://doi.org/10.2196/48659>.
- [13] A. Gilson, C. Safranek, T. Huang, V. Socrates, et al., "How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment," *JMIR Med. Educ.*, vol. 9, p. e45312, 2023, doi: <https://doi.org/10.2196/45312>.
- [14] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, et al., "Revolutionizing healthcare: The role of artificial intelligence in clinical practice," *BMC Med. Educ.*, vol. 23, p. 689, 2023, doi: <https://doi.org/10.1186/s12909-023-04698-z>.
- [15] M. Fernandes, S. M. Vieira, F. Leite, C. Palos, S. Finkelstein, and J. M. C. Sousa, "Clinical decision support systems for triage in the emergency department using intelligent systems: A review," *Artif. Intell. Med.*, vol. 102, p. 101762, Jan. 2020, doi: <https://doi.org/10.1016/j.artmed.2019.101762>.
- [16] C. Y. K. Williams, T. Zack, B. Y. Miao, M. Sushil, et al., "Use of a large language model to assess clinical acuity of adults in the emergency department," *JAMA Netw. Open*, vol. 7, no. 5, p. e248895, 2024, doi: <https://doi.org/10.1001/jamanetworkopen.2024.8895>.
- [17] J. H. Kim, S. K. Kim, J. Choi, and Y. Lee, "Reliability of ChatGPT for performing triage task in the emergency department using the Korean Triage and Acuity Scale," *Digit. Health*, vol. 10, art. no. 20552076241227132, 2024, doi: <https://doi.org/10.1177/20552076241227132>.

- [18] C. Colakca, M. Ergin, H. S. Ozensoy, A. Sener, S. Gures, and A. Ozhasenekler, "Emergency department triaging using ChatGPT based on emergency severity index principles: A cross-sectional study," *Sci. Rep.*, vol. 14, art. no. 22106, 2024, doi: <https://doi.org/10.1038/s41598-024-73229-7>.
- [19] B. Arslan, C. Nuhoglu, M. Saticio, and E. Altinbilek, "Evaluating LLM-based generative AI tools in emergency triage: A comparative study of ChatGPT Plus, Copilot Pro, and triage nurses," *Am. J. Emerg. Med.*, vol. 89, pp. 174–181, 2025, doi: <https://doi.org/10.1016/j.ajem.2024.12.024>.
- [20] J. M. Franc, L. Cheng, A. Hart, R. Hata, and A. Hertelendy, "Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale," *Can. J. Emerg. Med.*, vol. 26, no. 1, pp. 40–46, 2024, doi: <https://doi.org/10.1007/s43678-023-00616-w>.
- [21] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, et al., "STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies," *BMJ*, vol. 351, p. h5527, 2015, doi: <https://doi.org/10.1136/bmj.h5527>.
- [22] G. S. Collins, J. A. Hudson, R. D. Riley, et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, 2024, doi: <https://doi.org/10.1136/bmj-2023-078378>.
- [23] W. G. Cochran, *Sampling Techniques*, 3rd ed. New York, NY, USA: Wiley, 1977.
- [24] OpenRouter, "Create a chat completion," *OpenRouter API Reference*, OpenRouter, Inc., 2026. [Online]. Available: <https://openrouter.ai/docs/api/api-reference/chat/create-a-chat-completion>. [Accessed: Jul. 14, 2026]
- [25] DeepSeek, "DeepSeek V4 Pro 0423," *OpenRouter model card*, model identifier deepseek/deepseek-v4-pro, DeepSeek, released Apr. 24, 2026. [Online]. Available: <https://openrouter.ai/deepseek/deepseek-v4-pro>. [Accessed: Jul. 14, 2026]
- [26] DeepSeek, "DeepSeek V4 Flash 0423," *OpenRouter model card*, model identifier deepseek/deepseek-v4-flash, DeepSeek, released Apr. 24, 2026. [Online]. Available: <https://openrouter.ai/deepseek/deepseek-v4-flash>. [Accessed: Jul. 14, 2026]
- [27] Tencent, "Hy3," *OpenRouter model card*, model identifier tencent/hy3, Tencent, released Jul. 6, 2026. [Online]. Available: <https://openrouter.ai/tencent/hy3>. [Accessed: Jul. 14, 2026]
- [28] Xiaomi, "MiMo-V2.5," *OpenRouter model card*, model identifier xiaomi/mimo-v2.5, Xiaomi, released Apr. 22, 2026. [Online]. Available: <https://openrouter.ai/xiaomi/mimo-v2.5>. [Accessed: Jul. 14, 2026]
- [29] D. G. Altman and J. M. Bland, "Diagnostic tests 2: Predictive values," *BMJ*, vol. 309, no. 6947, p. 102, 1994, doi: <https://doi.org/10.1136/bmj.309.6947.102>.
- [30] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960, doi: <https://doi.org/10.1177/001316446002000104>.
- [31] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971, doi: <https://doi.org/10.1037/h0031619>.
- [32] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *J. Am. Stat. Assoc.*, vol. 22, no. 158, pp. 209–212, 1927, doi: <https://doi.org/10.1080/01621459.1927.10502953>.
- [33] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947, doi: <https://doi.org/10.1007/BF02295996>.
- [34] R. A. Armstrong, "When to use the Bonferroni correction," *Ophthalmic Physiol. Opt.*, vol. 34, no. 5, pp. 502–508, 2014, doi: <https://doi.org/10.1111/opo.12131>.
- [35] J. S. Hinson, D. A. Martinez, P. S. K. Schmitz, M. Toerper, et al., "Accuracy of emergency department triage using the Emergency Severity Index and independent predictors of under-triage and over-triage in Brazil: A retrospective cohort analysis," *Int. J. Emerg. Med.*, vol. 11, art. no. 3, 2018, doi: <https://doi.org/10.1186/s12245-017-0161-8>.
- [36] R. T. Sutton, D. Pincock, D. C. Baumgart, D. Sadowski, et al., "An overview of clinical decision support systems: Benefits, risks, and strategies for success," *NPJ Digit. Med.*, vol. 3, p. 17, 2020, doi: <https://doi.org/10.1038/s41746-020-0221-y>.
- [37] A. Rajaram, H. Li, J. K. Holodinsky, J. N. Hall, et al., "Opening the black box: Challenges and opportunities regarding interpretability of artificial intelligence in emergency medicine," *Can. J. Emerg. Med.*, vol. 27, no. 2, pp. 83–86, 2025, doi: <https://doi.org/10.1007/s43678-024-00827-9>.
- [38] C. Wang et al., "Patient triage and guidance in emergency departments using large language

- models: Multimetric study," *J. Med. Internet Res.*, vol. 27, p. e71613, 2025, doi: <https://doi.org/10.2196/71613>.
- [39] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, pp. 625–630, 2024, doi: <https://doi.org/10.1038/s41586-024-07421-0>.
- [40] T. Nadarzynski, O. Miles, A. Cowie, and D. Ridge, "Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study," *Digit. Health*, vol. 5, art. no. 2055207619871808, 2019, doi: <https://doi.org/10.1177/2055207619871808>.
- [41] I. Sarbay, G. B. Berikol, and I. U. Ozturan, "Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT)," *Turk. J. Emerg. Med.*, vol. 23, no. 3, pp. 156–161, 2023, doi: https://doi.org/10.4103/tjem.tjem_79_23.
- [42] S.-H. Moon, J. L. Shim, K.-S. Park, and C.-S. Park, "Triage accuracy and causes of mistriage using the Korean Triage and Acuity Scale," *PLoS One*, vol. 14, no. 9, p. e0216972, Sep. 2019, doi: <https://doi.org/10.1371/journal.pone.0216972>.
- [43] I. Nedos et al., "Is artificial intelligence ready for emergency department triage? A retrospective evaluation of multiple large language models in 39,375 patients at a university emergency department," *J. Clin. Med.*, vol. 15, no. 4, p. 1512, 2026, doi: <https://doi.org/10.3390/jcm15041512>.
- [44] G. Ben Haim, M. Saban, Y. Barash, D. Cirulnik, et al., "Evaluating large language model-assisted emergency triage: A comparison of acuity assessments by GPT-4 and medical experts," *J. Clin. Nurs.*, vol. 35, no. 10, pp. 4249–4255, 2026, doi: <https://doi.org/10.1111/jocn.17490>.
- [45] J. M. Hoppe, M. K. Auer, A. Struven, S. Massberg, et al., "ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: Retrospective analysis," *J. Med. Internet Res.*, vol. 26, p. e56110, 2024, doi: <https://doi.org/10.2196/56110>.
- [46] F. Gaber, M. Shaik, F. Allegra, A. J. Bilecz, et al., "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis," *NPJ Digit. Med.*, vol. 8, art. no. 263, 2025, doi: <https://doi.org/10.1038/s41746-025-01684-1>.
- [47] H. Fraser, D. Crossland, I. Bacher, M. L. Ranney, et al., "Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: Clinical data analysis study," *JMIR Mhealth Uhealth*, vol. 11, p. e49995, 2023, doi: <https://doi.org/10.2196/49995>.
- [48] Y. Raita, T. Goto, M. K. Faridi, D. F. Brown, et al., "Emergency department triage prediction of clinical outcomes using machine learning models," *Crit. Care*, vol. 23, art. no. 64, 2019, doi: <https://doi.org/10.1186/s13054-019-2351-7>.
- [49] T. Goto, C. A. Camargo, M. K. Faridi, R. J. Freishtat, et al., "Machine learning-based prediction of clinical outcomes for children during emergency department triage," *JAMA Netw. Open*, vol. 2, p. e186937, 2019, doi: <https://doi.org/10.1001/jamanetworkopen.2018.6937>.
- [50] B. M. Porto, "Improving triage performance in emergency departments using machine learning and natural language processing: A systematic review," *BMC Emerg. Med.*, vol. 24, art. no. 219, 2024, doi: <https://doi.org/10.1186/s12873-024-01135-2>.
- [51] A. Névél, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than English: Opportunities and challenges," *J. Biomed. Semantics*, vol. 9, no. 1, p. 12, Dec. 2018, doi: <https://doi.org/10.1186/s13326-018-0179-8>.
- [52] S. Nourbakhsh, "Skyer: A novel benchmark for evaluating the effectiveness of large language models in emergency department triage," *Can. J. Emerg. Med.*, vol. 28, 2026, doi: <https://doi.org/10.1007/s43678-026-01214-2>, published online Jul. 11, 2026.
- [53] H. Kareemi, H. Li, A. Rajaram, J. K. Holodinsky, J. N. Hall, et al., "Establishing methodological standards for the development of artificial intelligence-based clinical decision support in emergency medicine," *Can. J. Emerg. Med.*, vol. 27, no. 2, pp. 87–95, 2025, doi: <https://doi.org/10.1007/s43678-024-00826-w>.
- [54] P. Hager, F. Jungmann, R. Holland, K. Bhagat, et al., "Evaluation and mitigation of the limitations of large language models in clinical decision-making," *Nat. Med.*, vol. 30, pp. 2613–2622, 2024, doi: <https://doi.org/10.1038/s41591-024-03097-1>.
- [55] G. Yao, W. Zhang, Y. Zhu, U.-K. Wong, Y. Zhang, and C. Yang, "Comparing the accuracy of large language models and prompt engineering in diagnosing real-world cases," *Int. J. Med. Inform.*, vol. 203, p. 106026, 2025, doi: <https://doi.org/10.1016/j.ijmedinf.2025.106026>.
- [56] H. Kareemi, A. Colak, K. Yadav, W. Michalowski, and C. Vaillancourt, "Barriers and facilitators to developing and implementing

artificial intelligence-based clinical decision support in the emergency department: A qualitative study," *Can. J. Emerg. Med.*, vol. 28, 2026, doi: <https://doi.org/10.1007/s43678-026-01189-0>, published online Jun. 22, 2026.

- [57] N. Almehairi, G. Clark, and S. Davis, "How well does GPT-4 perform on an emergency medicine board exam? A comparative assessment," *Can. J. Emerg. Med.*, vol. 27, no. 12, pp. 969–973, 2025, doi: <https://doi.org/10.1007/s43678-025-00951-0>.

Author Biography



Ahmad Ridwan Fauzi received the B.E. degree in computer science from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, in 2017, the M.Sc. degree in information science and technology in 2020, and the M.B.A. degree from EDHEC Business School, Nice, France, in 2023. His major field of study encompasses computer science, information systems, and business administration.

He co-founded KitaLabs Co., Ltd. in Sapporo, Japan, in 2018, where he served as Chief Technology Officer and led the development of an online marketplace platform, raising seed funding from a regional newspaper publisher in Japan and growing the user base from zero to 1,000 active users within one year. He subsequently worked as a Senior Software Engineer with Xfers, Singapore, developing eKYC systems for banking applications using React and Ruby on Rails. He then served as Manager with Pradipa.co, Jakarta, Indonesia, overseeing international trade operations for Indonesian handicraft producers. He is currently a Lecturer with the Information Systems Study Program, Faculty of Industrial Engineering, Telkom University, Jakarta, Indonesia, where he teaches Enterprise Application Integration, Application Testing, and Algorithm courses. His current research interests include large language models, clinical decision support systems, and healthcare informatics, while his previous research focused on software engineering and business process optimization.



Adelia Chitra Sazkia received the B.S. degree in information systems from Telkom University, Bandung, Indonesia, in 2015, and the M.Sc. degree in computer science with a specialization in applied computing from the University of Malaya, Kuala Lumpur, Malaysia, in 2023. Her major field of study encompasses information systems and computer science.

She worked as an Integrated Channel Marketing Manager with PT XL Planet, Jakarta, Indonesia, from

2016 to 2018, where she oversaw multi-channel digital marketing operations including app push notifications, email newsletters, and LINE Official Account communications. She subsequently served as a Customer Lifecycle Management Planner with the same company from 2018 to 2019, managing end-to-end automation campaigns for customer retention and engagement. She is currently a Lecturer with the Information Systems Study Program, Faculty of Industrial Engineering, Telkom University, Jakarta, Indonesia. Her master's research included non-invasive stress detection using machine learning on multimodal datasets. Her current research interests include customer relationship management, marketing automation, and machine learning applications, while her previous research focused on gamification techniques for programming education.



Muhammad Ardiansyah was born in Jambi, Indonesia, on August 12, 1992. He received the B.S. degree in information systems from BINUS University, Jakarta, Indonesia, in 2014, and the M.S. degree in business management from BINUS Business School, Jakarta, Indonesia, in 2016. His major field of study encompasses information systems and business management.

He served as a Faculty Member with the Information Systems Study Program, Universitas Internasional Batam, Batam, Indonesia, from 2016 to 2025, where he also held the position of Head of the Academic Development Center. During his tenure, he led the Digital Learning Innovation Project funded by the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia in 2019. He is currently a Lecturer with the Information Systems Study Program, Faculty of Industrial Engineering, Telkom University, Jakarta, Indonesia. He has published scholarly articles on strategic information systems planning, Scrum methodology implementation for web-based systems, and self-service technology evaluations. His current research interests include strategic information systems, business process reengineering, and enterprise systems, while his previous research focused on information systems planning and agile methodologies.