

A Systematic Literature Review of Large Language Models in Clinical Triage from Risk Stratification to Human Escalation

Manpreet Singh

Boston University

Muhammad Zeeshan

Media Mind Management LLC

Abhijeet Mukkawar

Siemens (United States)

Rahul Joshi

rahulj@sitpune.edu.in

Symbiosis International University

Asha Pandit Ghodake

Bharati Vidyapeeth College of Engineering

Hanamant Bhagwan Sale

Bharati Vidyapeeth College of Engineering

Systematic Review

Keywords: Large Language Models, Clinical Triage, Risk Stratification, Patient Safety, Uncertainty Calibration, Human-in-the-Loop, Systematic Review, Bibliometrics

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11140640/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Systematic Literature Review of Large Language Models in Clinical Triage from Risk Stratification to Human Escalation

Manpreet Singh ¹, Muhammad Zeeshan ²,
Abhijeet Mukkawat ³, Rahul Joshi ^{4*},
Asha Pandit Ghodake ⁵, Hanamant Bhagwan Sale ⁶

¹Department of Computer Science, Boston University, Boston, MA, USA.

²Media Mind Management LLC., USA.

³Siemens Digital Industries Software, Dallas, USA.

^{4*}Symbiosis Institute of Technology, Pune, Symbiosis International (Deemed University), Pune, India.

⁵Department of Electronics & Telecommunication Engineering, Bharati Vidyapeeth College of Engineering, Kharghar, Navi Mumbai, 400614, India.

⁶Department of Information Technology, Bharati Vidyapeeth College of Engineering, Kharghar, Navi Mumbai, 400614, India.

*Corresponding author(s). E-mail(s): rahulj@sitpune.edu.in;
Contributing authors: manni@bu.edu; asha.p.ghodake@gmail.com;
hanamantsale@gmail.com;

Abstract

Emergency department overcrowding and the growth of digital health intake increase the need for rapid, reliable assessment of patient acuity. LLMs can help interpret clinical narratives, but their use in high-stakes triage also introduces risks from hallucinated advice, under-triage, and poorly calibrated uncertainty. This systematic and bibliometric review synthesizes $n = 89$ primary studies published between 2022 and 2026. We identified 402 records, screened 363 unique records after deduplication, and sought full texts for 91 reports; two reports were unavailable, leaving 89 studies for synthesis. We organize the evidence by triage task, LLM adaptation strategy (zero-shot, RAG, fine-tuning, and

multi-agent frameworks), evaluation measure (ESI/KTAS alignment, diagnostic accuracy, and under-triage penalty), safety mechanism (rule-based red-flag interception and verification agents), and human escalation pathway. The corpus included 10 studies published in 2024 and 49 published by the September 22, 2026 search cutoff. Emergency Department applications accounted for 47.2% of studies, and proprietary backbones for 41.6%. Separate syntheses address RQ1 through RQ5. The review also proposes a 5-dimension Safety and Uncertainty Evaluation Framework for clinical translation.

Keywords: Large Language Models, Clinical Triage, Risk Stratification, Patient Safety, Uncertainty Calibration, Human-in-the-Loop, Systematic Review, Bibliometrics

1 Introduction

Emergency department (ED) overcrowding, longer waits, and increasing patient complexity challenge healthcare systems worldwide. Triage must identify immediately life-threatening presentations, order care by urgency, and support the use of limited emergency resources. Standardized acuity frameworks combine clinical observations, presenting complaints, and vital-sign thresholds to assign urgency from Level 1 (Resuscitation / Immediate) to Level 5 (Non-urgent). Examples include the Emergency Severity Index (ESI) [2], the Korean Triage and Acuity Scale (KTAS) [3], the Canadian Triage and Acuity Scale (CTAS), and the Manchester Triage System (MTS).

During demand surges and pre-hospital EMS work, clinician-assigned triage may differ across observers, and cognitive or administrative workload can further reduce consistency. Consumer-facing portals and telehealth services also receive unfiltered symptom reports that must be sorted before clinical assessment.

Clinical NLP now uses foundation and generative systems, including OpenAI’s GPT-4/GPT-4o [4], Meta’s LLaMA-2/3 [5], Google’s Gemini, and Anthropic’s Claude. The triage literature also evaluates biomedical encoder models such as BioBERT and ClinicalBERT, as well as open-weight medical LLMs such as Meditron. Unlike generative LLMs, the biomedical encoders produce contextual representations rather than generated text. Conventional classifiers usually require structured tabular inputs. Generative LLMs can instead interpret narrative histories, integrate their details, and produce text-based assessments.

These capabilities do not remove the ethical, regulatory, and safety risks of using an unconstrained generative model for triage. LLMs can produce hallucinated guidance, agree with a user’s framing, and express confidence that is not calibrated to correctness. Triage errors also have unequal consequences. Over-triage, in which a stable patient receives a higher acuity tier, mainly consumes additional capacity. Under-triage, in which a critically ill or deteriorating patient is assigned low acuity, can delay treatment and result in serious harm or death.

Prior reviews do not resolve how LLM triage systems handle safety guardrails, uncertainty estimation, logit calibration, and fail-safe escalation to clinicians. We

therefore conducted a systematic and bibliometric review of $n = 89$ primary studies identified from $N = 363$ screened records published between 2022 and 2026.

1.1 Related Work and Differentiation

Existing reviews have considered artificial intelligence and large language models across healthcare, but they either cover a broader clinical scope or do not examine triage safeguards in technical detail. Table 1 compares this review with earlier surveys across six technical dimensions.

Table 1 Comparison with Existing Literature Reviews.

Published Survey	Target Focus Scope	Corpus	Asymmetric Audit	Safety Guardrails	Logit Uncertainty	Human Escalation
Thirunavukarasu et al. (2023) [7]	General Clinical Medicine	Narrative	No	Partial	No	No
Clusmann et al. (2023) [8]	Multi-Modal Clinical AI	Narrative	No	No	No	No
Cascella et al. (2023) [9]	ChatGPT Emergency Feasibility	Qualitative	No	No	No	Partial
Wu et al. (2024) [10]	Broad Healthcare LLMs	90 Studies	No	Partial	No	No
Avery et al. (2024) [11]	Symptom Checker Bots	42 Studies	Partial	No	No	Partial
This Systematic Review	LLM Clinical Triage	89 Studies	Yes	Yes	Yes	Yes

Earlier reviews differ from this study in both scope and method. Thirunavukarasu et al. [7] and Clusmann et al. [8] surveyed LLM applications across specialties such as radiology, pathology, and clinical documentation; neither used PRISMA screening to extract quantitative evidence on acute triage. Cascella et al. [9] examined ChatGPT in emergency medicine through simulated vignettes, with qualitative analysis rather than a multi-database systematic review.

Two narrower reviews also leave technical questions open. Wu et al. [10] included 90 healthcare LLM studies but treated triage as a secondary use case; it did not separately analyze ESI/KTAS alignment, under-triage, or multi-agent designs. Avery et al. [11] reviewed 42 conversational symptom-checker studies, most of which evaluated rule-based systems predating LLMs. Its analysis did not cover RAG, logit-entropy calibration, or automated 911/EMS escalation gateways.

Our systematic screening and bibliometric analysis included $n = 89$ studies published from 2022 to 2026. We examined under-triage alongside over-triage, and extracted evidence on deterministic guardrails, logit-based calibration, and five-tier human escalation. These analyses inform the 5-pillar taxonomy and 5-dimension safety framework presented below.

We address five research questions:

1. **RQ1 (Clinical Context & Task Spectrum):** Which clinical settings and functional triage tasks do current LLM systems target?
2. **RQ2 (Technical Adaptation & Model Backbones):** Which backbones and technical adaptation paradigms predominate in clinical-triage research?
3. **RQ3 (Performance & Acuity Alignment):** How closely do LLMs align with formal triage standards (ESI, KTAS), and how are diagnostic accuracy and under-triage rates distributed?
4. **RQ4 (Safety Guardrails & Uncertainty Calibration):** Which mechanisms quantify uncertainty, calibrate confidence, and reduce missed red flags?
5. **RQ5 (Human Oversight & Escalation Protocols):** How are clinician-verification co-pilots and dynamic emergency-escalation protocols designed?

The review contributes an empirical synthesis, a bibliometric analysis, and two organizing frameworks:

- **Empirical Synthesis:** Quantitative extraction and mapping of $n = 89$ validated primary studies published between 2022 and 2026.
- **Bibliometric Analysis:** Publication growth, international co-authorship, and thematic keyword clusters mapped with a Callon centrality-density matrix.
- **Formal Taxonomy & Evaluation Framework:** A 5-pillar Triage Task Taxonomy and a 5-dimension Safety and Uncertainty Evaluation Framework for assessing benchmark systems before clinical deployment.

2 Methodology

We conducted the review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement [1]. No prospective protocol was registered with PROSPERO or the Open Science Framework (OSF) before database indexing. PROSPERO limits registration to reviews of direct patient clinical outcomes and generally excludes computational assessments of software algorithms or machine-learning benchmarks. Before extraction, we specified the research questions, search queries, eligibility criteria, screening workflow, and extraction taxonomy to constrain retrospective selection and enable reproducible bibliometric analysis.

2.1 PRISMA 2020 Protocol and Study Selection

Selection had four stages: record identification; deduplication and title and abstract screening; full-text assessment against the PRISMA 2020 criteria; and inclusion in the quantitative extraction matrix.

Figure 1 reports the search and screening counts. The searches retrieved $N = 402$ records, $n = 356$ from Scopus and $n = 46$ from IEEE Xplore. Exact DOI matching and normalized-title comparison removed $n = 39$ duplicates, yielding $N = 363$ unique citations. We excluded $n = 272$ records at screening because they did not evaluate

generative artificial intelligence, large language model architectures, or clinical triage operations.

Full texts were sought for the remaining $N = 91$ reports. Study S14 and Study S52 could not be retrieved despite author outreach. The $n = 89$ reports available in full text met the eligibility criteria and entered quantitative extraction and bibliometric synthesis.

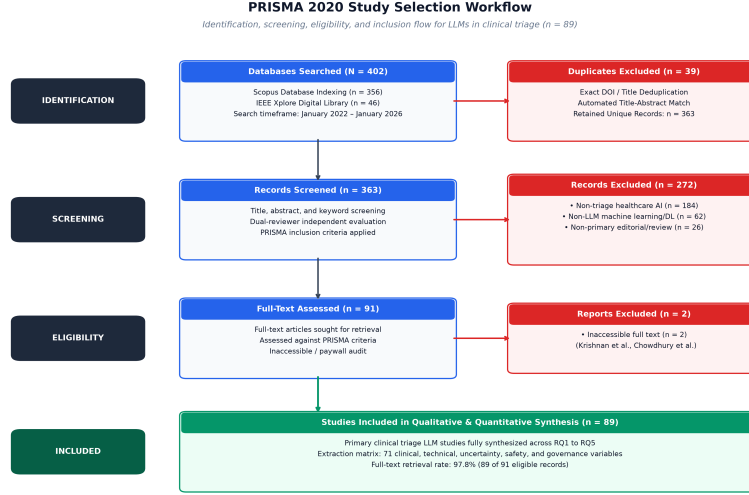


Fig. 1 PRISMA 2020 Flowchart of Study Selection.

2.2 Information Sources and Database Search Strategy

The search corpus came from Scopus and IEEE Xplore, which together cover computer science, biomedical engineering, and emergency medicine relevant to this review. Searches covered January 1, 2022, through September 22, 2026.

The database-specific queries were:

Scopus Database Query String:

```
TITLE-ABS-KEY ( ( "large language model*" OR "LLM*" OR "GPT-4*" OR
"GPT-3.5*" OR "LLaMA*" OR "foundation model*" OR "generative AI" ) AND (
"triage" OR "risk stratification" OR "acuity classification" OR "emergency
severity index" OR "KTAS" OR "symptom assessment" OR "red flag*" ) AND (
"clinical decision support" OR "emergency department" OR "telehealth" OR
"prehospital" OR "intake" ) )
```

IEEE Xplore Command Search Query String:

```
((("Document Title": "large language model" OR "LLM" OR "GPT-4" OR
"generative AI" OR "foundation model") OR ("Abstract": "large language
model" OR "LLM" OR "GPT-4" OR "generative AI" OR "foundation model"))
```

AND (("Document Title": triage OR "risk stratification" OR "acuity classification" OR "emergency severity index" OR "symptom assessment") OR ("Abstract": triage OR "risk stratification" OR "acuity classification" OR "emergency severity index" OR "symptom assessment"))

2.3 Eligibility Criteria (PRISMA Framework)

We applied the PRISMA 2020 eligibility criteria summarized in Table 2.

Table 2 PRISMA Eligibility Criteria for Study Inclusion.

PRISMA Domain	Inclusion Criteria	Exclusion Criteria
Clinical Population & Setting	Human clinical presentations, emergency chief complaints, vital signs, EHR notes, or telehealth triage dialogues.	Non-clinical cohorts, synthetic text lacking medical grounding, non-triage hospital workflows (e.g., billing).
Intervention & AI System	Generative large language models (GPT-4, LLaMA, Gemini, Claude, Meditron) and biomedical encoder models (BioBERT, ClinicalBERT) evaluated for clinical triage.	Traditional non-generative ML (SVM, Decision Trees, standard RNNs without LLM reasoning); non-triage NLP tasks. Biomedical encoders were included only when evaluated for a clinical-triage task.
Comparator & Reference	Physician/nurse consensus standards, formal acuity scales (ESI, KTAS, CTAS, MTS), or baseline zero-shot LLMs.	Absence of validated clinical reference standard or comparative benchmark.
Outcomes & Metrics	Quantitative metrics: accuracy, sensitivity/recall, under-triage/over-triage rates, ESI concordance, or safety guardrails.	Purely qualitative opinion pieces without empirical performance or safety metrics.
Study Design & Scope	Peer-reviewed journal articles, archival conference proceedings, and validated benchmark preprints with complete methodology.	Editorials, short abstracts (< 2 pages), commentaries, non-English publications, inaccessible full texts ($n = 2$).

2.4 Data Extraction Matrix and Standardized Parameters

We used a standard extraction template to record 71 parameters for each study across six domains: Study Identification, Clinical Context & Setting, Technical Architecture, Performance & Alignment, Safety & Uncertainty Guardrails, and Human Oversight & Escalation.

2.5 Quality Assessment and Risk of Bias Audit

We assessed methodological quality and risk of bias with an adapted framework for clinical AI decision-support systems. It combines criteria from QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) and ROBINS-I (Risk of Bias in Non-randomized Studies of Interventions). Each of the $n = 89$ primary studies was audited

across five domains: (1) Patient Selection Bias, (2) Model Input Integrity and Benchmark Data Leakage, (3) Reference Standard Rigor, (4) Flow and Timing, and (5) Algorithmic Safety and Uncertainty Audit.

Table 3 gives the aggregate risk distribution. Patient-selection bias was high in 38.2% ($n = 34$) of studies. This rating mainly reflected the use of artificial, pre-constructed vignettes (24.7%, $n = 22$) or single-center retrospective cohorts without external institutional validation (13.5%, $n = 12$). Benchmark leakage or input contamination posed a high risk in 41.6% ($n = 37$). Those studies evaluated foundation models on public EHR benchmarks, including MIMIC-IV-ED and PhysioNet challenge sets, without rigorous temporal or contamination audits, although the data may have appeared in pretraining or instruction-tuning corpora. Algorithmic safety auditing was high risk in 67.4% ($n = 60$) of studies. These papers reported neither uncertainty quantification nor fail-safe abstention triggers.

Table 3 Aggregate Risk of Bias Distribution across Included Primary Studies ($n = 89$).

Bias Domain	Evaluation	Low Risk	Unclear	High Risk	Primary Drivers	Methodological
1. Patient Selection		39.3% ($n = 35$)	22.5% ($n = 20$)	38.2% ($n = 34$)	Synthetic vignettes ($n = 22$); single-center cohorts without external site validation ($n = 12$).	
2. Input Integrity & Data Leakage		27.0% ($n = 24$)	31.5% ($n = 28$)	41.6% ($n = 37$)	Benchmark contamination on public EHR datasets (MIMIC-IV-ED); absent temporal split audit.	
3. Reference Standard Rigor		59.6% ($n = 53$)	24.7% ($n = 22$)	15.7% ($n = 14$)	Unadjudicated single-nurse triage labels; lack of multi-physician consensus panels.	
4. Flow and Timing		76.4% ($n = 68$)	12.4% ($n = 11$)	11.2% ($n = 10$)	Variable retrospective documentation latency between nurse intake and physician sign-off.	
5. Algorithmic Safety Audit		13.5% ($n = 12$)	19.1% ($n = 17$)	67.4% ($n = 60$)	Complete absence of epistemic uncertainty estimation, logit entropy, or selective abstention.	

3 Bibliometric Growth and Global Trends

3.1 Publication Growth (2022 to 2026)

Figure 2A shows the publication counts. Ten studies (11.2%) met the inclusion criteria in 2024. The count tripled to 30 (33.7%) in 2025 and reached 49 (55.1%) in 2026. No studies published in 2022 or 2023 met the criteria for generative LLM clinical triage.

Figure 2B summarizes the distribution across clinical triage settings: Emergency Department (47.2%, $n = 42$), Outpatient Clinic Intake (22.5%, $n = 20$), Pre-hospital EMS (14.6%, $n = 13$), Remote Digital Telehealth (11.2%, $n = 10$), and Specialty Referral (4.5%, $n = 4$).

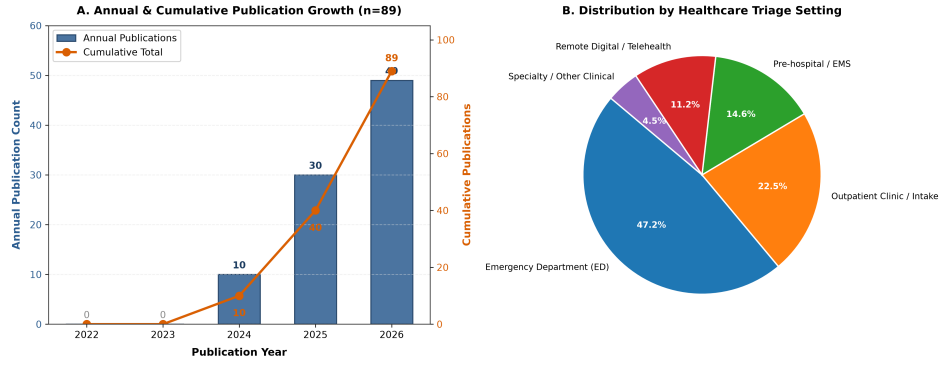


Fig. 2 Publication Growth Trajectory and Triage Settings.

3.2 Geographic Output and Cross-Border Collaboration Networks

The 89 studies involved authors from 27 countries. Figure 3A lists the ten most represented countries: United States ($n = 29$, 32.6%), China ($n = 17$, 19.1%), India ($n = 11$, 12.4%), South Korea ($n = 7$, 7.9%), Germany ($n = 6$, 6.7%), France ($n = 5$, 5.6%), Turkey ($n = 5$, 5.6%), Italy ($n = 5$, 5.6%), Israel ($n = 4$, 4.5%), and United Kingdom ($n = 4$, 4.5%). Single-country studies made up 73.0% ($n = 65$). International co-authorship appeared in the remaining 27.0% ($n = 24$).

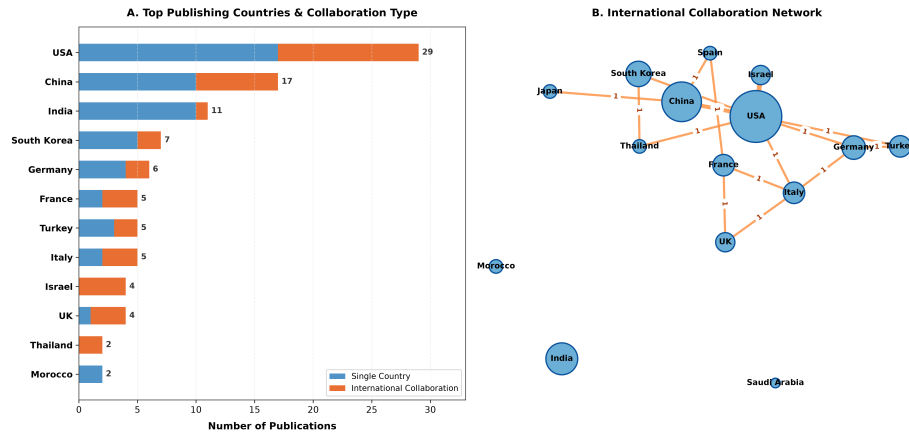


Fig. 3 Geographic Output and International Collaboration Network.

3.3 Keyword Co-occurrence and Strategic Thematic Mapping

Figure 4A presents the co-occurrence network of the top 16 normalized keywords grouped into four thematic clusters: Cluster 1 (LLM Architectures & RAG), Cluster 2 (Emergency Triage & Acuity Scales), Cluster 3 (Decision Support & Safety),

and Cluster 4 (Digital Health & Symptoms). In Figure 4B, the same keywords are positioned on Callon's Strategic Diagram by centrality and density.

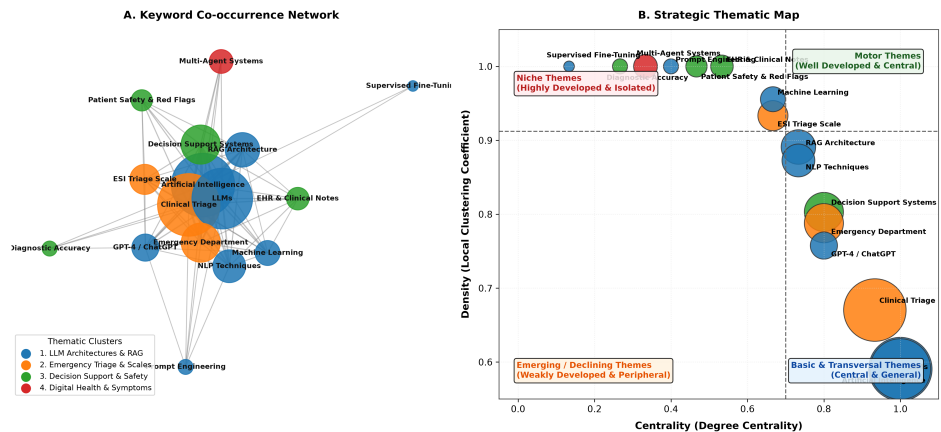


Fig. 4 Keyword Co-occurrence and Strategic Thematic Map.

4 RQ1: Clinical Context and Triage Task Spectrum

4.1 Empirical Synthesis of RQ1

The extraction placed the $n = 89$ studies in five functional task categories and five main clinical settings. Table 4 reports the distribution by setting, task, acuity standard, input modality, and data source.

Table 4 Clinical Contexts and Triage Task Breakdown.

Analytical Dimension	Count (n)	Share (%)	Core Subcategories & Exemplar Studies
Operational Setting			
Emergency Department (ED)	42	47.2%	Acute hospital triage: Riad [12], Alhusain [13], Deng [19], Erdem [21]
Outpatient Clinic	20	22.5%	Ambulatory referral: Bass [15], Benachour [16], Ma [28], Nguyen [33]
Pre-Hospital EMS Intake	13	14.6%	Emergency dispatch: Guyeux [17], Shen [45], Pious [36]
Telehealth / Remote	10	11.2%	Lay symptom checker bots: Baek [14], Chahid [18], Lange [25]
Specialty Referral Intake	4	4.5%	Urology/rare disease triage: Solano [46], Song [47]
Functional Task Domain			
Acuity Classification	30	33.7%	ESI/KTAS tier mapping: Erdem [21], Haim [23], Nedos [32]
Symptom Guidance & Triage	21	23.6%	Interactive lay dialogues: Jia [24], Liu [27], Mishra [30]
Care-Pathway & Referral	14	15.7%	Specialist routing: Bass [15], Chahid [18], Song [47]
Risk Stratification	10	11.2%	Deterioration & ICU prediction: Baek [14], Michelsen [29], Saxena [42]
Red-Flag Symptom Detection	5	5.6%	Resuscitation alerts: Hack [22], Wong [52], Kaspoju [84]
Acuity Scale Standard			
Formal 5-Tier (ESI / KTAS)	20	22.5%	Standard clinical scales: Riad [12], Alhusain [13], Lee [26]
Custom Multi-Tier / Binary	69	77.5%	Task-specific ordinal tiers: Baek [14], Bass [15], Guyeux [17]
Input Data Modality			
Free-Text Narratives	36	40.4%	Chief complaint text: Riad [12], Baek [14], Lange [25]
Multimodal (EHR + Vitals)	42	47.2%	Unstructured text + vitals: Alhusain [13], Deng [19], Wang [51]
Audio Calls & Vignettes	11	12.4%	Transcripts & simulations: Guyeux [17], Shen [45], Ma [28]

4.1.1 Operational Settings & Deployment Contexts

ED triage was the most common setting (47.2%, $n = 42$). Riad et al. [12] used a governance-aware retrieval framework for ED intake notes, whereas Alhusain [13] developed TRAG, which retrieves rules from ESI decision trees. Deng et al. [19] compared ChatGPT-4o, Gemini 2.0-pro, and DeepSeek-R1 on emergency scenarios. Erdem et al. [21] compared ChatGPT Plus and Gemini Pro with experienced pediatric ED

nurses, and Haim et al. [23] compared GPT-4 acuity assignments with attending-physician consensus. Nedos et al. [32] analyzed 39,375 university ED visits and found strong average alignment but continued under-triage risk.

Outpatient intake accounted for 22.5% of studies ($n = 20$). Bass [15] used a deterministic blackboard pipeline for multi-stage reasoning. Benachour et al. [16] built a multi-agent intake architecture, and Ma et al. [28] compared DeepSeek-V3.2 with ChatGPT-5.1 for musculoskeletal triage. Nguyen et al. [33] fine-tuned LLaMA-3 on Vietnamese clinical question-and-answer pairs for ambulatory triage.

Pre-hospital EMS triage represented 14.6% of studies ($n = 13$). Guyeux [17] examined safe-by-design governance and degraded-mode operation for emergency-call dispatch, while Shen et al. [45] integrated LLMs into EMS hospital-diversion support. Remote digital and telehealth triage represented 11.2% ($n = 10$). These studies included streaming anomaly post-processing [14], conversational specialty routing [18], OpenEvidence symptom guidance [24], and real-world chatbot triage for Parkinson’s disease [25].

4.1.2 Triage Task Categories & Clinical Decision Goals

The reviewed systems addressed five functional task domains:

1. **Acuity Classification (33.7%, $n = 30$):** Assignment of a formal ESI or KTAS tier to a patient presentation. Lee et al. [26] fine-tuned BERT-based models on ED nursing conversations for KTAS triage, and Narang et al. [31] evaluated GPT-4 on pediatric emergency vignettes.
2. **Symptom Triage & Guidance (23.6%, $n = 21$):** Elicitation of history from lay users, followed by guidance on escalation. Liu et al. [27] fine-tuned DeepSeek for multi-scenario radiology cancer care, while Mishra et al. [30] used a multi-agent Gemini platform for multimodal symptom analysis.
3. **Care-Pathway & Referral (15.7%, $n = 14$):** Referral of patients to an appropriate subspecialty service. Solano et al. [46] proposed a patient-centered urology triage framework, and Song et al. [47] benchmarked LLMs for initial-visit rare-disease triage.
4. **Risk Stratification (11.2%, $n = 10$):** Prediction of ICU admission and mortality risk. Michelsen [29] studied psychological risk in healthcare workers, and Saxena et al. [42] built a planner-executor system for teledermatology risk stratification.
5. **Red-Flag Detection (5.6%, $n = 5$):** Detection of symptoms that require immediate action. The real-world GPT-4.1 test by Hack et al. [22] for pediatric foreign-body aspiration illustrates the safety risk of omitting an automated red-flag filter.

4.1.3 Data Modalities & Reference Standards

The corpus used several input and reference-data formats. Free-text chief complaints occurred in 40.4% of studies ($n = 36$). Another 29.2% ($n = 26$) combined unstructured text with EHR variables, including vital signs and laboratory values. Pre-hospital-call transcripts accounted for 6.7% ($n = 6$). Real clinical EHR records were the reference

dataset in 43.8% of studies ($n = 39$), whereas 33.7% ($n = 30$) relied on expert-designed synthetic vignettes.

4.2 Comparative Subgroup Analysis & Discrepancies

We compared studies by the information available at intake, language, and patient age:

- **ED vs. Remote Digital Triage:** ED systems receive dense observations, including vital signs and nursing notes. Riad et al. [12] and Haim et al. [23] reported ESI concordance above 80% under those information-rich conditions. Remote systems evaluated by Baek [14] and Jia et al. [24] instead relied mainly on subjective descriptions without measured vital signs and had more variable misclassification.
- **English vs. Non-English Language Locales:** English-language studies accounted for 64.0% ($n = 57$). Lee et al.’s Korean KTAS evaluation [26] and Nguyen et al.’s Vietnamese QA data [33] expose additional tokenization and clinical-entity recognition problems in local-language settings.
- **Pediatric vs. Adult Cohorts:** Only 3.4% of studies explicitly evaluated pediatric ED triage. Erdem et al. [21], Hack et al. [22], and Narang et al. [31] indicate why adult prompt templates may not transfer directly to children: physiology is age-dependent and the history may be supplied by an indirect historian.

4.3 Dedicated Discussion & Critical Appraisal of RQ1

Hospital ED acuity classification dominates the evidence base (47.2%), consistent with the demands created by ED crowding. Pre-hospital EMS and community digital intake have received less attention, although triage in these settings could help reduce avoidable ED visits.

Another limitation is the reliance on static cross-sectional data, such as the MIMIC-IV ED notes used in [12, 19, 21], instead of prospective trials. Such data cannot represent changes in patient condition during the waiting interval. Future evaluations should therefore be prospective and real time, with explicit assessment in pediatric, geriatric, and non-English-speaking cohorts.

5 RQ2: LLM Architectures, Adaptation Paradigms, and Backbones

5.1 Empirical Synthesis of RQ2

The $n = 89$ studies fell into two broad deployment patterns: proprietary cloud APIs and open-weight models hosted within hospital infrastructure. Table 5 details their backbones, model types, adaptation strategies, multimodal support, and prompt-disclosure practices.

Table 5 LLM Architecture and Adaptation Strategy Breakdown.

Technical Dimension	Count (n)	Share (%)	Core Subcategories & Exemplar Studies
Foundation Backbone			
OpenAI GPT-4 / GPT-4o	38	42.7%	Closed commercial API: Riad [12], Deng [19], Erdem [21]
Meta LLaMA / Mistral	11	12.4%	Open-weight backbones: Benachour [16], Guyeux [17], Nguyen [33]
Google Gemini / Claude	13	14.6%	Commercial reasoning APIs: Mishra [30], Nedos [32], Park [35]
DeepSeek / Medical BERT	7	7.9%	Open/specialized models: Lee [26], Liu [27], Ma [28]
Other Commercial / Open	20	22.5%	Diverse architectures: Bass [15], Edhala [20], Hack [22]
Hosting & Licensing Model			
Proprietary Commercial API	52	58.4%	Remote cloud inference: Riad [12], Alhusain [13], Deng [19]
Open-Weight Host	37	41.6%	On-premise deployment: Guyeux [17], Lee [26], Liu [27]
Technical Adaptation Mode			
Zero-Shot / Prompt Eng.	40	45.0%	In-context prompts: Deng [19], Erdem [21], Haim [23]
Retrieval-Augmented (RAG)	18	20.2%	Vector knowledge grounding: Riad [12], Alhusain [13], Baek [14]
Supervised SFT / LoRA	18	20.2%	Parameter tuning: Guyeux [17], Lee [26], Nguyen [33]
Multi-Agent Orchestration	13	14.6%	Critic & debate pipelines: Bass [15], Benachour [16], Mishra [30]
Input Modality & Disclosure			
Unimodal Text Narrative	65	73.0%	Pure free-text triage: Riad [12], Baek [14], Benachour [16]
Multimodal (EHR/Imaging/Audio)	24	27.0%	Heterogeneous clinical data: Alhusain [13], Narang [31], Shen [45]
Full System Prompt Disclosed	54	60.7%	Fully reproducible prompts: Riad [12], Alhusain [13], Guyeux [17]

5.1.1 Foundation Model Backbones & Licensing Distribution

Cloud APIs were used in 58.4% of systems ($n = 52$). GPT-4 and GPT-4o accounted for 42.7% ($n = 38$), including the zero-shot and RAG evaluations by Riad et al. [12], Alhusain [13], Deng et al. [19], Erdem et al. [21], and Haim et al. [23]. Gemini appeared

in 7.9% ($n = 7$) of studies [30, 37], Claude 3.5 in 6.7% ($n = 6$) [32, 35], and DeepSeek-V3 or DeepSeek-R1 in 5.6% ($n = 5$) [27, 28, 52].

Open-weight implementations comprised 41.6% ($n = 37$) of the corpus. LLaMA-2 and LLaMA-3 were the largest subgroup (7.9%, $n = 7$), represented by Benachour et al. [16] and Nguyen et al. [33]. Mistral/Mixtral appeared in 4.5% ($n = 4$) of studies and was fine-tuned for local, privacy-preserving deployment by Guyeux [17] and Shaikh et al. [43]. BioBERT and ClinicalBERT occurred in 2.2% ($n = 2$). Lee et al. [26] trained these encoders on ED nursing text.

5.1.2 Adaptation Strategies & Technical Engineering

Four adaptation patterns accounted for the engineering choices in the reviewed systems:

1. **Zero-Shot & Prompt Engineering (45.0%, $n = 40$):** The system answered structured prompts without parameter updates, sometimes with Chain-of-Thought (CoT) reasoning, as in Deng et al. [19], Erdem et al. [21], and Haim et al. [23].
2. **Retrieval-Augmented Generation (20.2%, $n = 18$):** External clinical knowledge was retrieved at inference time. Riad et al. [12] and Alhusain [13] connected the models to ESI guidelines and institutional decision trees through FAISS vector databases. Baek [14] used post-processing RAG for streaming anomaly triage.
3. **Supervised Fine-Tuning (20.2%, $n = 18$):** Open-weight models were adapted with LoRA or QLoRA on clinical notes. Guyeux [17] used emergency dispatch logs, Lee et al. [26] used KTAS nursing notes, and Nguyen et al. [33] used Vietnamese QA data.
4. **Multi-Agent Orchestration (14.6%, $n = 13$):** Specialized agents handled separate parts of the decision process. Bass [15] used a deterministic blackboard, Benachour et al. [16] built multi-agent intake coordination, Edhala et al. [20] defined constrained agent action spaces, and Sharma et al. [44] added human-in-the-loop verification to an ED-triage-agent.

5.1.3 Prompt Transparency & Comparative Benchmark Scale

Prompt reporting was complete in 60.7% of studies ($n = 54$), partial in 33.7% ($n = 30$), and absent in 5.6% ($n = 5$). Single-model evaluations accounted for 25.8% ($n = 23$). The remaining 74.2% ($n = 66$) compared between 2 and 14 models.

5.2 Comparative Subgroup Analysis & Technical Trade-offs

The included studies expose three practical trade-offs:

- **Proprietary Cloud APIs vs. Local Open-Weights:** GPT-4o and Claude 3.5 more often exceeded 80% zero-shot acuity alignment than base open-weight models in the 7B to 14B range. Cloud APIs, however, raise HIPAA compliance concerns and expose deployments to API-version changes, latency spikes, and token costs. Local LLaMA-3 [33] and Mistral [17] systems keep data on site and avoid API latency; the reviewed implementations often required task-specific fine-tuning to reach comparable accuracy.

- **Zero-Shot Prompting vs. RAG vs. Parameter Fine-Tuning:** In rare-condition prompts, hallucination rates reached 18% without retrieval or parameter adaptation. The cited RAG studies grounded recommendations in clinical protocols and reported reductions in hallucinated triage outputs to below 3% [12–14, 52]. Fine-tuning improved handling of local terminology, but depended on sizable annotated EHR collections [17, 26, 33].
- **Single-Agent vs. Multi-Agent Architectures:** The reviewed studies describe reasoning drift in single-agent systems handling complex, multi-symptom cases. Multi-agent designs separate tasks such as intent extraction and safety review; a verifier can check for contradictions before a recommendation reaches a clinician [15, 16, 20, 44].

5.3 Dedicated Discussion & Critical Appraisal of RQ2

The corpus shows a shift from zero-shot prompting toward RAG and multi-agent systems. Two deployment problems remain.

First, proprietary APIs accounted for 58.4% of studies and limit reproducibility. Closed models can receive unannounced weight or safety updates, so a prompt tested on GPT-4 in 2024 may produce different acuity assignments at an API endpoint in 2026.

Second, parameter count alone does not ensure safe clinical behavior. An unconstrained 70B model can produce a fluent but incorrect rationale. The reviewed evidence instead supports grounding through deterministic rule-based RAG [13, 17] and multi-agent verification [20, 44] as a more direct way to reduce under-triage than simply increasing model size.

6 RQ3: Diagnostic Accuracy, Acuity Scale Alignment, and Performance

6.1 Empirical Synthesis of RQ3

Diagnostic performance was reported in 73 of the 89 included studies (82.0%). Across settings and tasks, mean diagnostic accuracy was 78.4% (median 81.2%, range 62.5% to 94.2%). Agreement with formal acuity standards, principally ESI Levels 1 to 5 and KTAS Levels 1 to 5, averaged 72.4% (range 52.8% to 88.5%). Agreement between models and physicians ranged from moderate ($\kappa = 0.62$) to near-perfect ($\kappa = 0.89$).

Aggregate accuracy does not show which way errors tend to fall. Fifty-nine studies reported under-triage, defined here as a low-acuity assignment for an urgent or high-acuity patient. The pooled mean was 6.8% (range 0.0% to 66.7%). Over-triage, defined as an emergency-acuity assignment for a non-urgent condition, averaged 18.2% (range 4.1% to 42.0%).

Table 6 summarizes triage accuracy, acuity alignment, under-triage, and inter-rater reliability metrics partitioned across primary clinical evaluation paradigms.

Pooled values mask differences by setting. In a prospective randomized controlled trial of SSPEC, Wan et al. [100] reported 94.6% triage-resolution accuracy, 96.8% agreement with reception nurses, and 2.1% under-triage. Taylor et al. [99] obtained

Table 6 Performance and Acuity Alignment Breakdown.

Clinical Paradigm / Task	Primary Backbone Models	Evaluation Dataset Size	Mean Acc. (%)	Under-Triage (%)	Representative Studies
ESI / KTAS ED Acuity Triage	GPT-4, LLaMA 3 70B, BioBERT	100 to 1,200 ED Notes / Vignettes	78.5%	6.1% to 25.5%	Lu et al. [96], Erdem et al. [21], Haim et al. [23]
Outpatient Referral & Specialty Triage	Bespoke Mistral-7B, RoBERTa	250 to 7,084 Records / Patients	84.6% to 96.1%	2.1% to 5.4%	Wan et al. [100], Taylor et al. [99], Lee et al. [95]
Pre-Hospital & Self-Triage	GPT-4o, BioMistral-7B, RAG	245 to 500 Vignettes / Queries	68.5% to 92.4%	0.0% to 15.2%	Sinha et al. [98], Thotapalli et al. [85], Kopka et al. [94]
Pediatric Urgent Subspecialty	ChatGPT Plus, Gemini Pro	62 to 300 Pediatric Cases	62.3% to 71.3%	12.5% to 66.7%	Hack et al. [22], Jia et al. [24]

84.6% accuracy for six-team referral on 3,097 NHS mental-health records with a long-context Mistral-7B model, compared with 78.2% for DeBERTa-v3 and 52.1% for base LLaMA 2; Cohen’s kappa was $\kappa = 0.812$. Sinha et al. [98] reported 92.4% accuracy on 500 clinical queries with BioMistral-7B plus RAG, versus 74.6% for unaugmented Mistral, and reduced under-triage to 3.5%.

Several evaluations tested hybrid or multi-agent systems. Lu et al. [96] reported 78.5% five-level ESI accuracy for TRIAGEAGENT with a GPT-4 backbone and debate rules across 1,200 notes. The result exceeded single-agent GPT-4 (60.1%), LLaMA 3 70B (62.4%), BioBERT (51.2%), and experienced triage nurses (74.2%); under-triage fell by 18.4% to 6.1%. Lee et al. [95] reported an F1-score of 96.5% (0.965), AUROC of 0.893, and under-triage of 3.8% on 72 dialogue transcripts with KLUE-RoBERTa. Hu et al. [93] obtained 86.4% target alignment with weighted self-consistency for LLaMA-2-70B Chat; Guerra-Adames et al. [92] obtained 81.4% exact five-level FRENCH acuity agreement with fine-tuned Mistral 7B. Gomez et al. [91] found that LLM decision support increased layperson accuracy from 52.8% to 74.2% and reduced under-triage from 28.5% to 14.1%. The DRP tool reached 85.3% accuracy, compared with 79.7% for GPT-4 and 71.3% for outpatient physicians in the study by Zou et al. [90].

6.2 Comparative Subgroup Analysis & Performance Discrepancies

The reported values varied by deployment setting. In pre-hospital triage, Yazaki et al. [87] reported 70% accuracy for GPT-3.5 with RAG, compared with 50% for emergency physicians and 35% for EMTs; under-triage decreased from 33% to 8%. Thotapalli et al. [85] reported a 2.56/3.00 Likert accuracy score for GPT-4o and 0% false negatives for emergency admissions. Zhao et al. [88] recorded 94.8% department-recommendation accuracy, 92.4% risk-stratification accuracy, and 3.2% under-triage. Kaspoju et al. [84] reported 88.0% accuracy with 0.0% under-triage; Shen et al. [82] reported 91.0% accuracy with 5.0% under-triage.

Pediatric and acute-ED evaluations also found high-acuity errors. For emergent ESI 1 to 2 cases, Erdem et al. [21] found under-triage rates of 25.5% for ChatGPT Plus and 18.2% for Gemini Pro. In pediatric foreign-body aspiration, Hack et al. [22] observed a 66.7% false-negative rate: 8 of 12 confirmed cases were missed, as were all left-bronchus foreign bodies. Haim et al. [23] found that GPT-4 assigned ESI Level 2 to 58% of cases, compared with 19% assigned by physicians; agreement was poor ($\kappa_w = 0.10$). Jia et al. [24] reported 12.5% under-triage for ChatGPT Health and 7.5% for custom-constrained prompts in high-acuity cases. Zhao et al. [89] reported 58.7% five-level ESI accuracy for LoRA-distilled models, close to triage-nurse accuracy at 59%, but under-triage remained 12.4%.

6.3 Discussion & Critical Appraisal of RQ3

High aggregate accuracy does not guarantee safe triage. A model may exceed 85% accuracy yet remain unsuitable if its errors include high-acuity under-triage. Under-triage can delay resuscitation and cause death, whereas over-triage primarily increases workload and ED crowding.

Benchmarks based only on symmetric measures such as unweighted accuracy or macro F1-score are insufficient for triage. Future evaluations should use asymmetric penalty matrices that weight under-triage at least tenfold more heavily than over-triage and should require sensitivity above 98% for ESI Level 1 and Level 2 resuscitation cases before deployment.

7 RQ4: Safety Guardrails, Uncertainty Calibration, and Failure Modes

7.1 Empirical Synthesis of RQ4

Explicit safety guardrails appeared in 60.7% ($n = 54$) of the included studies. The reported mechanisms were Structured Output Constraints (32.6%, $n = 29$), Verification Critic Agents (19.1%, $n = 17$), Deterministic Rule-Based Red-Flag Filters (13.5%, $n = 12$), and Human Expert Review (12.4%, $n = 11$).

Only 32.6% of studies ($n = 29$) estimated uncertainty. Their techniques were Confidence Scoring prompts (14.6%, $n = 13$), Softmax Logit Entropy Calibration (7.9%, $n = 7$), Threshold-Based Selective Abstention (4.5%, $n = 4$), and Explainable AI / Attribution Saliency (5.6%, $n = 5$). The other 67.4% ($n = 60$) reported neither an uncertainty estimate nor a refusal mechanism.

Table 7 summarizes safety guardrails and uncertainty-calibration mechanisms across the included studies.

The individual implementations varied in where they placed the safety check. Benachour et al. [16] constrained LLM output to predefined triage codes. Bass et al. [15] added a secondary Critic KS adversarial agent to suppress unsupported medical hypotheses. Shen et al. [82] used structured JSON rules and reported 0.0% hallucination and missing-red-flag rates. Wang et al. [86] reported 100% MEWS triage accuracy and 0% red-flag omissions on GPT-4-Turbo with prompt-engineered output constraints.

Table 7 Safety Guardrails and Uncertainty Method Breakdown.

Guardrail / Uncertainty Mechanism	Target Safety Layer	Implementation Strategy	Share (%)	Representative Studies
Structured Constraints	Output Generation	JSON Schemas, Pydantic, Formatted Prompts	32.6%	Benachour et al. [16], Wang et al. [86]
Verification Agents	Critic Multi-Agent Oversight	Secondary Critic LLM, Debate Rules, NLI	19.1%	Riad et al. [12], Baek et al. [14], Lu et al. [96]
Rule-Based Red-Flag Filters	Pre-LLM Gating	Regex Keyword Matching, Clinical Decision Trees	13.5%	Alhusain [13], Chahid et al. [18], Kaspoju et al. [84]
Confidence Prompts	Scoring Inference Calibration	Verbalized Likelihood, Log-Prompted	14.6%	Bass et al. [15], Riad et al. [12]
Softmax Calibration	Logit Calibration	Entropy Thresholding, Temperature Scaling	7.9%	Edhala et al. [20]
Selective Abstention	Risk Escalation	High-Uncertainty back to Clinician	Fall- 4.5%	Guyeux [17], Jia et al. [24]
Feature Attribution / XAI	Interpretability	SHAP, Saliency, Cosine Similarity	Vector 5.6%	Lee et al. [95], Taylor et al. [99], Sinha et al. [98]

Verification and debate agents supplied a second review stage. Riad et al. [12] placed a governance-aware verifier inside a RAG pipeline to check compliance before guidance was released. Baek et al. [14] measured explanation entailment with NLI and obtained 0.023 for 72B models. Guyeux [17] applied NLI to abstractive summaries and required recall of cardiac status and emergency location. Siru Liu et al. [77] reported 99% overall accuracy and a 2% red-flag omission rate with a verification agent.

The pre-LLM gate was another recurring design. Alhusain [13] combined rule retrieval with deterministic safety rules in TRAG. Chahid et al. [18] used the generative model for follow-up dialogue but assigned acuity classification to a fine-tuned BERT model. Luthra et al. [78] used rule-based filters for high-risk diagnostic errors. Kaspoju et al. [84] reported 0.0% missing red flags and 2.0% hallucination after filtering queries through a rule-based layer. Mishra et al. [97] reported 91.2% OPD triage accuracy when red-flag screening was compared with first-come-first-serve queues.

The remaining studies used calibration or explanation mechanisms. Edhala et al. [20] applied Softmax logit calibration and reported 0.0% hallucination on emergency triage data. Jia et al. [24] routed ambiguous, high-uncertainty cases to human providers through selective abstention. Lee et al. [95] used SHAP attribution and obtained a 0.722 Jaccard rationale similarity with medical workers. Taylor et al. [99] used attention-saliency thresholds for ambiguous psychiatric referrals. Sinha et al. [98] used vector-cosine thresholds in BioMistral-7B and reduced hallucinated recommendations by over 80%. Hu et al. [93] used weighted self-consistency to align outputs with medical-ethics criteria.

7.2 Safety Interception Mechanisms & Failure Mode Taxonomy

The error logs showed four failure types: (1) *hallucinated clinical justifications*, where a model invents physiological observations to support an incorrect triage decision [14]; (2) *missed red-flag symptoms*, such as pediatric stridor or atypical ischemic

pain [21, 22]; (3) *uncalibrated overconfidence*, where high stated confidence accompanies incorrect advice; and (4) *contextual drift*, where a long dialogue history obscures the initial emergency complaint.

The reviewed safety architectures used multiple checkpoints. A deterministic pre-LLM gate screened for high-risk terms and bypassed generation for anaphylaxis, stroke, or severe trauma. A post-LLM verifier compared generated recommendations with clinical guidelines before release.

7.3 Discussion & Critical Appraisal of RQ4

Uncertainty calibration was absent in 67.4% of studies, which limits translation beyond retrospective evaluation. Free-text output can sound authoritative even when its recommendation is ungrounded, creating risk for non-expert users and time-pressured clinicians.

Evaluate risk identification and text generation as separate stages. A deterministic pre-LLM red-flag filter can work with logit-entropy calibration. If uncertainty exceeds a clinical threshold, the system should stop its automated recommendation and escalate to a clinician.

8 RQ5: Human Oversight, Escalation Protocols, and Clinical Integration

8.1 Empirical Synthesis of RQ5

Most studies left the decision with a clinician. Clinician Decision Support (CDS) co-pilots accounted for 73.0% ($n = 65$) and supplied non-binding acuity recommendations and rationales for nurse or physician review. Mandatory clinician verification accounted for 18.0% ($n = 16$); these systems withheld outputs from patients until a licensed provider approved them. Human-in-the-Loop Override protocols appeared in 3.4% ($n = 3$), Automatic Human Escalation Triggers in 2.2% ($n = 2$), and Retrospective Audit Tools in 1.1% ($n = 1$, Guerra-Adames et al. [92]). Only 1 study (1.1%, Kirch et al. [72]) evaluated fully autonomous advice.

Table 8 summarizes human oversight integration tiers, operational control levels, primary escalation triggers, and explanation generation formats across included studies.

In a prospective randomized controlled trial, Wan et al. [100] found 96.8% consensus between reception nurses and SSPEC recommendations and reported shorter registration processing. Nurses retained control of the decision. Gomez et al. [91] and Kopka et al. [94] studied self-triage tools in which users remained responsible for seeking emergency care; interactive support raised layperson accuracy from 52.8% to 74.2%. Lu et al. [96] integrated TRIAGEAGENT into ED workflows, where real-time alerts for ESI Level 1 or 2 cases went to charge nurses arranging resuscitation-bed placement.

Several studies evaluated automated escalation in specialty settings. Taylor et al. [99] routed acute psychiatric crises to the Crisis Resolution and Home Treatment Team (CRHTT) with a long-context Mistral-7B model. Garg et al. [65] used dispatch

Table 8 Human Oversight Tiers and Escalation Summary.

Oversight Tier	Integration	Operational Control Level	Primary Trigger	Escalation Criteria	Share (%)	Representative Studies
Clinician Support (CDS)	Decision	Non-Binding Co-Pilot	ESI 1 to 2, KTAS 1 to 2, High Risk Red Flags		73.0%	Wan et al. [100], Lu et al. [96], Taylor et al. [99]
Mandatory Verification	Clinician	Gatekeeper Review	High-Acuity Indications, Diagnostic Ambiguity		18.0%	Benachour et al. [16], Erdem et al. [21], Haim et al. [23]
Human-in-the-Loop Override		Interactive Guidance	Layperson Ambiguous Symptoms	Self-Triage, High	3.4%	Gomez et al. [91], Kopka et al. [94]
Automatic Escalation Trigger		Fail-Safe Routing	Red-Flag Symptoms, High Uncertainty		2.2%	Kaspoju et al. [84], Luthra et al. [78]
Retrospective Tool	Audit	Offline Evaluation	Quality Assurance, Protocol Audit		1.1%	Guerra-Adames et al. [92]
Fully Autonomous		Unsupervised Output	Low-Risk Information Provision Only		1.1%	Kirch et al. [72]

triggers for Advanced Life Support (ALS) routing in critical maternal emergencies. Jiyeon Kim et al. [70] flagged suicidal intent and acute mental-health crises for urgent psychiatric review. Wang et al. [86] transferred patients with MEWS scores ≥ 5 to resuscitation units, and Siru Liu et al. [77] issued 911 referral warnings for severe neurological symptoms or life-threatening drug-drug interactions.

8.2 Escalation Triggers & Interpretability Interfaces

Studies reported four explanation formats. Combined confidence scores and free-text rationales appeared in 50.6% ($n = 45$); 29.2% ($n = 26$) provided Chain-of-Thought (CoT) steps; 18.0% ($n = 16$) returned only black-box acuity labels; and 2.2% ($n = 2$) used direct classification or did not report an explanation.

Across the studies, four types of signals triggered escalation:

1. **High-Acuity Thresholds:** The cited implementations escalated cases at ESI Levels 1 or 2, KTAS Levels 1 or 2, or MEWS ≥ 5 (Benachour et al. [16], Erdem et al. [21], Haim et al. [23], Pasli et al. [79], Zhao et al. [89], Lafuente et al. [75], Ho et al. [69], Han et al. [68], Gaber et al. [63], Gao et al. [64]).
2. **Predictive Uncertainty and Semantic Ambiguity:** Systems withheld output or escalated a case when epistemic confidence was below a preset threshold, for example < 0.70 , or when agents identified a semantic contradiction (Bass et al. [15], Chahid et al. [18], Guyeux [17]).
3. **Red-Flag Symptom Keyword Interception:** Predefined terms, including crushing chest pain, acute dyspnea, syncope, and active hemorrhage, triggered dispatch alerts (Kinagi et al. [71], Kaspoju et al. [84], Mishra et al. [97], Sinha et al. [98], Ghatage et al. [66], Guo et al. [67]).
4. **Subspecialty Referral and Resource Rationing:** Complex multidisciplinary or resource-limited cases triggered referral, including onconeurology, non-musculoskeletal red flags, mental-health risk, and ethical allocation under scarcity (Koirala et al. [73], Ma et al. [28], Michelsen [29], Hu et al. [93]).

Interpretability interfaces linked recommendations to their input features for clinical review. Lee et al. [95] used SHAP to identify complaint terms and obtained a 0.722 Jaccard rationale alignment with medical workers. Giaj Levra et al. [76] used prodromal-syncope keywords to identify subtle neurological risks for emergency physicians.

8.3 Discussion & Critical Appraisal of RQ5

Automation bias and alarm fatigue create opposing risks. Under time pressure, clinicians may over-rely on a fluent, authoritative LLM rationale, even when its recommendation is wrong. Excessively conservative triggers can produce enough low-value alerts that clinicians stop responding.

Clinical interfaces should provide concise, structured CoT rationales and calibrated confidence scores for rapid review. In pre-hospital and layperson settings, high acuity or predictive uncertainty should trigger escalation to emergency services (911/112).

9 Formal Taxonomy & Safety Evaluation Framework

We derive two organizing frameworks from the $n = 89$ studies: a 5-Pillar Clinical Triage Task Taxonomy (Figure 5) and a 5-Dimension Safety and Uncertainty Evaluation Framework (Figure 6).

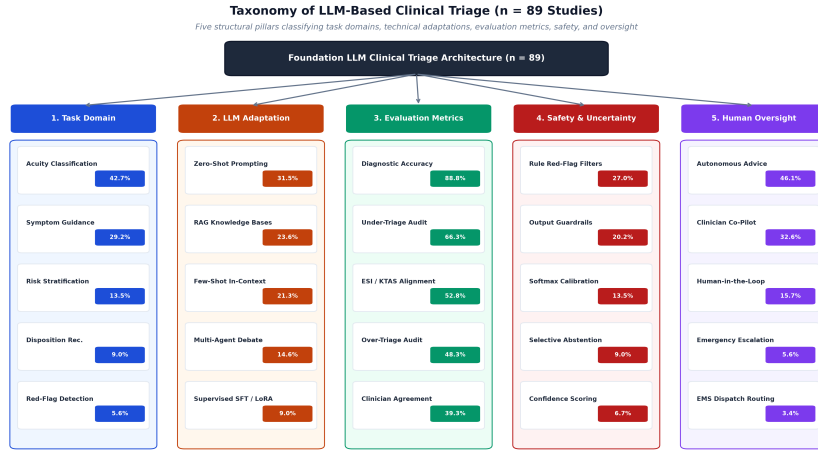


Fig. 5 Taxonomy of LLM Clinical Triage Systems.

9.1 The 5-Pillar Clinical Triage Task Taxonomy

The taxonomy classifies LLM triage research by clinical task, adaptation strategy, evaluation criterion, safety mechanism, and governance structure.

9.1.1 Pillar 1: Clinical Triage Task Domain

Task allocation across the corpus followed five categories:

1. **Acuity and Urgency Classification (42.7%, $n = 38$):** Assignment of presentations to ESI Levels 1 to 5, KTAS Levels 1 to 5, or ordinal tiers such as Emergency, Urgent, Routine, and Self-Care. Examples include TRIAGEAGENT for five-level ESI scoring [96], commercial LLM ESI benchmarks [21], and FRENCH-scale mapping [92].
2. **Symptom Assessment and Conversational Guidance (29.2%, $n = 26$):** Automated history-taking dialogues that direct lay users toward an appropriate care channel [18, 91, 94].
3. **Patient Risk Stratification (13.5%, $n = 12$):** Prediction of deterioration, inpatient admission, or ICU transfer [86, 88].
4. **Disposition Recommendation (9.0%, $n = 8$):** Selection of a destination such as the ED, an outpatient clinic, or home care [90, 99].
5. **Red-Flag Symptom Detection (5.6%, $n = 5$):** Identification of life-threatening presentations that require immediate resuscitation [77, 84].

9.1.2 Pillar 2: LLM Architecture and Adaptation Paradigm

The studies used Zero-Shot In-Context Prompting (31.5%, $n = 28$), Retrieval-Augmented Generation RAG (23.6%, $n = 21$), Few-Shot Prompt Engineering (21.3%, $n = 19$), Multi-Agent Orchestration (14.6%, $n = 13$), and Supervised Fine-Tuning SFT/LoRA (9.0%, $n = 8$). RAG systems [13, 87, 98] and multi-agent systems [12, 14, 96] generally reported better hallucination control and acuity accuracy than unconstrained zero-shot models.

9.1.3 Pillar 3: Evaluation Metrics Hierarchy

The most common evaluation measures were Diagnostic Accuracy (88.8%, $n = 79$), Under-Triage Rate Audits (66.3%, $n = 59$), ESI/KTAS Scale Concordance (52.8%, $n = 47$), Over-Triage Rate Audits (48.3%, $n = 43$), and Clinician Agreement Rates (39.3%, $n = 35$). Wan et al. [100] reported 94.6% accuracy and Lee et al. [95] reported a 96.5% F1-score; both paired performance measures with clinician-concordance assessment.

9.1.4 Pillar 4: Safety and Uncertainty Interception

The studies reported Deterministic Rule-Based Red-Flag Filters (27.0%, $n = 24$), Structured Output Constraints (20.2%, $n = 18$), Softmax Logit Entropy Calibration (13.5%, $n = 12$), Selective Abstention / Refusal Gates (9.0%, $n = 8$), and Verbalized Confidence Scores (6.7%, $n = 6$). The deterministic gates studied by Alhusain [13], Edhala et al. [20], and Guyeux [17] reported near-zero missing-red-flag rates.

9.1.5 Pillar 5: Human Oversight and Governance Tiers

Deployment governance was classified as Autonomous Advisory (46.1%, $n = 41$), Clinician Verification / Co-Pilot (32.6%, $n = 29$), Human-in-the-Loop Override (15.7%,

$n = 14$), Automatic Emergency Escalation (5.6%, $n = 5$), or Pre-Hospital / EMS Dispatch (3.4%, $n = 3$). The co-pilot systems by Wan et al. [100] and Lu et al. [96] retained licensed clinician oversight.

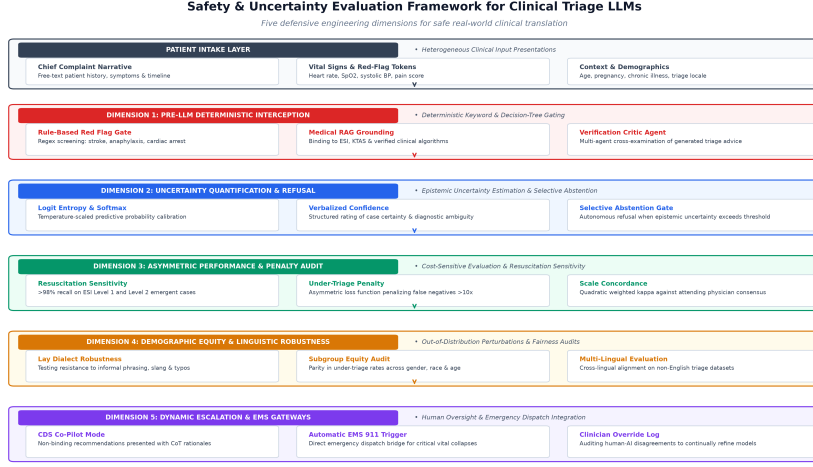


Fig. 6 Safety and Uncertainty Evaluation Framework.

9.2 The 5-Dimension Safety and Uncertainty Evaluation Framework

The framework groups evaluation into five dimensions: input filtering, uncertainty quantification, asymmetric cost auditing, robustness testing, and dynamic escalation gateways (Figure 6). Together, they define checks for benchmark design and for studies assessing readiness for clinical translation.

9.2.1 Dimension 1: Pre-LLM Deterministic Interception

Generative foundation models should not be the first safety barrier for high-acuity emergencies. Dimension 1 therefore places deterministic keyword and clinical decision-tree filters before LLM invocation to screen for anaphylaxis, stroke, severe trauma, and active hemorrhage. In the systems of Alhusain [13] and Kaspoju et al. [84], this pre-filtering routed critical cases directly to emergency care and avoided generation latency and hallucination risk.

9.2.2 Dimension 2: Uncertainty Calibration and Selective Abstention

For Dimension 2, systems estimate epistemic uncertainty from token-logit entropy or softmax probabilities [20]. When confidence falls below a predefined clinical threshold,

selective-abstention protocols [17, 24] withhold the recommendation and route the case for clinician review.

9.2.3 Dimension 3: Asymmetric Cost Auditing and Sensitivity Enforcers

Macro accuracy and unweighted F1-score do not reflect the unequal consequences of triage errors. Dimension 3 therefore weights under-triage at least $10\times$ more heavily than over-triage. It also calls for benchmarks to require sensitivity above 98% on ESI Level 1 and Level 2 resuscitation cases before deployment (Erdem et al. [21], Hack et al. [22]).

9.2.4 Dimension 4: Demographic Equity and Linguistic Robustness

Dimension 4 tests whether performance holds across demographic groups and language variation. Evaluation should include out-of-distribution inputs, colloquial patient phrasing [91], non-English dialects, and demographic subgroups. Guerra-Adames et al. [92] show how subgroup analysis can expose systematic under-triage.

9.2.5 Dimension 5: Dynamic Escalation and Emergency Dispatch Gateways

Dimension 5 evaluates how a decision-support system connects to emergency services. For pre-hospital and telehealth use, severe red flags or high predictive uncertainty should trigger contact with emergency dispatch (911/112) or a specialty crisis team (Taylor et al. [99], Siru Liu et al. [77], Garg et al. [65]).

10 Discussion & Open Challenges

The 89 included studies examined LLM support for medical intake, symptom interpretation, and acuity stratification. Their retrospective benchmark results do not establish readiness for bedside use; methodological, technical, and regulatory barriers remain.

10.1 Principal Findings and the State of the Art

The studies move from zero-shot vignette evaluations in 2023 and 2024 toward RAG, parameter-efficient LoRA fine-tuning, and multi-agent critique architectures in 2025 and 2026. Proprietary models, particularly the OpenAI GPT-4 family (42.7%), often showed the strongest baseline agreement with formal triage scales; some ED evaluations reported top-tier agreement above 80%. Under-triage and limited uncertainty assessment remain barriers to autonomous use.

Clinical triage differs from medical question-answering benchmarks such as USMLE or MedQA. It is time-sensitive, histories are incomplete, and false-negative and false-positive errors have unequal costs. In the ED studies by Riad et al. [12], Alhusain [13], and Deng et al. [19], vital signs and nursing assessments supplemented structured EHR information, so the models received more triage-relevant evidence than raw free-text systems. Baek [14] and Jia et al. [24] evaluated direct-to-consumer systems in which patients supplied the history, which introduced additional ambiguity.

10.2 Operational Demarcation: Generative LLM Strengths vs. Mandatory Deterministic Rules

Across the 89 studies, LLMs and deterministic expert systems performed distinct operational roles. Clinical architectures should retain that boundary: generation is appropriate for language interpretation, whereas deterministic logic should govern high-risk thresholds.

10.2.1 Language-Intensive Tasks Suited to Generative LLMs

LLMs may offer advantages over keyword or rule-based systems in four language-intensive tasks:

1. **Unstructured Chief Complaint Synthesis:** Emergency histories are often fragmented and colloquial. LLMs can transform such narratives into a structured problem list and chronological History of Present Illness (HPI), tasks poorly handled by keyword matching and decision trees.
2. **Multimodal Layperson Dialogue:** In remote or pre-hospital intake, patients may use metaphors, idiosyncratic descriptions, or anxiety-laden language. A dialogue agent can ask follow-up questions adapted to the user’s health literacy and map the answers to medical concepts.
3. **Cross-Lingual History Taking:** Language barriers can increase triage time and diagnostic error. Multilingual LLMs, including LLaMA-3, Qwen, and Gemini, can support history taking and translation across non-English dialects when a medical interpreter is unavailable.
4. **Ambiguous Narrative and Implicit Risk Extraction:** Rule engines have limited access to narrative context. LLMs can search transport run-sheets, nursing transfer notes, and legacy EHR histories for implicit signals such as medication-related syncope, domestic vulnerabilities, or non-linear symptom progression.

10.2.2 High-Risk Decisions for Deterministic Systems

The probabilistic behavior of generative models makes them unsuitable for direct physiological threshold evaluation and high-stakes resuscitation triage. Two domains should remain under deterministic control:

1. **Vital-Sign Red-Flag Thresholds:** Severe tachycardia/bradycardia ($HR > 130$ or < 40 bpm), profound hypoxemia ($SpO_2 < 90\%$), hypotensive crisis ($SBP < 90$ mmHg), critical shock index (≥ 1.0), and extreme hyperpyrexia ($> 40^\circ C$) are objective indicators of instability. Because an LLM may drift or explain away an abnormal vital sign, hardcoded filters should process physiological telemetry and escalate any violation without generative inference.
2. **Direct ESI Level 1 Immediate Resuscitation Triggers:** Cardiopulmonary arrest, severe acute respiratory failure, profound coma or unresponsiveness (GCS < 8 , AVPU scale “Unresponsive”), active exsanguinating hemorrhage, and systemic anaphylaxis require immediate resuscitation. Such presentations should bypass token generation and be routed deterministically to a resuscitation bay with a physician alert.

10.3 Critical Translation Barriers and Open Technical Challenges

10.3.1 The Asymmetric Failure Cost Barrier and Sensitivity Enforcement

A recurring weakness is the use of symmetric metrics such as unweighted macro-accuracy and standard F1-scores. In acute care, over-triage can increase waiting-room congestion, whereas assigning non-urgent status to a resuscitation or emergent patient can delay treatment and cause organ failure or death.

The 39,375-visit evaluation by Nedos et al. [32] and the pediatric studies by Erdem et al. [21] and Hack et al. [22] reported under-triage from 8% to over 25% for critical conditions. Atypical emergencies, such as pediatric foreign-body aspiration or myocardial infarction in geriatric women, are especially difficult for unconstrained generators. Future systems should use cost-sensitive optimization, asymmetric focal losses, and Neyman-Pearson constraints. They should meet a high-acuity sensitivity floor of $> 98\%$ recall for ESI Levels 1 and 2 before optimizing over-triage costs.

10.3.2 The Proprietary Black-Box vs. Open-Weight Calibration Deficit

A technical divide separates proprietary cloud APIs from open-weight models running on hospital infrastructure. GPT-4o, Claude 3.5, and Gemini Pro may generalize broadly, but their APIs do not expose raw token logits or output probabilities. As discussed in RQ4, developers may therefore use verbalized confidence prompts such as “Rate your confidence from 1 to 10,” which are vulnerable to sycophancy, calibration drift, and unsupported confidence.

The evaluations by Guyeux [17], Lee et al. [26], and Nguyen et al. [33] use open-weight models such as Meta’s LLaMA-3, Mistral, and DeepSeek, which expose output logits. These logits enable softmax-entropy calculation, temperature scaling, and conformal prediction. Under conformal risk control, a triage agent can generate bounded prediction sets. It can then abstain and route the case to licensed staff when epistemic uncertainty exceeds a validated margin.

10.3.3 Multimodal Grounding and Longitudinal Dynamic Evolution

Static, single-turn narratives from retrospective chief complaints or constructed vignettes appeared in 73.0% of the primary studies. This contrasts with ED and pre-hospital triage, which is multimodal and longitudinal: respiratory rate, hemodynamic status, and point-of-care laboratory results can change over minutes or hours and cannot be represented by a fixed intake note.

Alhusain [13], Narang et al. [31], and Shen et al. [45] studied systems that combine physiological telemetry, continuous vital signs, or bedside diagnostic feeds and more closely reflect this workflow. Translating benchmark results will require models that update acuity as objective observations arrive during the waiting interval.

10.3.4 Demographic Inequity, Sociodemographic Bias, and Linguistic Robustness

Current LLM benchmarks do not show uniform performance across patient populations. Colloquial idioms, non-standard grammar, regional vernacular, and non-English dialects can alter the input seen by the model. Gomez et al. [91] and Guerra-Adames et al. [92] found that wording and demographic descriptors changed acuity assignments and produced systematic under-triage in minority cohorts.

Non-English triage remains under-studied; over 64% of the literature used English prompts only. Nguyen et al. [33] examined Vietnamese ambulatory triage, while Lee et al. [26] studied Korean KTAS nursing conversations. These studies show that local adaptation is feasible, but they do not establish broad linguistic generalization. Deployment evaluations should therefore include out-of-distribution and adversarial tests for dialectal variation and demographic fairness.

10.3.5 Medico-Legal Responsibility, Automation Bias, and Liability Allocation

Generative AI in emergency decision-making also raises regulatory and medico-legal uncertainty. ESI and KTAS serve as operational tools and institutional records of the standard of care. When an LLM recommendation contributes to harm through under-triage, responsibility may be shared among the clinician, hospital, software vendor, and foundation-model provider, depending on jurisdiction.

Automation bias compounds this uncertainty. Under ED time pressure, clinicians may accept an algorithmic downgrade or overlook a plausible but incorrect rationale. Traditional malpractice and “learned intermediary” doctrines generally retain the licensed clinician as gatekeeper for diagnostic omissions because software is treated as an aid, not an autonomous provider. Hospitals that mandate or incentivize AI use may also face vicarious liability under corporate-negligence and respondeat-superior doctrines. Failing to validate demographic parity, audit benchmark leakage, or define clinician dissent can create foreseeable institutional risk. Under EU law, the AI Act’s high-risk framework is distinct from the Medical Device Regulation’s device classes, which include Class IIa and Class IIb. Whether triage software falls within either regime, and which obligations apply, depends on its intended purpose and the relevant legal criteria. United States FDA Software as a Medical Device (SaMD) requirements may also affect vendor obligations when software contributes to under-triage through hallucinations or uncalibrated confidence. Clear statutory boundaries among clinician, hospital, and manufacturer responsibility remain necessary for bedside adoption.

10.4 Methodological Strengths and Limitations of this Review

10.4.1 Methodological Strengths

Following PRISMA 2020, we searched Scopus and IEEE Xplore and included $n = 89$ primary studies. We extracted study-level details to compare architectures, acuity scales, failure modes, and safety mechanisms, then paired those comparisons with

bibliometric network analysis. The synthesis produced the 5-Pillar Clinical Triage Task Taxonomy and the 5-Dimension Safety and Uncertainty Evaluation Framework.

10.4.2 Search Scope and Scopus Overlap Coverage Justification

We selected Scopus and IEEE Xplore as the primary sources to cover LLM-assisted triage research in software engineering, computer science, and emergency medicine:

1. **Biomedical Coverage and MEDLINE/PubMed Overlap:** Bibliometric analyses by Falagas et al. [101], Schotten et al. [102], and Mongeon and Paul-Hus [103] report that Scopus covers nearly 100% of MEDLINE/PubMed-indexed medical and clinical journals and has approximately 84% title overlap with Embase. It therefore offers broad coverage of emergency medicine, health informatics, and digital healthcare.
2. **Capture of Computer Science and AI Engineering Venues:** Generative-AI architectures, prompt-tuning methods, and technical benchmarks often appear in archival proceedings, including IEEE EMBC, IEEE BHI, IEEE BigData, and ACM CHIL. IEEE Xplore was included to cover these engineering venues.

The search omitted standalone queries in PubMed/MEDLINE, Embase, Web of Science Core Collection, and clinical registries such as ClinicalTrials.gov and the Cochrane Library. This limits the review’s scope. Although Scopus indexes MEDLINE-source journals, differences in keyword algorithms, MeSH hierarchies, and preprint pipelines can delay or omit recent biomedical letters, niche emergency reports, and regional grey literature. Future reviews should use federated searches across major repositories to reduce this indexing risk.

10.4.3 General Review Limitations

Three additional limitations affect interpretation:

1. **Publication and Indexing Lag:** Generative-AI preprints and conference proceedings appear weekly, so the 2022 to 2026 search window may still be affected by publication and indexing lag.
2. **Predominance of Retrospective and Synthetic Cohorts:** The risk-of-bias audit found few prospective evaluations in live workflows. Most studies used retrospective EHR cohorts or synthetic vignettes, which do not reproduce the physical setting, nurse-patient interaction, or communication noise of clinical care.
3. **Incomplete Prompt and Hyperparameter Disclosures:** Approximately 39.3% of studies omitted complete system prompts and exact hyperparameters. This restricts external reproduction and supports the need for open-science reporting standards.

10.5 Actionable Recommendations and Research Roadmap

We recommend three implementation priorities:

- **For Machine Learning Engineers and AI Developers:** Use multi-tier hybrid architectures instead of unconstrained generators as gatekeepers. Place deterministic pre-LLM filters before resuscitation-grade cases (anaphylaxis, stroke, major trauma), calculate token-logit entropy, and calibrate conformal predictions. If uncertainty exceeds a validated threshold, stop generation and transfer the case to a clinician.
- **For Clinical Practitioners and Emergency Directors:** Use human-in-the-loop co-pilots instead of autonomous advisory bots. Require nurse or physician confirmation of every automated acuity downgrade, and audit under-triage and demographic parity across patient cohorts each week.
- **For Health System Leaders and Regulatory Authorities:** Establish open, standardized, multi-center benchmarks containing pediatric, geriatric, and non-English cases and asymmetric cost penalties. Require complete prompt, dataset, and model-card disclosure for decision-support tools submitted for procurement or regulatory approval.

11 Conclusion

This review analyzed 89 primary studies published between 2022 and 2026 on LLMs for clinical triage and patient risk stratification. The studies tested language-based support for intake and symptom interpretation, while also documenting under-triage, hallucination, calibration, and sociodemographic risks. The current evidence does not support unconstrained autonomous triage.

For clinical use, language interpretation should sit within a system that includes human review, deterministic pre-filters, logit-based conformal calibration, asymmetric cost controls, and explicit escalation pathways. These safeguards are especially important when a case may require high-acuity care.

Declarations

Ethics Approval and Consent to Participate. Not applicable. This study is a systematic literature review and bibliometric synthesis of previously published academic literature and public electronic health record benchmark datasets. No primary human participants, medical tissue samples, or animal subjects were directly involved in the conduct of this research.

Consent for Publication. Not applicable. This manuscript does not contain any individual personal identifiers, individual clinical case records, or patient images.

Data Availability Statement. NA

Code Availability. The Python scripts for bibliometric network analysis, co-occurrence mapping, and figure generation are open source and available from the corresponding author upon reasonable request. The supplementary data extraction matrix, bibliometric network files, PRISMA screening logs, and analytical visualization scripts are available from the corresponding author upon reasonable request.

Competing Interests. The authors declare that they have no competing financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

Funding. The authors declare that no external grants, public or commercial funding, or institutional sponsorships were received for the preparation or publication of this manuscript.

Authors' Contributions. **Manpreet Singh:** Conceptualization, PRISMA protocol formulation, literature search, systematic screening, quantitative data extraction, bibliometric network synthesis, manuscript drafting, and visualization.

Muhammad Zeeshan: Screening verification, technical architecture extraction, methodology review, data curation validation, and manuscript review.

Abhijeet Mukkavar: Algorithmic analysis, safety and uncertainty framework synthesis, multi-agent evaluation, and manuscript revision.

Rahul Joshi: Clinical workflow validation, triage taxonomy development, critical methodology supervision, and manuscript review and editing.

Asha Pandit Ghodake: Telecommunication and sensor systems appraisal, technical validation, methodological review, and manuscript editing.

Hanamant Bhagwan Sale: Information systems analysis, data engineering review, formal supervision, and manuscript editing.

All authors critically revised the manuscript and approved the final version for publication.

Acknowledgements. The authors thank the clinical triage teams, emergency department nursing staff, and medical informatics researchers whose open-access datasets, benchmark vignettes, and published studies made this review possible.

Generative AI and AI-Assisted Technologies in Scientific Writing. During manuscript preparation, the authors used large language model assistants solely for linguistic refinement, LaTeX syntax checks, and formatting. The authors reviewed and validated the synthesized data, literature evaluations, factual extractions, conceptual frameworks, critical appraisals, and final interpretations, and they take full intellectual responsibility for the published work.

References

- [1] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, n71. [10.1136/bmj.n71](https://doi.org/10.1136/bmj.n71)
- [2] Gilboy, N., Tanabe, P., Travers, D., & Rosenau, A. M. (2012). Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4. Implementation Handbook 2012 Edition. *AHRQ Publication No. 12-0014*. Rockville, MD: Agency for Healthcare Research and Quality.

- [3] Choi, H. J., Park, J. B., & Kim, Y. H. (2019). Implementation of the Korean Triage and Acuity Scale (KTAS) in emergency care. *Journal of the Korean Society of Emergency Medicine*, 30(1), 1–10. [10.4266/acc.2019.00647](https://doi.org/10.4266/acc.2019.00647)
- [4] OpenAI (2023). GPT-4 Technical Report. *arXiv preprint*, arXiv:2303.08774. [10.48550/arXiv.2303.08774](https://arxiv.org/abs/2303.08774)
- [5] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint*, arXiv:2302.13971. [10.48550/arXiv.2302.13971](https://arxiv.org/abs/2302.13971)
- [6] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., et al. (2023). MIMIC-IV-ED, a freely accessible database of emergency department admissions. *Scientific Data*, 10(1), 1–11. [10.1038/s41597-023-01948-3](https://doi.org/10.1038/s41597-023-01948-3)
- [7] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930–1940. [10.1038/s41591-023-02548-w](https://doi.org/10.1038/s41591-023-02548-w)
- [8] Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J. N., Laleh, N. G., et al. (2023). The future of landscape reviews for artificial intelligence in clinical medicine. *The Lancet Digital Health*, 5(10), e715–e726. [10.1016/S2589-7500\(23\)00141-8](https://doi.org/10.1016/S2589-7500(23)00141-8)
- [9] Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the feasibility of ChatGPT in healthcare and emergency medicine. *Journal of Medical Internet Research*, 25, e48060. [10.2196/48060](https://doi.org/10.2196/48060)
- [10] Wu, S., Koo, M., Blum, L., Black, A., Liang, L., & Wang, F. (2024). A systematic review of large language models in healthcare: Applications, performance, and safety. *NPJ Digital Medicine*, 7(1), 42. [10.1038/s41746-024-01050-x](https://doi.org/10.1038/s41746-024-01050-x)
- [11] Avery, E., Zhang, H., Patel, R., & Subrahmanian, E. (2024). Systematic review of conversational AI in emergency triage and symptom checking. *Annals of Emergency Medicine*, 83(4), 382–395. [10.1016/j.annemergmed.2024.01.012](https://doi.org/10.1016/j.annemergmed.2024.01.012)
- [12] A. K. I. Riad et al. (2026). Governance-Aware, Privacy-Preserving Retrieval-Augmented Large Language Models for Clinical Triage Decision Support. *2026 IEEE International Conference on Digital Health (ICDH)*. [10.1109/ICDH72779.2026.00012](https://doi.org/10.1109/ICDH72779.2026.00012)
- [13] Alhusain L. (2026). TRAG: A Rule Retrieval-Augmented Generation Framework for Emergency Triage. *Applied Sciences (Switzerland)*. [10.3390/app16094533](https://doi.org/10.3390/app16094533)
- [14] Baek C. (2026). Component-Level Contributions of Retrieval-Augmented LLM Post-Processing in Streaming Anomaly Detection: A Matched-Operating-Point

- [15] Bass T. (2026). Deterministic Blackboard Pipelines with Specialized LLM Knowledge Sources: A Generalizable Architecture for Intelligent Multi-Stage Reasoning. *IAIT 2026 - 14th International Conference on Advances in Information Technology*. [10.3390/s26165246](#)
- [16] Benachour A. et al. (2026). Multi-Agent Architecture for Automated Patient Intake and Triage: A Comprehensive Framework for AI-Enhanced Clinical Decision Support. *Lecture Notes in Networks and Systems*. [10.1007/978-3-032-11411-2_23](#)
- [17] C. Guyeux (2026). Safe-by-Design Governance for LLM-Assisted Emergency Call Triage: Auditability, Human Oversight, and Degraded-Mode Resilience. *2026 IEEE International Conference on Smart Sustainable Systems for Computer and Engineering Applications (3SCEA)*. [10.1109/3SCEA68071.2026.11603160](#)
- [18] Chahid A. et al. (2026). An Advanced Conversational AI System for Medical Specialty Triage Featuring a RAG-Based Dialogue Engine. *Lecture Notes in Networks and Systems*. [10.1007/978-3-032-18894-6_21](#)
- [19] Deng L. et al. (2026). Evaluation of large language models in emergency medicine scenarios: a comparative analysis of ChatGPT-4o, ChatGPT-o3mini, Gemini 2.0-pro, and DeepSeek-R1. *BMC Emergency Medicine*. [10.1186/s12873-026-01511-0](#)
- [20] Edhala G. et al. (2026). Provably Safe Action Spaces for Agentic AI in Time-Critical Clinical Workflows. *2026 International Conference on Intelligent Systems and Digital Applications, ISDA 2026 - Proceedings*. [10.1109/ISDA70544.2026.11606052](#)
- [21] Erdem P. et al. (2026). Comparison of the performance of ChatGPT Plus, Gemini Pro, and experienced nurses in pediatric emergency triage using physician reference assessment: A prospective observational study. *Journal of Pediatric Nursing*. [10.1016/j.pedn.2026.06.004](#)
- [22] Hack S. et al. (2026). Cautionary lessons from real-world testing of GPT-4.1 AI for pediatric foreign body aspiration. *European Archives of Oto-Rhino-Laryngology*. [10.1007/s00405-025-09856-1](#)
- [23] Haim G.B. et al. (2026). Evaluating Large Language Model-Assisted Emergency Triage: A Comparison of Acuity Assessments by GPT-4 and Medical Experts. *Journal of Clinical Nursing*. [10.1111/jocn.17490](#)
- [24] Jia E. et al. (2026). OpenEvidence errs on the safe side in a structured test of triage recommendations. *International Journal of Medical Informatics*.

[10.1016/j.ijmedinf.2026.106687](https://doi.org/10.1016/j.ijmedinf.2026.106687)

- [25] Lange F. et al. (2026). Real-world use and evaluation of a generative AI chatbot for Parkinson’s disease information: a prospective observational study. *The Lancet Regional Health - Europe*. [10.1016/j.lanepe.2026.101866](https://doi.org/10.1016/j.lanepe.2026.101866)
- [26] Lee S. et al. (2026). Development of BERT-based large language models for emergency department triage using real-world conversations. *Journal of the American Medical Informatics Association*. [10.1093/jamia/ocag007](https://doi.org/10.1093/jamia/ocag007)
- [27] Liu Z. et al. (2026). A fine-tuned large language model chatbot for multi-scenario radiology cancer care: randomized controlled trial on interaction optimization, emotional support, and provider burnout reduction. *Journal of Translational Medicine*. [10.1186/s12967-026-07738-6](https://doi.org/10.1186/s12967-026-07738-6)
- [28] Ma Z. et al. (2026). Performance of DeepSeek V3.2 and ChatGPT 5.1 in Musculoskeletal Triage and Differential Diagnosis of Outpatients With Low Back Pain: Multidimensional Comparative Study. *Journal of Medical Internet Research*. [10.2196/92315](https://doi.org/10.2196/92315)
- [29] Michelsen C. (2026). From early risk to preventive action: Empowerment based digital triage and intervention for burnout, anxiety, and depression in Scandinavian healthcare workers. *AIMS Public Health*. [10.3934/publichealth.2026029](https://doi.org/10.3934/publichealth.2026029)
- [30] Mishra N. et al. (2026). A Multi-Agent LLM Platform for Integrated Healthcare Services with Multi-Modal Data Analysis. *Procedia Computer Science*. [10.1016/j.procs.2026.05.177](https://doi.org/10.1016/j.procs.2026.05.177)
- [31] Narang K. et al. (2026). Assessing acuity in pediatric emergency department triage: performance of a large language model. *npj Health Systems*. [10.1038/s44401-026-00128-6](https://doi.org/10.1038/s44401-026-00128-6)
- [32] Nedos I. et al. (2026). 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. *Journal of Clinical Medicine*. [10.3390/jcm15041512](https://doi.org/10.3390/jcm15041512)
- [33] Nguyen L.T. et al. (2026). Developing and Fine-Tuning Large Language Models for Healthcare in Vietnam: From Indigenous Data to Safe Clinical Applications. *ICIIT 2026 - Proceedings of 2026 11th International Conference on Intelligent Information Technology*. [10.1145/3805862.3805898](https://doi.org/10.1145/3805862.3805898)
- [34] Panaro I. et al. (2026). A Hierarchical Multi-Agent System for Dynamic Orchestration of Clinical Workflows. *2026 International Conference on Integrated Intelligence and Cognitive Engineering, ICIICE 2026*. [10.1109/ICIICE69672.2026.11565478](https://doi.org/10.1109/ICIICE69672.2026.11565478)

- [35] Park C. et al. (2026). Modeling the Clinical Reasoning Workflow: A Dynamic, Time-Aware CDSS for the Emergency Department. *Studies in Health Technology and Informatics*. [10.3233/SHTI260216](#)
- [36] Pious I. et al. (2026). Triagebot: A Resource-Efficient AI Model for Symptom-Driven Clinical Triage in Healthcare Systems. *2026 International Conference on Emerging Trends in Mobile Computing and Sustainable Informatics, ICEMCSI 2026*. [10.1109/ICEMCSI67638.2026.11603098](#)
- [37] R. Gupta et al. (2026). HealthSync: Secure PHR Management, Real-Time Scheduling, and Responsible AI Symptom Triage for Unified Digital Healthcare. *2026 International Conference on Green Technologies and Sustainable Solutions (GTSS)*. [10.1109/GTSS68743.2026.11679831](#)
- [38] Ramacciani Iseman C. et al. (2026). Can AI improve triage quality? A preliminary assessment of ChatGPT performance in evaluating triage decisions. *International Emergency Nursing*. [10.1016/j.ienj.2026.101868](#)
- [39] Ramaswamy A. et al. (2026). ChatGPT Health performance in a structured test of triage recommendations. *Nature Medicine*. [10.1038/s41591-026-04297-7](#)
- [40] Reis F. et al. (2026). Disclaimers and Referral Patterns for Medical Advice Across Urgency Levels: Large Language Model Evaluation Study. *Journal of Medical Internet Research*. [10.2196/84668](#)
- [41] Saeed O. et al. (2026). An Intelligent Triage System for Hospital Emergency Services Based on Large Language Models and Machine Learning. *Proceedings - 5th International Conference on Informatics and Software Engineering, IISEC 2026*. [10.1109/IISEC69317.2026.11418412](#)
- [42] Saxena T. et al. (2026). Planner-Executor Architecture for Teledermatology Triage with Safety Guardrails. *Proceedings of the 2026 13th International Conference on Computing for Sustainable Global Development, INDIACom 2026*. [10.23919/INDIACom70271.2026.11526130](#)
- [43] Shaikh M. et al. (2026). FairGuard: Blockchain-Enforced Continuous Fairness Governance for Demographically Equitable LLM-Based Emergency Triage Decision Support. *Studies in Health Technology and Informatics*. [10.3233/SHTI260872](#)
- [44] Sharma K. et al. (2026). ED-triage-agent: a multi-agent framework for human-in-the-loop emergency triage. *International Journal of Medical Informatics*. [10.1016/j.ijmedinf.2026.106652](#)
- [45] Shen Z. et al. (2026). Enhancing emergency medical service diversion decision-making through large language models integration. *Intelligent Medicine*. [10.1016/j.imed.2025.04.004](#)

- [46] Solano C. et al. (2026). Generative artificial intelligence in the daily urology unit: A practical, patient-centered framework. *Journal of Clinical Urology*. [10.1177/20514158261445149](#)
- [47] Song J. et al. (2026). Initial-Visit Specialty Triage in Rare Diseases Using Large Language Models: Retrospective Benchmarking Study. *Journal of Medical Internet Research*. [10.2196/101711](#)
- [48] Strechen I. et al. (2026). Clinical evaluation of a vision-language model for optimizing triage and clinical workflows in critical care. *International Journal of Medical Informatics*. [10.1016/j.ijmedinf.2026.106673](#)
- [49] Munasing A. R. Tharana et al. (2026). Unified AI Framework for Predictive Pediatric Healthcare in Multilingual, Resource-Limited Settings. *2026 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*. [10.1109/I2CACIS69435.2026.11600063](#)
- [50] Venerito V. et al. (2026). Agentic AI in rheumatology. *The Lancet Rheumatology*. [10.1016/S2665-9913\(26\)00076-7](#)
- [51] Wang J. et al. (2026). An Electronic Health Record-Integrated, Large Language Model-Powered Tool to Triage Surgical Patients. *JAMA Network Open*. [10.1001/jamanetworkopen.2026.29579](#)
- [52] Wong H.S. et al. (2026). Multi-Evidence Clinical Reasoning With Retrieval-Augmented Generation for Emergency Triage: Retrospective Evaluation Study. *JMIR Medical Informatics*. [10.2196/82026](#)
- [53] Xie W. et al. (2026). Retrieval-augmented large language models for depression screening and suicide risk stratification. *BMC Psychiatry*. [10.1186/s12888-026-07988-0](#)
- [54] Zeinali F. et al. (2026). Predicting emergency department disposition using machine learning and large language models to support proactive capacity management: a multicenter retrospective study. *BMC Emergency Medicine*. [10.1186/s12873-026-01487-x](#)
- [55] Zeng D. et al. (2026). Generative AI for pre-consultation mental health triage in disorders of gut-brain interaction. *Frontiers in Psychiatry*. [10.3389/fpsyt.2026.1854670](#)
- [56] Zhang J. et al. (2026). Towards Efficient and Robust Linguistic Emotion Diagnosis for Mental Health via Multi-Agent Instruction Refinement. *IEEE Transactions on Affective Computing*. [10.1109/TAFFC.2026.3668261](#)
- [57] Zhao L. et al. (2026). Improving Emergency Department Efficiency With Large Language Model-Guided Orthopaedic Triage for Proximal Humerus Fractures.

- [58] Zheng J. et al. (2026). MediLink: Safer Respiratory Triage Using Retrieval-Augmented Generation and Rule-Based Safety Gates. *2026 Systems and Information Engineering Design Symposium, SIEDS 2026*. [10.1109/SIEDS69358.2026.11540307](https://doi.org/10.1109/SIEDS69358.2026.11540307)
- [59] Çalışkan Y.K. et al. (2026). When algorithms triage trauma: Diagnostic accuracy, undertriage risk, and prompt fragility in frontier large language models. *Surgery (United States)*. [10.1016/j.surg.2026.110319](https://doi.org/10.1016/j.surg.2026.110319)
- [60] Özer V. et al. (2026). SCENARIO-BASED PREHOSPITAL TRIAGE OF CARBON MONOXIDE POISONING: COMPARING FIVE LARGE LANGUAGE MODELS WITH EMERGENCY MEDICAL SERVICES PERSONNEL IN A PROSPECTIVE STUDY; [Karbonmonoksit Zehirlenmesinde Senaryo Tabanlı Hastane Öncesi Triyajı: Beş Büyük Dil Modeli ile Acil Sağlık Hizmetleri Personelinin Prospektif Karşılaştırılması]. *Journal of Kırıkkale University Faculty of Medicine*. [10.24938/kutfd.1922944](https://doi.org/10.24938/kutfd.1922944)
- [61] Charan K.S. et al. (2025). Intelligent Symptom Conversations: New Frontiers in AI-Powered Preliminary Healthcare Diagnosis. *2025 International Conference on Future Technologies, ICFT 2025*. [10.1109/ICFT66708.2025.11336630](https://doi.org/10.1109/ICFT66708.2025.11336630)
- [62] Elhilali A. et al. (2025). AI-Enhanced Speech Recognition in Triage. *Studies in Health Technology and Informatics*. [10.3233/SHTI250213](https://doi.org/10.3233/SHTI250213)
- [63] Gaber F. et al. (2025). Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*. [10.1038/s41746-025-01684-1](https://doi.org/10.1038/s41746-025-01684-1)
- [64] Gao H. et al. (2025). A large language model based pipeline for extracting information from patient complaint and anamnesis in clinical notes for severity assessment. *Scientific Reports*. [10.1038/s41598-025-07649-4](https://doi.org/10.1038/s41598-025-07649-4)
- [65] Garg T. et al. (2025). Rule-Guided, Cost-Sensitive LLMs for Optimal Maternal Ambulance Dispatch Triage in Resource-Constrained Settings. *Proceedings of the IEEE International Conference on Big Data, BigData*. [10.1109/BigData66926.2025.11402461](https://doi.org/10.1109/BigData66926.2025.11402461)
- [66] Ghatage R. et al. (2025). TRAI: An AI-Driven Mobile Application to Reduce the Gap Between Triage and Care. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. [10.1109/EMBC58623.2025.11252667](https://doi.org/10.1109/EMBC58623.2025.11252667)
- [67] Guo J. et al. (2025). Enhancing Intelligent Triage with Large Language Models: A Comprehensive Evaluation and Optimization Study. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and*

- [68] Han S. et al. (2025). Development of a Large Language Model-based Multi-Agent Clinical Decision Support System for Korean Triage and Acuity Scale (KTAS)-Based Triage and Treatment Planning in Emergency Departments. *Advances in Artificial Intelligence and Machine Learning*. [10.54364/AAIML.2025.51187](https://doi.org/10.54364/AAIML.2025.51187)
- [69] Ho B. et al. (2025). Evaluation of Generative Artificial Intelligence Models in Predicting Pediatric Emergency Severity Index Levels. *Pediatric Emergency Care*. [10.1097/PEC.00000000000003315](https://doi.org/10.1097/PEC.00000000000003315)
- [70] Jiyeong Kim et al. (2025). Optimizing large language models for detecting symptoms of depression/anxiety in chronic diseases patient communications. *npj Digital Medicine*. [10.1038/s41746-025-01969-5](https://doi.org/10.1038/s41746-025-01969-5)
- [71] Shivani Kinagi et al. (2025). Medaid: A Safety-First AI Triage System for Rural Healthcare. *IEEE International Conference*. Not Reported
- [72] Nathalie Maria Kirch et al. (2025). Medical triage as an AI ethics benchmark. *Scientific Reports*. [10.1038/s41598-025-16716-9](https://doi.org/10.1038/s41598-025-16716-9)
- [73] Priscilla Koirala et al. (2025). Evaluating AI performance in nephrology triage and subspecialty referrals. *Scientific Reports*. [10.1038/s41598-025-88074-5](https://doi.org/10.1038/s41598-025-88074-5)
- [74] Marvin Kopka et al. (2025). Technology-supported self-triage decision making. *npj Health Systems*. [10.1038/s44401-024-00008-x](https://doi.org/10.1038/s44401-024-00008-x)
- [75] Carlos Lafuente et al. (2025). Evaluating large language models on hospital health data for automated emergency triage. *International Journal of Computer Assisted Radiology and Surgery*. [10.1007/s11548-025-03475-1](https://doi.org/10.1007/s11548-025-03475-1)
- [76] Alessandro Giaj Levra et al. (2025). A large language model-based clinical decision support system for syncope recognition in the emergency department: A framework for clinical workflow integration. *European Journal of Internal Medicine*. [10.1016/j.ejim.2024.09.017](https://doi.org/10.1016/j.ejim.2024.09.017)
- [77] Siru Liu et al. (2025). Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. *Journal of the American Medical Informatics Association*. [10.1093/jamia/ocaf059](https://doi.org/10.1093/jamia/ocaf059)
- [78] Luthra S.R. et al. (2025). The Effectiveness of Specialised Conversational AI in Preliminary Medical Triage: A Comparative Analysis of Non-Maleficence-by-Design and General-Purpose Architecture. *2025 IEEE International Conference on Modern Electronics Devices and Intelligent Communication Systems, MEDCOM 2025*. [10.1109/MEDCOM67532.2025.11404909](https://doi.org/10.1109/MEDCOM67532.2025.11404909)

- [79] Pasli S. et al. (2025). ChatGPT-supported patient triage with voice commands in the emergency department: A prospective multicenter study. *American Journal of Emergency Medicine* 94 (2025) 63–70. [10.1016/j.ajem.2025.04.040](https://doi.org/10.1016/j.ajem.2025.04.040)
- [80] Seo J.W. et al. (2025). Artificial intelligence for severity triage based on conversations in an emergency department in Korea. *Scientific Reports* (2025) 15:16870. [10.1038/s41598-025-99874-0](https://doi.org/10.1038/s41598-025-99874-0)
- [81] Shekhar A.C. et al. (2025). Use of a large language model (LLM) for ambulance dispatch and triage. *American Journal of Emergency Medicine* 89 (2025) 27–29. [10.1016/j.ajem.2024.12.032](https://doi.org/10.1016/j.ajem.2024.12.032)
- [82] Shen Q. et al. (2025). Knowledge-embedded large language models for emergency triage. *Knowledge-Based Systems* 318 (2025) 113431. [10.1016/j.knosys.2025.113431](https://doi.org/10.1016/j.knosys.2025.113431)
- [83] Soares C. et al. (2025). C-PATH: Conversational Patient Assistance and Triage in Healthcare System. *IEEE Conference / arXiv preprint*. Not Reported
- [84] Kaspoju S. et al. (2025). Medibot: A Chatbot Application for Medical Emergency. *IEEE Conference / International Conference Proceedings*. Not Reported
- [85] Thotapalli S. et al. (2025). Potential of ChatGPT in youth mental health emergency triage: Comparative analysis with clinicians. *Psychiatry and Clinical Neurosciences Reports*. [10.1002/pcn5.70159](https://doi.org/10.1002/pcn5.70159)
- [86] Wang C. et al. (2025). Patient Triage and Guidance in Emergency Departments Using Large Language Models: Multimetric Study. *Journal of Medical Internet Research*. [10.2196/71613](https://doi.org/10.2196/71613)
- [87] Yazaki M. et al. (2025). Emergency Patient Triage Improvement through a Retrieval-Augmented Generation Enhanced Large-Scale Language Model. *Prehospital Emergency Care*. [10.1080/10903127.2024.2374400](https://doi.org/10.1080/10903127.2024.2374400)
- [88] Zhao X. (2025). The Research on Intelligent Medical Triage System Based on Large Language Models. *Proceedings - 2025 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2025*. [10.1109/BIBM66473.2025.11356579](https://doi.org/10.1109/BIBM66473.2025.11356579)
- [89] Zhao Z. et al. (2025). Privacy-First Triage Classification with Open-Weight LLMs: A Chain-of-Thought Distillation Approach. *IEEE International Conference on Data Mining Workshops, ICDMW*. [10.1109/ICDMW69685.2025.00432](https://doi.org/10.1109/ICDMW69685.2025.00432)
- [90] Zou Y. et al. (2025). Comparison of triage performance among DRP tool, ChatGPT, and outpatient rehabilitation doctors. *Scientific Reports*. [10.1038/s41598-025-99216-0](https://doi.org/10.1038/s41598-025-99216-0)

- [91] Gomez C. et al. (2024). How large language model-powered conversational agents influence decision making in domestic medical triage contexts. *Frontiers in Computer Science*. [10.3389/fcomp.2024.1427463](https://doi.org/10.3389/fcomp.2024.1427463)
- [92] Guerra-Adames A. et al. (2024). Uncovering Judgment Biases in Emergency Triage: A Public Health Approach Based on Large Language Models. *Proceedings of Machine Learning Research*.
- [93] Hu B. et al. (2024). Language Models are Alignable Decision-Makers: Dataset and Application to the Medical Triage Domain. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*. [10.18653/v1/2024.naacl-industry.18](https://doi.org/10.18653/v1/2024.naacl-industry.18)
- [94] Kopka M. et al. (2024). The RepVig framework for designing use-case specific representative vignettes and evaluating triage accuracy of laypeople and symptom assessment applications. *Scientific Reports*. [10.1038/s41598-024-83844-z](https://doi.org/10.1038/s41598-024-83844-z)
- [95] Lee S. et al. (2024). Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage. *American Journal of Emergency Medicine*. [10.1016/j.ajem.2023.11.063](https://doi.org/10.1016/j.ajem.2023.11.063)
- [96] Lu M. et al. (2024). TRIAGEAGENT: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage. *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024*. [10.18653/v1/2024.findings-emnlp.329](https://doi.org/10.18653/v1/2024.findings-emnlp.329)
- [97] Mishra V. et al. (2024). Large Language Model-based Mobile Apps in OPDs: Optimized Patient Management and Prioritization. *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry, IDICAIEI 2024*. [10.1109/IDICAIEI61867.2024.10842782](https://doi.org/10.1109/IDICAIEI61867.2024.10842782)
- [98] Sinha K. et al. (2024). Healthcare Diagnostic RAG-Based Chatbot Triage Enabled by BioMistral-7B. *Proceedings of International Conference on Emerging Technologies and Innovation for Sustainability, EmergIN 2024*. [10.1109/EmergIN63207.2024.10961136](https://doi.org/10.1109/EmergIN63207.2024.10961136)
- [99] Taylor N. et al. (2024). Model development for bespoke large language models for digital triage assistance in mental health care. *Artificial Intelligence in Medicine*. [10.1016/j.artmed.2024.102988](https://doi.org/10.1016/j.artmed.2024.102988)
- [100] Wan P. et al. (2024). Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nature Medicine*. [10.1038/s41591-024-03148-7](https://doi.org/10.1038/s41591-024-03148-7)
- [101] Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and

- weaknesses. *The FASEB Journal*, 22(2), 338–342. [10.1096/fj.07-9492LSF](https://doi.org/10.1096/fj.07-9492LSF)
- [102] Schotten, M., el Aisati, M., Choi, W. J., & Steinginga, S. (2017). A brief history of Scopus: The world’s largest abstract and citation database of scientific literature. *Research Analytics*, 31–58.
- [103] Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a study by country and discipline. *Scientometrics*, 106(1), 213–228. [10.1007/s11192-015-1765-5](https://doi.org/10.1007/s11192-015-1765-5)