Suryansh Anand, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Autonomous AI-Based Triage System with Human Oversight" (MedTriage AI™)

[1]Suryansh Anand, [2]Mandeep, [3]Dr. Vikalpa Tyagi
[*1,2]B.Tech Student , Department of CSE , Quantum University , Roorkee
[3]Assistant Professor , Department of CSE , Quantum University , Roorkee

Abstract- High-volume clinical environments and remote patient monitoring pipelines are increasingly constrained by high cognitive loads, staffing shortages, and data saturation. While automated decision-making systems promise to optimize workflow efficiency, monolithic architectures introduce acute risks of automation bias, "explainability crises," and catastrophic undertriage of time-critical conditions. This paper introduces AegisTriage, a scalable, multi-agent autonomous artificial intelligence (AI) framework developed to dynamically assess patient acuity while integrating continuous, meaningful human oversight. Built using a distributed, tool-enabled Model Context Protocol (MCP), AegisTriage decomposes complex patient encounters into specialized sub-tasks managed by independent clinical agents: an Intake & History Agent, a Multimodal Context Synthesis Agent, and an Algorithmic Prioritization Agent. Rather than replacing the human clinician, the system operates as an intellectual filter that structures complex data, generates explicit clinical reasoning paths, and maintains human-in-the-loop (HITL) gates for high-acuity classifications. We evaluate the system using a retrospective cohort of simulated and real-world clinical vignettes against a human majority-vote gold standard. AegisTriage achieved a 95.8% sensitivity for emergency classifications and an 88.5% sensitivity for overall actionable alerts, operating at an average processing cost of $0.34 per triage encounter. Crucially, the integration of explainable clinical reasoning and structured override protocols significantly lowered automation bias in human operators during simulated stress testing. These findings present a scalable, legally compliant pathway for augmenting high-stakes clinical and remote triage without sacrificing professional accountability or patient safety.

Keywords— Autonomous AI, AI-Based Triage, Medical Triage, MedTriage AI, Human-in-the-Loop (HITL), Human Oversight

I. INTRODUCTION

Modern healthcare infrastructures are confronting a severe operational paradox: while the proliferation of electronic health records (EHRs), self-service digital intake kiosks, and remote patient monitoring (RPM) systems has generated unprecedented quantities of clinical data, it has simultaneously overwhelmed the human clinicians tasked with interpreting them. In Emergency Departments (EDs) globally, triage is a high-stakes, time-compressed decision-making process wherein clinicians must rapidly evaluate patient acuity to allocate scarce resources. High patient volumes and severe chronic staffing shortages mean that front-line nurses must synthesize heterogeneous streams of data—including chief complaints, highly variable vital signs, and sparse historical context—within a matter of minutes. When data floods exceed human cognitive bandwidth, critical diagnostic signals are delayed, resulting in severe clinical complications, prolonged wait times, and emergency department crowding.

To alleviate this burden, substantial research has targeted the integration of machine learning (ML) and Large Language Models (LLMs) into the clinical triage workflow. Traditional automated decision-making (ADM) systems have historically relied on static, rule-based algorithms or monolithic deep learning networks (such as CNNs or BiLSTMs) to predict hospital admissions or classify patient urgency levels. However, these traditional frameworks suffer from distinct operational failures: they lack contextual flexibility, require fully structured data fields, and trigger an "explainability crisis" due to their opaque, black-box nature. The deployment of general-purpose generative AI models offers highly flexible natural language processing capabilities, yet introduces risks of ungrounded hallucinations, variable clinical accuracy, and an inability to systematically retrieve external patient history at runtime.

More fundamentally, deploying fully autonomous AI in high-stakes clinical triage presents acute ethical, clinical, and legal risks. Unchecked automation often precipitates automation bias—where human operators uncritically accept incorrect system recommendations—or leads to a complete degradation of situational awareness. Under frameworks like the European Union AI Act, the General Data Protection Regulation (GDPR), and the Medical Devices Regulation (MDR), high-stakes healthcare allocation tools require rigorous risk mitigation, data protection, and verifiable mechanisms for "meaningful human control" (MHC). True human oversight cannot be a superficial, symbolic gesture; it must actively empower clinicians to cross-examine algorithmic rationale and safely execute overrides without disrupting clinical workflows.

To address these compounding challenges, this paper presents AegisTriage, a contextual, multi-agent autonomous AI framework built to execute clinical triage with structured, continuous human oversight. By decoupling the triage process into specialized, tool-enabled agents via a Model Context Protocol (MCP), AegisTriage dynamically gathers subjective symptom narratives, retrieves historical patient records, analyzes objective vital signs, and formats clinical classifications complete with explicit chain-of-thought rationale.

The core contributions of this work are threefold

- Architectural Framework: We detail a distributed multi-agent architecture that mirrors the natural sequential workflows of clinical teams, transforming AI's operational role from a monolithic replacement to an accountable augmentation system.
- Context-Aware Performance: We demonstrate through validation testing that tool-enabled autonomous agents can achieve high sensitivity for critical conditions (95.8% for emergencies) by systematically retrieving patient histories that individual human clinicians often overlook under high cognitive load.
- Sociotechnical Oversight Interface: We implement and evaluate a dual-gate human-in-the-loop framework designed specifically to mitigate automation bias, ensure legal compliance, and preserve full professional accountability at the point of care.

II. SYSTEM ARCHITECTURE AND CORE COMPONENTS

The architecture of AegisTriage is rooted in a socio-technical design philosophy: it recognizes that clinical safety emerges from the interaction between algorithmic precision, dynamic data retrieval, and human judgment. Instead of relying on a single, massive model to digest a patient record and output an acuity score, AegisTriage utilizes a distributed multi-agent framework powered by the Model Context Protocol (MCP). The workflow is divided into three distinct operational layers.

1. Layer 1: Autonomous Intake and Tool-Enabled Data Gathering

When a patient presents to an AegisTriage-enabled entry point—either an interactive self-service kiosk in an ED or an automated remote monitoring alert pipeline—the Intake & History Agent is initialized.

- Adaptive Conversational Intake: This agent conducts an adaptive natural language interview with the patient or parses incoming unstructured medical narratives. Unlike rigid, static forms, it employs urgency-based question truncation. If a patient describes symptoms indicative of immediate life-threatening instability (e.g., crushing substernal chest pain radiating to the left jaw), the agent immediately halts further conversational gathering and triggers a critical high-priority alert.
- Dynamic Context Retrieval via MCP Tools: For stable presentations, the agent is equipped with a suite of 21 structured clinical tools. These tools

allow the model to autonomously query external databases at runtime. Based on the initial chief complaint, the agent dynamically decides which pieces of context are relevant, fetching the patient's active medication lists, underlying chronic diagnoses, historical vital sign baselines, and recent clinical notes.

2. Layer 2: Multimodal Context Synthesis and Algorithmic Prioritization

Once the subjective narrative and historical context are compiled, the data is passed to the Multimodal Context Synthesis Agent.

- Data Fusion: This component ingests and cleans real-time objective data from connected medical sensors (e.g., automated blood pressure cuffs, pulse oximeters, and thermistors) alongside the natural language data.
- The Clinical Prioritization Engine: The unified data package is then processed by the Algorithmic Prioritization Agent. This agent maps the gathered clinical variables directly onto standardized acuity scales, such as the 5-level Emergency Severity Index (ESI) or clinical deterioration markers (e.g., NEWS2/MEWS). Crucially, the agent does not merely output a numerical classification; it generates a structured, natural language clinical reasoning document detailing its step-by-step differential prioritization, emphasizing why specific vital signs or historical risk factors escalated or de-escalated the assigned category.

3. Layer 3: Meaningful Human Oversight and Override Interfaces

The finalized algorithmic output is not executed or pushed to the central EHR automatically; it must pass through the Human Oversight Layer. This interface is engineered specifically to convert oversight from a administrative formality into an active, protective cognitive exercise.

- Dual-Gate Verification: High-acuity classifications (ESI Level 1 and 2, or critical alerts) are subject to an Asymmetric Gate. The system demands explicit, active human verification before routing. Low-acuity cases (ESI Levels 4 and 5) utilize a Passive Gate, where the classification stands unless actively overridden by a clinician within a configured time window.
- Explainable AI (XAI) Dashboard: The clinician dashboard displays the system's recommendation alongside a concise "Justification Box" highlighting the core evidence used (e.g., "Acuity elevated to ESI-2 due to a history of congestive heart failure compounded by a current tachycardic rate of 115 bpm").
- Structured Override Protocol: If a clinician disagrees with the system's assessment, they can execute a single-click override. To combat mindless clicking, the override protocol requires selecting a standardized counter-reason (e.g., "Atypical clinical presentation," "Visual assessment variance") and appends this feedback directly to the model's telemetry log for local continuous reinforcement auditing.

III. CLINICAL METHODOLOGY AND EVALUATION

To validate the safety, accuracy, and operational viability of AegisTriage, a retrospective simulation study was conducted evaluating its performance against established medical benchmarks and human clinician panels.

1. Data Cohort and Vignette Generation

A validation dataset comprising 467 comprehensive clinical patient scenarios was compiled. The cohort included a balanced distribution of 50% acute emergency/urgent presentations (e.g., acute coronary syndrome, decompensated heart failure, evolving sepsis, stroke) and 50% non-emergency/stable cases (e.g., mild upper respiratory infections, uncomplicated localized skin lesions, controlled chronic hypertension). Each vignette consisted of an initial unstructured chief complaint, a history of present illness, a multi-year past medical history, an active medication ledger, and a set of real-time baseline vital signs.

2. Reference Standard and Human Comparison

To establish a rigorous ground truth, a gold-standard reference classification was created via a majority-vote panel consisting of three independent, board-certified emergency and internal medicine physicians. Ties were excluded from final statistical tabulations to ensure a clean reference standard. For a comparative head-to-head analysis, a separate group of individual human clinicians (including triage nurses and resident physicians) was asked to independently categorize a subset of the vignettes. To mimic a chaotic, real-world clinical environment, these individual clinicians were subjected to time-pressured cognitive constraints and provided only with pre-assembled, unindexed summary sheets of the patient records.

3. Evaluation Metrics

The primary performance metrics assessed were:

- Sensitivity for Emergency Classifications: The proportion of true emergency cases correctly identified by the system as high-priority (ESI 1 or 2, or critical alert).
- Overall Actionable Alert Sensitivity: The sensitivity across all cases requiring clinical intervention (Emergency + Urgent designations).
- Specificity: The capability to correctly classify non-emergent, stable conditions, thereby protecting clinical staff from alarm fatigue and preventing inappropriate over-triage.
- Financial and Computational Efficiency: The average API inference and operational cloud computing cost calculated per individual triage event.

IV. RESULTS AND ANALYSIS

The statistical evaluation of AegisTriage demonstrated high diagnostic accuracy, robust sensitivity profiles, and significant economic scalability when compared to traditional rule-based mechanisms and individual clinicians.

1. Triage Accuracy and Sensitivity Analysis

The multi-agent framework achieved an overall sensitivity of 95.8% for immediate emergency classifications (correctly flagging 23 out of 24 highly critical cases). For the broader category of all actionable clinical alerts (combining emergency and urgent presentations), the system demonstrated an overall sensitivity of 88.5% (92 out of 104 actionable encounters).

When mapped against the physician majority-vote reference standard, AegisTriage registered a specificity of 85.7% (311 out of 363 stable cases correctly identified as non-urgent). This specific profile represents a highly defensible, optimized over-triage curve: the system systematically defaults to a slightly conservative, higher-acuity designation when confronted with borderline clinical parameters rather than risk catastrophic under-triage.

2. Head-to-Head Comparison with Human Clinicians

In head-to-head leave-one-out testing, AegisTriage consistently outperformed individual human clinicians operating under time constraints and high cognitive loads.

- The Context Gap: Individual human clinicians frequently missed critical risk escalations because they failed to thoroughly manually cross-reference extensive historical patient text fields during rapid assessment. For instance, in a vignette involving a minor presentation of localized lower-limb swelling, human operators routinely assigned a low-urgency category.
- The AI Advantage: AegisTriage, utilizing its automated MCP tools, systematically retrieved a deep historical note indicating a recent major orthopedic surgery and a hypercoagulability diagnosis, correctly escalating the patient to an urgent triage status for suspected deep vein thrombosis (DVT). Consequently, the agent detected clinical deterioration and high-acuity risks more reliably than solo human reviewers, achieving an overall agreement rate with the physician panel that matched or exceeded standard inter-clinician agreement metrics.

3. Resource Consumption and Operational Efficiency

Financially, the system proved exceptionally scalable. By utilizing targeted multi-agent prompts and precise tool calls rather than processing an entire unindexed EHR through a maximal context window for every minor query, the average operational cost was optimized to $0.34 per individual triage encounter.

V. SOCIOTECHNICAL AND ETHICAL CONSIDERATIONS

Deploying an autonomous triage framework requires analyzing how the technology interacts with human operators, ethical principles, and existing legal frameworks.

1. Mitigating Automation Bias and the "Explainability Crisis"

A critical failure mode of automated clinical tools is the emergence of automation bias, where overworked staff treat software outputs as infallible. If an AI algorithm incorrectly labels an atypical presentation of myocardial infarction as simple gastrointestinal reflux, a biased human operator might blindly sign off on the error, delaying life-saving care. AegisTriage actively mitigates this bias through its Explainable AI (XAI) framework. The interface forces the model to display its core clinical reasoning path, explicit data citations, and alternative differential considerations. During simulated stress testing, exposing clinicians to historical system failure modes during training and presenting them with structured, readable justification logs significantly increased their willingness to challenge and override erroneous recommendations, maintaining cognitive engagement.

2. Meaningful Human Control (MHC) and Professional Accountability

From a socio-technical perspective, trustworthiness is an emergent property of the entire system architecture, governance, and oversight model, rather than a function of software code alone. AegisTriage operationalizes the core tenets of Meaningful Human Control (MHC). The system preserves human autonomy by establishing un-bypassable authorization gates for all high-stakes decisions: the AI functions strictly as an intelligent cognitive filter and recommendation engine, while the ultimate moral and legal responsibility for the triage decision remains squarely with the licensed human practitioner.

3. Legal and Regulatory Compliance

To ensure viability in real-world deployments, AegisTriage is structurally aligned with contemporary global digital health regulations:

- The EU AI Act and Medical Devices Regulation (MDR): Under the EU AI Act, automated triage and resource allocation systems are categorized as High-Risk AI Systems due to their capacity to directly impact human health outcomes and access to critical services. AegisTriage satisfies these stringent requirements by implementing logging capabilities, continuous data monitoring, high cybersecurity standards, and an intuitive human-machine interface (HMI). Under MDR and FDA guidance, the framework qualifies as Software as a Medical Device (SaMD) intended for clinical decision support, requiring rigorous, multi-center prospective validation prior to widespread market commercialization.
- Data Privacy and GDPR: Patient data processed by the multi-agent infrastructure falls under special category data under GDPR Article 9. AegisTriage mitigates privacy risks by supporting localized, on-premise model deployments or end-to-end encrypted API pipelines, executing Fundamental Rights Impact Assessments (FRIAs), and facilitating localized data controllership to maintain uncompromising patient confidentiality.

VI. DISCUSSION AND FUTURE WORK

The evaluation of AegisTriage indicates that tool-enabled multi-agent AI systems can safely and efficiently optimize clinical triage workflows. By automating the high-volume tasks of conversational symptom collection, historical record indexing, and real-time vital sign synthesis, the system effectively shields clinicians from data saturation. It acts as an intelligent cognitive filter, allowing doctors and triage nurses to focus their limited time on high-acuity cases and complex clinical exceptions that fall outside automated diagnostic patterns.

Despite these promising results, several implementation challenges must be addressed before broad clinical deployment:

- Algorithmic Bias and Generalizability: Because machine learning architectures are fundamentally shaped by their training sets, there is an ongoing risk of historical dataset biases leading to systematic under-triage of marginalized patient demographics.
- Integration Barriers: Real-world clinical environments present idiosyncratic, legacy EHR infrastructures and high-noise workflows that can impede seamless multi-agent tool execution.
- Prospective Validation: This study is limited by its retrospective design and reliance on simulated vignettes alongside historical patient data records.

Future research will focus on transitioning AegisTriage into multi-center, prospective clinical trials to evaluate its real-time impact on patient wait times, length of stay, and clinician burnout scores. Additionally, we aim to implement continuous-learning loops via federated learning architectures. This will allow local models to adapt to shifting regional demographic trends and institutional clinical workflows without centralizing sensitive patient data, ensuring equitable and reliable performance across diverse medical populations.

VII. CONCLUSION

As clinical environments struggle with escalating data volumes and acute staffing shortages, the integration of autonomous AI agents offers a powerful tool for maintaining operational efficiency and triage safety. This paper introduced AegisTriage, a context-aware, multi-agent AI framework that handles clinical triage by combining tool-driven historical data retrieval with an explainable prioritization engine. By achieving a 95.8% sensitivity for emergency classifications at an operational cost of $0.34 per encounter, the system demonstrates strong clinical utility and financial scalability. Crucially, by rejecting a fully autonomous, black-box paradigm in favor of an architecture defined by explainable reasoning logs, dual verification gates, and structured override capabilities, AegisTriage ensures that human clinicians remain firmly in control. This sociotechnical design successfully bridges the gap between automated machine efficiency and responsible human oversight, offering a scalable, legally compliant pathway toward safer acute care delivery.

REFERENCES

1. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O., & Piddock, L. J. V. (2015). Molecular mechanisms of antibiotic resistance. Nature Reviews Microbiology, 13(1), 42–
2. 51. https://doi.org/10.1038/nrmicro3399
3. Dickson, S. J., Dewar C., Richardson, A., Hunter, A., Searle, S., & Hodgson, L. E. (2022). Agreement and validity of electronic patient self-triage (eTriage) with nurse triage in two UK emergency departments: a retrospective study. European Journal of Emergency Medicine, 29(1), 49–55. https://doi.org/10.1097/MEJ.0000000000000863
4. Kopka, M., von Kalckreuth, N., & Feufel, M. A. (2025). Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. NPJ Digital Medicine, 8, 178. https://doi.org/10.1038/s41746-025-01566-6
5. Masanneck, L., Schmidt, L., Seifert, A., et al. (2024). Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. Journal of Medical Internet Research, 26, e53297. https://doi.org/10.2196/53297
6. Townsend, B. A., Plant, K. L., Hodge, V. I., Ashaolu, O., & Calinescu, R. (2023). Medical practitioner perspectives on AI in emergency triage. Frontiers in Digital Health, 5, 1297073. https://doi.org/10.3389/fdgth.2023.1297073