Analyst-Supervised Multi-Agent AI for Context-Aware Cyber Alert Triage

Kristo Kapshtica
Applied Artificial Intelligence
Alma Mater Europaea University
Maribor, Slovenia
kristo.kapshtica@almamater.si

Abstract

Security Operations Centers (SOCs) face high alert volumes, fragmented contextual evidence, and limited analyst time. This paper presents an analyst-supervised multi-agent workflow that separates alert normalization, context enrichment, reasoning, clarification, and case summarization. A bounded-autonomy gate requests analyst input when the evidence is insufficient; official closure or escalation remains subject to analyst review. A controlled pilot compares manual triage, single-agent support, and the proposed workflow on 15 cases derived from AIT alert and log data. The multi-agent workflow matches the manual reference on disposition and clarification decisions (15/15 each), compared with 11/15 for the single-agent baseline. Reported mean time estimates decrease from 7.27 to 2.40 minutes per case, and the analyst effort proxy decreases from 8.73 to 1.53. These preliminary results suggest benefits from structured evidence preparation and explicit uncertainty handling. The small constructed corpus and estimated workload measures limit conclusions about operational performance.

Keywords

multi-agent systems, alert triage, bounded autonomy, human-in-the-loop, context enrichment

1 Introduction

Security Operations Centers (SOCs) must turn heterogeneous alerts into defensible decisions under time pressure. Alert fatigue reflects both the volume of detections and the effort required to interpret them: analysts consult asset inventories, identities, recent alerts, allowlists, and logs before deciding whether activity is benign, suspicious, or malicious [1, 2]. Ranking alerts helps allocate attention, but does not by itself assemble the evidence or produce a reviewable case note.

AlertPro [3] and Context2Vector [4] show the value of historical and environmental evidence. Large language models (LLMs) offer additional support for interpreting and summarizing that evidence, but plausible explanations can mask unsupported conclusions. Cybersecurity surveys and human–AI teaming research therefore emphasize evaluation, traceability, and analyst control [5, 6, 7].

We investigate a staged, analyst-supervised workflow rather than a new alert classifier. Its contribution is the integration of specialized processing roles, contextual evidence, and a clarification gate into an analyst-ready case. Three research questions guide the evaluation: **RQ1**, can this workflow reduce estimated triage time and analyst effort relative to manual and single-agent support? **RQ2**, does staged enrichment improve required evidence coverage? **RQ3**, does explicit gating improve the decision to request clarification when evidence is insufficient? Disposition correctness and severity agreement are retained as checks that efficiency does not come at the expense of decision quality.

The hypothesis is that this decomposition reduces repetitive evidence preparation while preserving reviewability and appropriate uncertainty. The 15-case pilot tests that hypothesis under shared case conditions; it does not establish performance across SOCs or isolate the contribution of each individual agent.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.14

2 Related Work

Research on alert fatigue and prioritization identifies repeated evidence gathering, contextual interpretation, and decision documentation as important sources of analyst workload [1, 2]. Prioritization can reduce the cost of selecting the next alert, while leaving substantial investigation work after selection.

AlertPro applies contextual information and reinforcement learning to prioritization in multi-step attack detection [3]. Context2Vector learns representations of event context to support security-event triage [4]. These studies motivate the use of context; the present work focuses on organizing evidence into an auditable case and preserving ambiguity when a definitive verdict is not justified.

LLM and agentic-AI research discusses opportunities for cyber defense alongside concerns about reliability and evaluation [5, 6]. Human–AI teaming emphasizes complementary roles and analyst situation awareness [7]. Our contribution is a compact workflow and preliminary comparison that connect these concerns to explicit outputs: a disposition, evidence bundle, explanation, and clarification decision. The evaluation compares the complete workflow with manual and single-agent baselines; it is not an ablation study of multi-agent reasoning.

3 Methodology

3.1 Case Representation

Let a be a trigger alert, T the available contextual evidence sources, and h optional analyst feedback. The workflow produces an analyst-ready case:

$$c = f(a, T, h), \qquad c = (a', e, d, s, \gamma, j), \tag{1}$$

where a' is the normalized alert, e the evidence bundle, d the recommended disposition, s the severity, γ the confidence estimate, and j the evidence-grounded justification. Confidence is an advisory output, not a calibrated probability of correctness. A case must be reviewable through its evidence and rationale, rather than accepted on confidence alone.

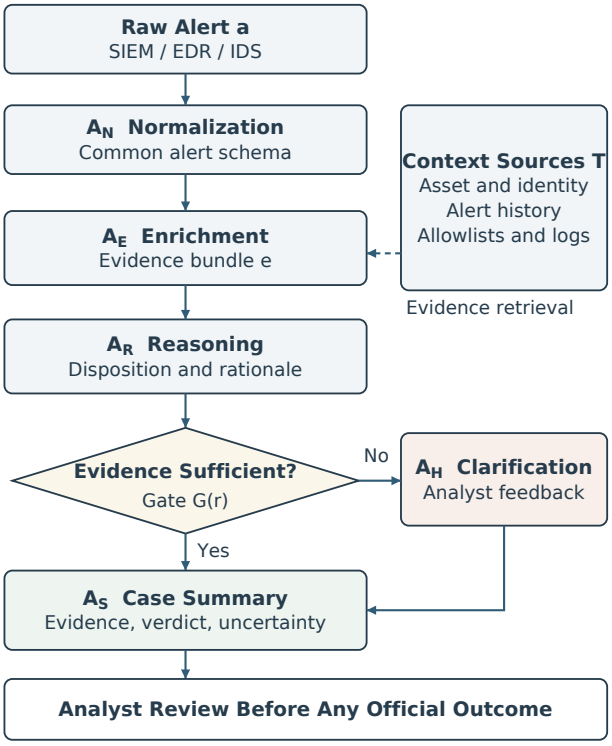

Figure 1: Analyst-supervised workflow. The dashed link supplies context only to enrichment; solid arrows show processing and clarification paths. Both branches lead to a case summary and mandatory analyst review.

3.2 Agent Roles and Evidence Flow

The workflow assigns five processing roles (Figure 1). The normalization agent A_N maps timestamps, sources, affected assets, accounts, and observables into a common schema. The enrichment agent A_E retrieves asset and identity information, alert history, allowlist status, and supporting logs. The reasoning agent A_R uses this bundle to propose disposition, severity, confidence, and justification. The analyst-interaction agent A_H requests clarification when required. Finally, A_S summarizes the evidence, recommendation, uncertainty, and any analyst feedback.

These are functional roles in a lightweight workflow; their separation does not imply independently trained models. The distinction matters because the pilot evaluates orchestration, evidence preparation, and gating together. It does not establish that a particular number of LLM instances is necessary for the reported behavior.

3.3 Bounded Autonomy and Analyst Authority

Let r denote the intermediate reasoning result. The gate $G(r)$ distinguishes evidence sufficient for preparing a recommendation from evidence requiring clarification:

$$c = \begin{cases} A_S(r), & G(r) = 1, \\ A_S(r, h^*), \quad h^* = A_H(r), & G(r) = 0. \end{cases} \tag{2}$$

Missing, contradictory, or insufficient evidence triggers the clarification path. A Suspicious disposition preserves uncertainty when the available case context does not justify a definitive benign or malicious verdict. Passing the gate authorizes summarization, not autonomous closure or escalation. The analyst reviews the final case before any official SOC outcome is recorded.

The prototype logic uses a webhook for alert ingestion, normalization into the common schema, contextual tool outputs, a reasoning stage, a clarification checkpoint, and final summarization. Case packets provide a controlled representation of those tool outputs during evaluation. This separates the study of workflow behavior from the availability or latency of live SIEM, EDR, identity, inventory, and ticketing connectors. Operational integration would need to retain evidence provenance and enforce the same analyst approval boundary.

Table 1: Composition of the 15-case seed corpus.

Reference Class	Case IDs	Triage Purpose
Benign (4)	C01, C02, C08, C10	Avoid unnecessary escalation.
Suspicious (8)	C04, C06, C07, C09, C11, C12, C14, C15	Preserve uncertainty; seek clarification.
Malicious (3)	C03, C05, C13	Escalate supported harmful activity.

4 Experimental Setup

4.1 Dataset and Reference Outcomes

The controlled seed corpus contains 15 cases (C01–C15) constructed from public AIT-ADS/AIT-LDSv2-related alert material, including records from the *fox* scenario [8, 9, 10, 11]. AIT-ADS contains alerts generated from simulated enterprise-network logs; the underlying resources include benign activity and multi-step attacks. The case corpus is a selected triage exercise, not a random sample of the full datasets.

Each case packet contains a trigger alert, timestamp, affected host or asset, account or user, source address where available, alert description, log location, and supporting context. The scoring reference specifies disposition, severity, and required evidence. Case-level Benign, Suspicious, and Malicious outcomes reflect triage-oriented judgment; they should not be confused with the original datasets' event-level attack labels.

Table 1 lists the composition. Benign cases represent expected activity, such as maintenance or routine access. Suspicious cases contain security-relevant but inconclusive evidence and test whether clarification is requested. Malicious cases represent harmful or strongly attack-relevant sequences requiring escalation. For example, a login attempt alone is not sufficient proof of compromise, whereas correlated remote access and privilege escalation may justify escalation when supported by the surrounding records.

4.2 Configurations and Execution Protocol

Three configurations process the same cases and supporting evidence. The *manual reference* follows careful analyst-style triage: inspect the alert and context, determine disposition and severity, and prepare a case note. The *single-agent baseline* performs that

task in one end-to-end AI step. The *multi-agent workflow* separates the roles in Figure 1 and includes an explicit clarification checkpoint.

All configurations use the same reference outcomes, required-evidence checklist, and output fields. Their outputs comprise a disposition, severity, evidence summary, clarification decision, and case note. Holding the evidence constant supports comparison of the complete workflows; it does not independently establish a causal benefit from enrichment, agent count, or gating alone. No production-timing claim is made: the pilot's time and workload values are controlled protocol estimates.

4.3 Evaluation Metrics

The unit of evaluation is one complete case. Table 2 reports seven metrics, grouped below by what they measure. Higher values are desirable for correctness, agreement, completeness, and clarification accuracy; lower values are desirable for false closures, estimated time, and effort.

Decision quality. Triage correctness is the percentage of the 15 cases whose recommended disposition exactly matches the reference Benign, Suspicious, or Malicious label. Severity agreement is the percentage whose severity exactly matches its reference. For either metric,

$$\text{Agreement} = 100\,\frac{\text{number of exact matches}}{15}. \tag{3}$$

Thus, 100% denotes 15/15 matches and 73.3% denotes 11/15. These are case-level agreement measures, not clustering scores or estimates of deployment-wide accuracy.

Evidence coverage (RQ2). Context completeness (CC) is the fraction of required evidence items correctly included in an output:

$$\text{CC} = \frac{\text{CEI}}{\text{TREI}}, \qquad 0 \le \text{CC} \le 1, \tag{4}$$

where CEI is the number of correctly included evidence items and TREI is the total number required by the scoring checklist. Mentioning an unrelated or unsupported item does not satisfy a required item. CC=1 means complete required-evidence coverage; it does not by itself prove that every generated statement is correct. Table 2 gives the reported aggregate coverage scores. Grounding and summary usefulness inform qualitative interpretation and are not additional numerical outcomes in that table.

Uncertainty handling (RQ3). Clarification accuracy is the percentage of cases for which the yes/no decision to request analyst input matches the reference need for clarification. It is scored separately from disposition: a plausible label does not compensate for failing to ask for missing information. False closure rate is the proportion of reference non-benign cases incorrectly recommended for benign closure. The corpus contains 11 such cases (eight Suspicious and three Malicious); zero observed false closures is not evidence of zero operational risk.

Estimated efficiency (RQ1). Time to triage is reported as the mean estimated minutes per case. Analyst effort score (AES) combines the pilot's manual-lookup (ML), documentation-effort (DE), and human-action (HA) components:

$$\text{AES} = \text{ML} + \text{DE} + \text{HA}. \tag{5}$$

AES is an unweighted, dimensionless workload proxy, not elapsed time or a validated cognitive-load scale. The table reports mean estimated AES per case. Time and AES are interpreted separately because they have different units. Their values compare workflows under the shared pilot protocol; they are not stopwatch measurements or a controlled study of analyst productivity.

Table 2: Aggregate pilot results on 15 cases. Arrows indicate the preferred direction. Time and AES are estimates.

Metric	Manual	Single	Multi
Triage correctness ↑	100%	73.3%	100%
Severity agreement ↑	100%	100%	100%
Context completeness ↑	1.00	0.865	1.00
Clarification accuracy ↑	100%	73.3%	100%
False closure rate ↓	0%	0%	0%
Mean time (min/case) ↓	7.27	2.73	2.40
Mean AES (points/case) ↓	8.73	3.00	1.53

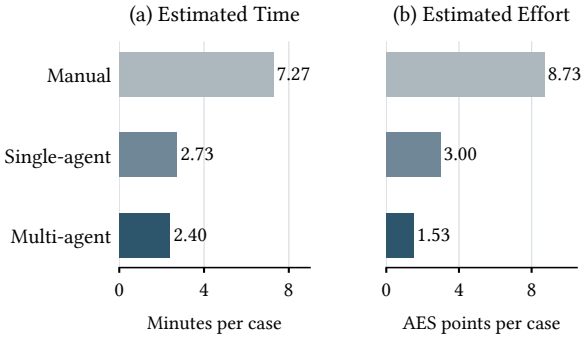

Figure 2: Pilot estimates of mean time and analyst effort per case. Separate axes preserve the distinction between minutes and AES points. Values reproduce Table 2.

4.4 Scoring Scope and Reproducibility

The shared packets, references, and output schema reduce differences in evidence availability and scoring criteria. They do not remove subjectivity in selecting cases or assigning reference judgments. Exact replication also requires the case-level evidence counts, the effort-coding rubric, prompts, and model configuration; the aggregate scores alone do not determine these details. The workflow description, case-construction framework, and evaluation materials are available from the author upon reasonable request.

5 Results and Discussion

Table 2 summarizes the pilot's aggregate results. Manual and multi-agent triage each match all 15 reference dispositions, while the single-agent baseline matches 11. Severity agreement is 15/15 for all three configurations and therefore does not distinguish them in this corpus.

Estimated efficiency (RQ1). Relative to the manual reference, the multi-agent workflow reduces mean estimated time from 7.27 to 2.40 minutes and mean AES from 8.73 to 1.53. Using $100(M-P)/M$, where M is the manual value and P the proposed value, the reductions are 67.0% and 82.5%, respectively. Relative to the single-agent baseline, the corresponding reductions are 12.1% and 49.0%. Figure 2 separates time and effort into panels to avoid treating minutes and proxy points as interchangeable measurements.

Evidence coverage (RQ2). Reported CC is 1.00 for manual and multi-agent triage and 0.865 for single-agent support. This is consistent with structured enrichment helping the final case

retain required evidence. Because all configurations have access to the same context, the result concerns how that context is processed and presented, rather than access to a richer dataset. Coverage should not be interpreted as an independent assessment of all factual claims in the generated summaries.

Uncertainty handling (RQ3). Clarification decisions match the reference in 15/15 manual and multi-agent cases and 11/15 single-agent cases. The pilot attributes the single-agent disagreements to overcommitment on suspicious cases instead of seeking clarification. All configurations have zero reported false closures, so that metric provides no comparative advantage here. The clarification result supports explicit uncertainty handling within this corpus, while analyst review remains necessary in every configuration.

Overall, the findings are consistent with the proposed hypothesis: structured preparation and gating can improve the usefulness of a triage workflow without lowering observed disposition agreement. However, the three-configuration comparison cannot determine which component caused the differences, and the workload estimates require validation through observed analyst behavior.

6 Limitations and Conclusion

The 15 selected cases constitute a preliminary evaluation. Their composition, reference judgments, and simulated source environment limit representativeness. The Suspicious class depends on evidence sufficiency and analyst judgment; independent reviewers may disagree. Aggregate CC and AES provide limited visibility into variation across cases, and the effort proxy has not been calibrated against measured workload. The study does not report repeated model runs, confidence intervals, or an independent multi-analyst assessment. Perfect agreement on this corpus therefore cannot establish robustness or operational safety.

The contribution is an analyst-supervised architecture that makes evidence preparation, clarification, and decision authority explicit. The pilot suggests lower estimated workload and improved clarification behavior compared with single-agent support, while matching the manual reference on observed dispositions. Future work should expand the corpus and telemetry sources, release detailed scoring and execution artifacts, repeat model runs, and measure time and effort with independent analysts. Deployment should preserve evidence provenance and analyst approval of official outcomes.

References

[1] S. Tariq, M. Baruwal Chhetri, S. Nepal, and C. Paris. 2025. Alert Fatigue in Security Operations Centres: Research Challenges and Opportunities. *ACM Computing Surveys* 57, 9. https://doi.org/10.1145/3723158.

[2] F. Jalalvand, M. Baruwal Chhetri, S. Nepal, and C. Paris. 2024. Alert Prioritisation in Security Operations Centres: A Systematic Survey on Criteria and Methods. *ACM Computing Surveys* 57, 2. https://doi.org/10.1145/3695462.

[3] X. Wang et al. 2024. Combating alert fatigue with AlertPro: Context-aware alert prioritization using reinforcement learning for multi-step attack detection. *Computers & Security* 137, 103583. https://doi.org/10.1016/j.cose.2023.103583.

[4] J. Liu et al. 2022. Context2Vector: Accelerating security event triage via context representation learning. *Information and Software Technology* 146, 106856. https://doi.org/10.1016/j.infsof.2022.106856.

[5] H. Xu et al. 2025. Large Language Models for Cyber Security: A Systematic Literature Review. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3769676.

[6] N. Kshetri and J. Voas. 2025. Agentic Artificial Intelligence for Cyber Threat Management. *Computer* 58, 5, 86–90. https://doi.org/10.1109/MC.2025.3544797.

[7] M. Baruwal Chhetri, S. Tariq, R. Singh, F. Jalalvand, C. Paris, and S. Nepal. 2024. Towards Human-AI Teaming to Mitigate Alert Fatigue in Security Operations Centres. *ACM Transactions on Internet Technology* 24, 3, Article 12. https://doi.org/10.1145/3670009.

[8] M. Landauer, F. Skopik, and M. Wurzenberger. 2024. Introducing a New Alert Data Set for Multi-Step Attack Analysis. In *Proceedings of CSET 2024*. https://doi.org/10.1145/3675741.3675748.

[9] M. Landauer, F. Skopik, and M. Wurzenberger. 2023. *AIT Alert Data Set*. Zenodo. https://doi.org/10.5281/zenodo.8263181.

[10] M. Landauer et al. 2022. *AIT Log Data Set V2.0*. Zenodo. https://doi.org/10.5281/zenodo.5789064.

[11] M. Landauer et al. 2023. Maintainable Log Datasets for Evaluation of Intrusion Detection Systems. *IEEE Transactions on Dependable and Secure Computing* 20, 4, 3466–3482. https://doi.org/10.1109/TDSC.2022.3201582.