International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Retrieval-Augmented Detection: Grounding Intrusion Analysis in Historical Incident Corpora

Vishal B[1], Neha Patil[2]

Abstract

Modern security operations centers (SOCs) are saturated with alerts whose disposition depends less on the alert itself than on institutional memory: whether a comparable pattern has been seen before, what it was found to mean, and how it was resolved. That memory exists — in closed tickets, investigation notes, and analyst dispositions — but it is unstructured, weakly indexed, and effectively unavailable at triage time. This paper introduces Retrieval-Augmented Detection (RAD), an architecture in which retrieval over a historical incident corpus participates in the detection decision itself rather than only in post-hoc narration. RAD contributes four mechanisms: (i) a normalization stage that maps heterogeneous detector output into a common schema suitable for embedding; (ii) hybrid dense–sparse retrieval over analyst-adjudicated incident records with an explicit temporal-decay weighting that discounts precedent drawn from superseded infrastructure; (iii) an evidence-conditioned scoring function in which retrieved precedent modulates detection confidence and every generated assertion carries a provenance link to a specific incident identifier; and (iv) a retrieval-support threshold below which the system abstains rather than emitting an unsupported verdict. We distinguish RAD from prior retrieval-augmented work in IT operations, which applies retrieval after an incident is declared in order to explain it; RAD applies retrieval before disposition in order to decide it. We identify abstention calibration as the central open problem for safe deployment, together with the failure surface a precedent-based architecture inherits — corpus poisoning, precedent staleness, and the systematic bias of a corpus that records only what was previously detected. This paper presents the architecture and formal problem definition; empirical validation is left to future work.

Keywords: Intrusion detection, retrieval-augmented generation, security operations, alert triage, large language models, incident response, hallucination, provenance, abstention.

I. Introduction

A. The Triage Bottleneck

Network intrusion detection has been an active field for three decades, and the detection component is, in a narrow sense, solved: signature engines match known patterns reliably [1], and supervised and unsupervised learning methods identify statistical deviation from baseline [2], [3]. The unsolved component is disposition — deciding what a firing alert means and whether it warrants response.

The scale of the mismatch is well documented. Qualitative study of SOC practitioners reports that the overwhelming majority of alerts reaching analysts are ultimately closed as benign, and that the cognitive burden of that filtering is a primary driver of analyst attrition [4]. Automated provenance triage work has framed the same problem as one of ranking rather than detection: the security-relevant signal is present,

but buried [5]. Contextual sequence analysis of security events has shown that alerts interpreted in isolation lose most of their diagnostic value, and that the surrounding event context is what carries disposition-relevant information [6].

What each of these lines of work implies is that the missing input to triage is not a better detector. It is context — and specifically historical context. An analyst confronting an alert asks, in effect, a retrieval question: has this happened before, and what did it turn out to be?

B. Institutional Memory as an Unexploited Corpus

Organizations that have operated a SOC for any length of time possess a large, high-quality, and almost entirely unexploited corpus: the record of every prior investigation. Each closed incident ticket pairs a set of observed telemetry with an analyst-adjudicated verdict, a remediation, and frequently a mapping to an adversary technique taxonomy [7], [8]. This is, structurally, a labeled dataset of exactly the mapping triage requires — from observation to disposition — assembled at considerable cost and then archived.

It goes unused for a mechanical reason. The corpus is prose. Ticket text, analyst notes, and post-incident summaries are not indexed in a form that supports similarity search at alert time, and keyword search over them returns either nothing or everything. The knowledge is present and unreachable.

C. From Explanation to Decision

Retrieval-augmented generation (RAG) [9] provides the obvious mechanism for reaching it, and RAG has already been applied to IT operations. Prior work has proposed retrieval over historical incidents and knowledge bases to generate natural-language root-cause narratives, within a layered observability architecture combining telemetry ingestion, normalization, multimodal fusion, and a generative reasoning engine [10].

That work is positioned downstream of declaration. An incident has already been recognized as an incident; retrieval supplies context so that a language model can explain it. The contribution of the present paper is to move retrieval upstream of disposition, so that retrieved precedent is an input to the decision of whether an alert constitutes an incident at all. This relocation changes the requirements substantially. A narrative generated after the fact is reviewed by an analyst who already has the incident in front of them and can discount a poor explanation. A disposition recommendation issued at triage time may determine whether an alert is ever reviewed at all, which makes ungrounded output not merely unhelpful but actively dangerous. Hallucination in language models is a general and persistent failure mode [11]; in a detection path it manifests as confident suppression of a true positive.

RAD is therefore designed around the constraint that the system must be able to decline. Retrieval does not only supply context; the quality of retrieval is itself a first-class signal, and when the corpus offers no adequate precedent the correct output is abstention with escalation rather than a verdict. Fig. 1 contrasts the two positions.

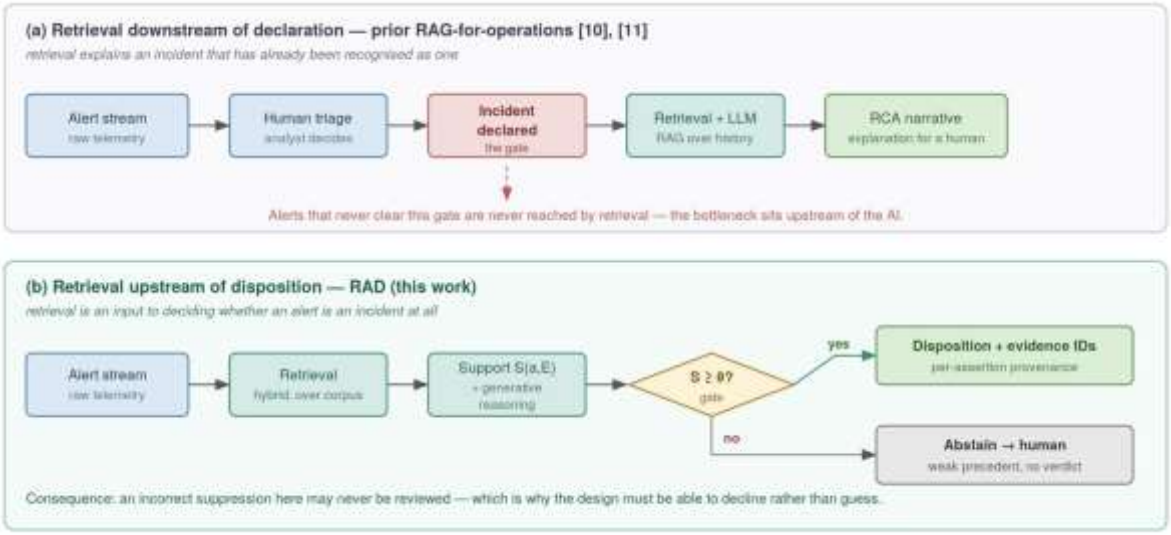

Fig. 1. Positioning of retrieval relative to the disposition decision. In (a), retrieval is invoked on an incident that a human has already declared, and alerts closed during triage are never reached by it. In (b), retrieval is invoked on every alert prior to disposition, which relocates the value but also removes the human review that (a) implicitly assumes.

D. Contributions

1. A formal problem definition for retrieval-conditioned alert disposition that treats retrieval support as an explicit term rather than an implicit preprocessing step (§III), together with the RAD architecture (§IV) that operationalizes it: schema normalization, incident-corpus construction, hybrid dense–sparse retrieval with temporal decay, evidence-conditioned scoring, and threshold-governed abstention.
2. A provenance discipline in which every assertion in generated analysis carries a link to the specific incident record supporting it, enabling per-assertion verification rather than whole-output trust (§IV-E).
3. A failure-surface analysis specific to retrieval in a security context (§VI): corpus poisoning, precedent staleness, and the systematic bias of a corpus that records only what was previously detected, together with the abstention-calibration problem this failure surface makes unavoidable.

II. Background and Related Work

A. Intrusion Detection and Its Evaluation

Signature-based detection remains the operational backbone of network security monitoring [1]. Machine-learning approaches have been proposed continuously since the early 2000s, but their operational adoption has lagged their publication record. Sommer and Paxson [2] identified the structural reasons: intrusion detection differs from other ML application domains in the extreme cost asymmetry between error types, the difficulty of obtaining representative training data, and the semantic gap between a statistical outlier and a security-relevant event. Arp et al. [12] catalogued recurring methodological faults in security ML — sampling bias, label inaccuracy, spurious correlation, and inappropriate performance measures — that inflate reported results and impede reproducibility.

RAD is designed with both critiques in view. The semantic gap identified in [2] is precisely what a corpus of analyst dispositions addresses: the corpus supplies the human judgment that converts statistical deviation into security meaning. The methodological cautions in [12] — particularly against temporal leakage and unqualified aggregate accuracy — apply directly to any future evaluation of this architecture.

B. Alert Triage and Contextual Analysis

NoDoze [5] approaches alert fatigue by scoring alert-associated provenance paths against historical frequency, propagating anomaly scores through a dependency graph to rank alerts by likely significance. HOLMES [13] correlates suspicious information flows into high-level kill-chain stages [7], producing a compact summary of attacker progression rather than a stream of primitive events. DEEPCASE [6] uses sequence modeling over security events to identify contextually similar event sequences and transfers analyst decisions across them semi-supervised.

DEEPCASE is the closest antecedent to RAD and the comparison is worth stating precisely. Both leverage prior analyst decisions and both operate on similarity between a current situation and past ones. DEEPCASE learns a fixed similarity model over event sequences and clusters them. RAD performs retrieval at query time over a corpus of full incident records — telemetry, prose notes, disposition, remediation — which admits heterogeneous and unstructured evidence that a sequence model cannot ingest, allows the corpus to be updated without retraining, and preserves the retrieved records as citable artifacts for analyst verification. The trade-off is that RAD inherits the failure modes of retrieval and of generative synthesis, which §VI treats at length.

C. Retrieval-Augmented Generation

RAG [9] couples a parametric generator with a non-parametric retrieval index, allowing knowledge to be updated without retraining and enabling attribution of output to source documents. Dense passage retrieval [14] established learned bi-encoder retrieval as competitive with and complementary to lexical methods such as BM25 [15]; hybrid retrieval combining both is now standard practice, as surveyed in [16]. Sentence-level encoders [17] provide efficient embedding for corpus-scale similarity search, and approximate nearest-neighbour indexing over hierarchical navigable small-world graphs [18] makes such search tractable at operational latency.

Two developments are directly relevant to RAD's abstention mechanism. REALM [19] and REPLUG [20] demonstrate that retrieval quality can be treated as a differentiable or at least tunable component of the overall system rather than a fixed preprocessing stage. Self-RAG [21] introduces reflective tokens by which a model assesses whether retrieved evidence supports its own output, providing a template for the support-scoring function of §IV-D. Automated RAG evaluation frameworks [22] supply the methodological template a grounding-fidelity metric for RAD's provenance discipline (§IV-E) would need to adopt.

D. Language Models in Operations and Security

Application of large language models to operational telemetry has proceeded rapidly. The four-layer generative observability architecture of [10] — ingestion, normalization, multimodal fusion, and a generative reasoning engine with agentic escalation on low confidence — provides the structural template that RAD adapts to the detection path, and its explicit treatment of hallucination risk, context-window limits, and the absence of human-operator baselines identifies the constraints that any operational deployment of generative reasoning must confront. Chain-of-thought prompting [23] and the transformer architecture underlying these models [24] are assumed as background. Predictive, AI-driven observability for IT operations is surveyed in [25].

The distinction RAD draws from [10] is architectural rather than incremental. In [10], the generative engine is invoked on a declared incident and produces an explanation; an agentic module gathers additional context when confidence falls below threshold. In RAD, retrieval is invoked on every alert prior to disposition, the retrieval support score is a term in the disposition decision, and low support produces

abstention rather than autonomous context-gathering — a deliberately more conservative response chosen because the detection path lacks the human-in-the-loop review that a declared incident already has.

III. Threat Model and Problem Formulation

A. Threat Model

We assume an enterprise network instrumented with one or more detectors producing an alert stream, and a SOC with a history of adjudicated investigations. The adversary is external or internal, and may employ techniques both represented and unrepresented in the historical corpus. We explicitly assume the adversary may have knowledge of the RAD system's existence and general design.

Two adversarial capabilities are in scope. First, evasion by novelty: an adversary who selects techniques absent from the corpus in order to obtain weak retrieval support and thereby a low-confidence disposition. RAD's abstention mechanism converts this from a suppression attack into an escalation, which is the intended behaviour, though it also constitutes a denial-of-service surface against analyst attention. Second, corpus poisoning: an adversary with the ability to influence incident records — for example by generating benign-appearing activity that is investigated and closed as benign, establishing precedent that a later malicious campaign retrieves. Both are analysed in §VI-B.

Out of scope: compromise of the detector layer itself, compromise of the embedding model supply chain, and adversarial manipulation of the language model through prompt injection embedded in telemetry, which we identify as important future work.

B. Formal Definition

Let a denote an alert, comprising a normalized feature vector and associated raw telemetry. Let $H = \{h_1, h_2, \ldots, h_n\}$ denote the historical incident corpus, where each record h_i is a tuple

$h_i = \langle \tau_i, d_i, \rho_i, t_i, q_i \rangle$

with τ_i the associated telemetry and analyst prose, d_i the adjudicated disposition, $\rho_i \in \mathscr{R}$ the remediation record, t_i the closure timestamp, and $q_i \in [0,1]$ a record-quality weight reflecting the depth of the original investigation.

Let $R_k(a, H) = E \subseteq H$, $|E| = k$, denote the retrieval function returning the k most relevant records. The disposition objective is

$d* = \arg\max_{d \in \mathcal{D}} P(d \mid a, E, c)$

where c is deployment context (asset criticality, business calendar, current threat posture).

This differs from the root-cause formulation of [10], which maximizes $P(C \mid T, S, R, I)$ over candidate causes given telemetry, components, relationships, and a declared incident I. Here no incident is presumed; $\mathcal{D}$ includes a benign disposition, and the evidence set E appears as an explicit conditioning term whose adequacy is separately assessed.

C. Retrieval Support and the Abstention Condition

Define the retrieval support score

$S(a, E) = \Sigma_{hi \in E}\, \mathrm{sim}(a, h_i) \cdot w(\Delta t_i) \cdot q_i$

where $\mathrm{sim}(\cdot,\cdot)$ is the hybrid similarity of §IV-C, q_i the record-quality weight, and $w(\Delta t_i)$ a temporal decay factor

$w(\Delta t_i) = \exp(-(\ln 2 / \lambda) \cdot \Delta t_i), \quad \Delta t_i = t_{now} - t_i$

with λ the precedent half-life: the interval over which a prior incident retains half its evidentiary weight. The rationale is that enterprise infrastructure changes continuously, and a two-year-old disposition may

describe a system topology that no longer exists. We treat λ as an environment-specific parameter to be calibrated per deployment rather than a universal constant.

The system emits a disposition only where S(a, E) ≥ θ. Below threshold, RAD emits an abstention with the retrieved evidence attached and escalates for human review. The setting of θ is the principal safety parameter, trading verdicts issued on thin precedent against analysts flooded with abstentions; §VI-A argues that a single scalar is ultimately inadequate to that trade-off.

IV. The RAD Architecture

RAD comprises five stages, shown in Fig. 2. Stages A and B are offline or incremental; stages C through E execute per alert.

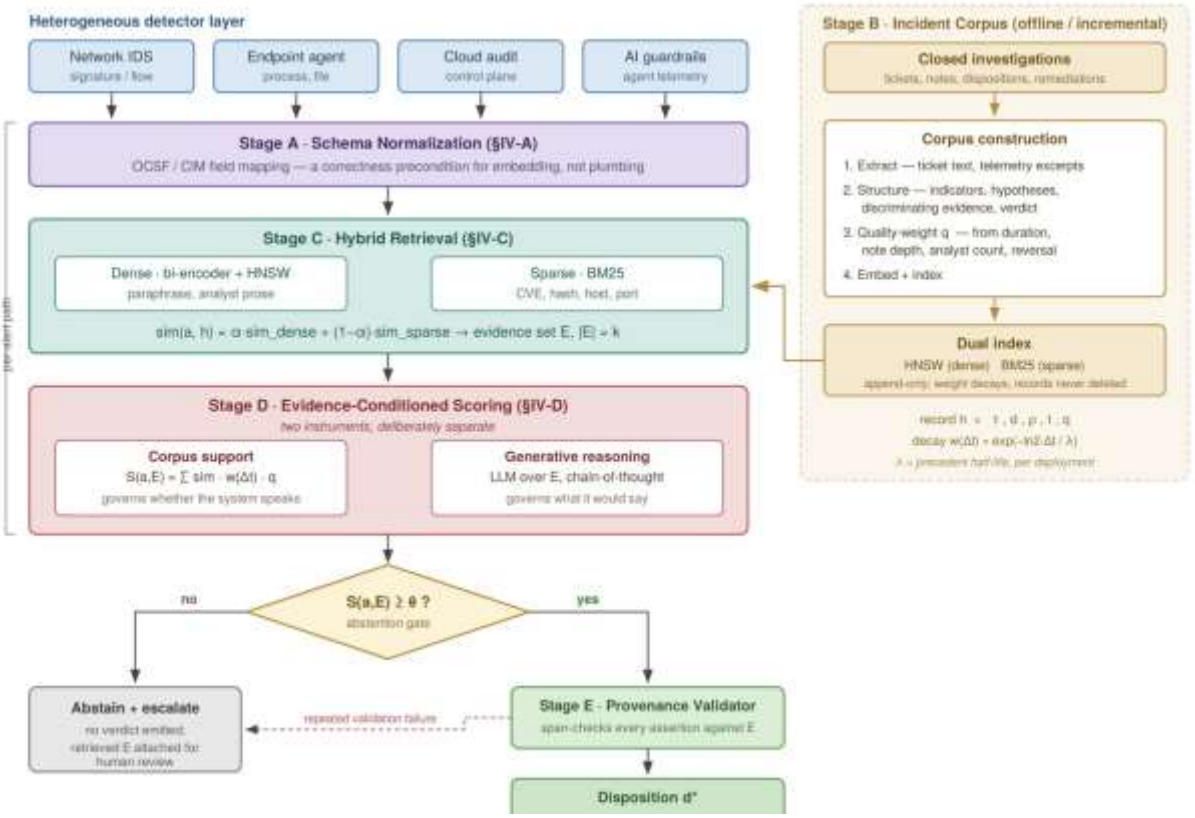

Fig. 2. RAD reference architecture. The corpus path (right) is maintained offline; the alert path (left) executes per alert. The two boxes in Stage D are deliberately separate instruments: the generator determines what the system would say, while the corpus support score S(a,E) determines whether it says it. A fluent generation over weak evidence cannot pass the gate on fluency alone.

A. Schema Normalization

Alerts arrive from heterogeneous detectors — network IDS, endpoint agents, cloud audit logs, application-layer guardrails — with incompatible field naming and semantics. RAD normalizes to a common security schema. Two candidates are appropriate: the Common Information Model used in Splunk deployments, and the Open Cybersecurity Schema Framework [26], which is vendor-neutral and increasingly adopted. The reference implementation targets both.

Normalization is not merely plumbing. Embedding quality depends directly on field consistency: an embedding model presented with src_ip in one record and source.address in another will encode a spurious distinction. Normalization is therefore a correctness requirement of the retrieval stage, not a convenience.

The layered treatment of ETL and semantic enrichment in [10] applies directly here, with the difference that RAD's normalization target is a security-specific schema rather than a general observability one.

B. Incident Corpus Construction

Each closed investigation yields one corpus record. Construction involves four steps.

Extraction. Ticket text, analyst notes, attached telemetry excerpts, and final disposition are extracted from the case management system.

Structuring. Free-text notes are parsed into a semi-structured form: observed indicators, hypotheses considered, evidence that discriminated among them, and final verdict. Where the organization maps incidents to an adversary technique taxonomy [8], those mappings are retained as filterable metadata.

Quality weighting. Not all closed tickets represent equal evidence. A ticket closed in ninety seconds with the note "false positive" and a ticket closed after a four-hour investigation with documented reasoning should not carry equal weight in S(a,E). We assign q_i from observable proxies — investigation duration, note length, number of analysts involved, whether the disposition was later reversed — and treat the resulting weight as an explicit and auditable quantity rather than a hidden heuristic.

Embedding and indexing. Records are embedded with a sentence-level encoder [17] and indexed for approximate nearest-neighbour search [18], with a parallel lexical index [15] supporting the sparse component of §IV-C.

The corpus is append-only with respect to history but its weights are time-varying through $w(\Delta t)$, so no record is ever deleted for age; it simply decays.

C. Hybrid Retrieval

Purely dense retrieval underperforms on the exact-match content that dominates security telemetry — specific CVE identifiers, hostnames, hash values, port numbers — because embedding models compress precisely the distinctions that matter. Purely lexical retrieval fails on the paraphrase and conceptual similarity that analyst prose exhibits. RAD uses a convex combination:

$$\mathrm{sim}(a, h_i) = \alpha \cdot \mathrm{sim}_{\mathrm{dense}}(a, h_i) + (1 - \alpha) \cdot \mathrm{sim}_{\mathrm{sparse}}(a, h_i)$$

with $\mathrm{sim}_{\mathrm{dense}}$ the cosine similarity of encoder representations [14], [17] and $\mathrm{sim}_{\mathrm{sparse}}$ a BM25 score [15] over the normalized field text. The mixing parameter α is intended to be tuned on a held-out development split, and we expect its optimal value to vary with the lexical richness of a given organization's incident notes rather than to transfer as a fixed constant.

Retrieval is filtered by hard constraints before scoring: asset class compatibility, and where available, technique-taxonomy category. This prevents high-similarity retrieval across semantically incompatible contexts — for instance, matching a database server alert to a precedent involving a developer workstation.

D. Evidence-Conditioned Scoring

The retrieved set E conditions disposition in two distinct ways, and the separation is deliberate.

First, E enters the generative reasoning step as context, allowing the model to produce a disposition recommendation informed by precedent. Second, and independently, S(a, E) enters as a scalar governing whether that recommendation is emitted at all. Keeping these separate prevents the failure mode in which a fluent, confident generation over weak evidence is mistaken for a well-supported conclusion — the language model's expressed confidence and the corpus's actual support are measured by different instruments.

The generation step is prompted to produce a structured output containing: recommended disposition, a reasoning chain [23], a per-assertion evidence mapping (§IV-E), a set of discriminating observations that would change the recommendation, and an explicit statement of what the retrieved precedent does not

cover. The last field is unusual and we consider it important: it converts the model's blind spot from an invisible property into a reviewable output.

Two parameters govern this stage: the precedent half-life λ (an exact decay function, §III-C) and the abstention threshold θ, which trades verdicts on thin precedent against analysts flooded with abstentions — two failure regimes that are not commensurable, which is why §VI-A treats their reconciliation as open rather than solved.

E. Provenance Discipline

Every factual assertion in the generated analysis must carry a reference to the record identifier supporting it. Assertions that cannot be so linked are marked as model inference rather than retrieved fact.

This is enforced structurally rather than requested rhetorically. The output schema requires an evidence array on each assertion; a post-generation validator rejects outputs in which an assertion claims support from a record identifier not present in E, or in which the claimed supporting text does not appear in the referenced record. Rejected outputs trigger regeneration, and repeated rejection triggers abstention.

The purpose is to make verification cheap. An analyst reviewing a RAD output need not evaluate the whole narrative for plausibility; they can spot-check individual assertions against named records. This shifts the review burden from a judgment about a language model's general reliability to a series of local, checkable claims.

V. Implementation

The reference implementation is specified as follows. Values are stated for reproducibility and are not claimed to be optimal.

Component	Selection	Notes
Normalization	OCSF v1.x mapping layer [26]	CIM mapping maintained in parallel
Embedding	Sentence-transformer bi-encoder [17]	384-dim; domain adaptation left for future evaluation
Dense index	HNSW [18]	M = 32, ef_construction = 200
Sparse index	BM25 [15]	k_1 = 1.2, b = 0.75
Generator	Instruction-tuned open-weight LLM	Local inference; no telemetry leaves the boundary
Validator	Deterministic schema + span checker	Rejects unsupported assertions (§IV-E)
Store	Vector DB + relational store	Records retained for audit

Local inference is a requirement rather than a preference: security telemetry frequently contains material that cannot cross an organizational boundary, and the regulatory exposure of transmitting it to a third-party inference endpoint is prohibitive in the regulated environments where SOCs are most mature. This constraint bounds the model scale available and is a real limitation on achievable reasoning quality; §VI-C addresses the resulting trade-off.

Latency budget is set at the alert arrival rate rather than at human reaction time. For a deployment receiving r alerts per second, per-alert retrieval and generation must complete within 1/r seconds amortized, or the

system requires batching or selective invocation. We anticipate selective invocation — RAD applied only to alerts that pass a cheap pre-filter — as the practical deployment pattern; the pre-filter's own error contribution would need to be measured separately, since it bounds what RAD can see before RAD's own accuracy is even assessed.

VI. Discussion and Limitations

A. Abstention Calibration Is the Central Problem

The threshold θ trades two error modes that are not commensurable. A high θ produces frequent abstention, returning the alert volume to human analysts and forfeiting the system's purpose. A low θ produces confident dispositions on thin evidence, and because those dispositions are consumed as triage decisions, an incorrect suppression may never be reviewed.

We do not believe a single scalar threshold θ is ultimately adequate. A more defensible design conditions on asset criticality — demanding stronger precedent before dispositioning an alert on a payment system than on a test host. We identify criticality-conditioned abstention as the most important open problem arising from this work and do not claim to have solved it.

B. Corpus-Specific Failure Modes

Survivorship bias. The corpus records only what was previously detected and investigated. Attack techniques that historically evaded detection are absent from it, and RAD will find no precedent for them. This is not a defect of the retrieval mechanism but a property of the evidence, and it means RAD structurally cannot improve detection of never-before-seen technique classes. It improves disposition of alerts on techniques with precedent. Conflating these would substantially overstate the contribution.

Error propagation. A misadjudicated incident becomes precedent. If an intrusion was historically closed as benign, RAD will retrieve that closure and reproduce the error, with the added harm that the error now carries apparent evidentiary support. Quality weighting q_i partially mitigates this where reversals are recorded, but a consistently missed technique produces consistent and confidently wrong precedent. Periodic corpus audit against retrospective threat intelligence is necessary and is not automated by this design.

Poisoning. Per §III-A, an adversary able to generate activity that is investigated and closed as benign can plant precedent. The temporal-decay mechanism modestly raises the cost of this attack — planted precedent decays and must be refreshed — but does not defeat a patient adversary. Detecting anomalous corpus growth patterns is a plausible countermeasure and is unevaluated here.

C. Model and Scale Constraints

Local inference bounds model capability. The reasoning quality achievable from an open-weight model that fits a SOC's hardware budget is materially below that of frontier hosted models, and this is a real ceiling on RAD's synthesis quality — the same constraint identified in [10] with respect to context-window limits and computational resource requirements. Where regulatory posture permits hosted inference, RAD's architecture is unchanged and its quality should improve; we simply cannot assume that posture.

D. What This Paper Does Not Establish

This paper contributes an architecture and a formal treatment of retrieval support and abstention. It does not specify an evaluation protocol, does not report empirical validation, does not establish that RAD outperforms existing triage automation, and does not demonstrate operation on a genuine SOC corpus. Readers should treat the contribution as a specification and a research agenda: the architectural claims of

§IV, the failure modes of §VI-A and §VI-B, and the abstention-calibration problem of §VI-A are testable, but none has been tested here. We regard the honest statement of this scope as more useful than a proof-of-concept result over derived data presented as validation.

VII. Future Work

1. Evaluation on genuine SOC corpora under data-use agreement, which is the single highest-value next step and the only route to establishing external validity.
2. Criticality-conditioned abstention (§VI-A), replacing the scalar θ with a policy over asset and business context.
3. Prompt-injection resistance. Telemetry is attacker-influenced text that enters a language model's context. This is an obvious and currently unaddressed attack surface, and we consider it the most urgent security question raised by the design.
4. Corpus audit and sharing. Detecting historically misadjudicated records by replaying current threat intelligence against closed cases, and cross-organizational precedent sharing to mitigate the survivorship-bias problem of §VI-B — one organization's detected technique is another's blind spot.

VIII. Conclusion

Alert triage is a retrieval problem that has been treated as a classification problem. The knowledge required to disposition most alerts already exists inside the organization, recorded in prior investigations, and is unavailable at the moment of decision only because it was never indexed for that use.

RAD proposes to close that gap by moving retrieval upstream of disposition rather than downstream of incident declaration, and by treating the adequacy of retrieval as a measured quantity that governs whether the system speaks at all. The architecture's distinguishing commitments are provenance at the level of individual assertions and abstention as a first-class output. Both follow from the same observation: a system that participates in detection decisions is not adequately constrained by being usually right, because the cases where it is wrong are precisely the cases no one will review.

We present the architecture and formal problem definition, and identify abstention calibration, prompt-injection resistance, and empirical validation against genuine incident corpora as the work that must follow before deployment can be responsibly recommended.

References

1. M. Roesch, "Snort — Lightweight intrusion detection for networks," in Proc. 13th USENIX Conf. Syst. Admin. (LISA), Seattle, WA, USA, 1999, pp. 229–238.
2. R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symp. Security and Privacy (S&P), Oakland, CA, USA, 2010, pp. 305–316.
3. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surveys, vol. 41, no. 3, pp. 1–58, Jul. 2009.
4. B. A. Alahmadi, L. Axon, and I. Martinovic, "99% false positives: A qualitative study of SOC analysts' perspectives on security alarms," in Proc. 31st USENIX Security Symp., Boston, MA, USA, 2022, pp. 2783–2800.
5. W. U. Hassan, S. Guo, D. Li, Z. Chen, K. Jee, Z. Li, and A. Bates, "NoDoze: Combatting threat alert fatigue with automated provenance triage," in Proc. Netw. Distrib. Syst. Security Symp. (NDSS), San

Diego, CA, USA, 2019.
6. T. van Ede, H. Aghakhani, N. Spahn, R. Bortolameotti, M. Cova, A. Continella, M. van Steen, A. Peter, C. Kruegel, and G. Vigna, "DEEPCASE: Semi-supervised contextual analysis of security events," in Proc. IEEE Symp. Security and Privacy (S&P), San Francisco, CA, USA, 2022, pp. 522–539.
7. E. M. Hutchins, M. J. Cloppert, and R. M. Amin, "Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains," in Proc. 6th Int. Conf. Inf. Warfare and Security, 2011, pp. 113–125.
8. B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, and C. B. Thomas, "MITRE ATT&CK: Design and philosophy," MITRE Corporation, Bedford, MA, USA, Tech. Rep. MP180360R1, 2018.
9. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. 34th Conf. Neural Inf. Process. Syst. (NeurIPS), 2020, pp. 9459–9474.
10. Bisht, K. S. Generative AI-Driven Observability for Automated Root Cause Analysis in Modern IT Systems: Architecture and Vision. Journal of Computer Science and Technology Studies, 7(9), 549-560. https://doi.org/10.32996/jcsts.2025.7.9.63
11. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," ACM Comput. Surveys, vol. 55, no. 12, pp. 1–38, Dec. 2023.
12. D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, and K. Rieck, "Dos and don'ts of machine learning in computer security," in Proc. 31st USENIX Security Symp., Boston, MA, USA, 2022, pp. 3971–3988.
13. S. M. Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, and V. N. Venkatakrishnan, "HOLMES: Real-time APT detection through correlation of suspicious information flows," in Proc. IEEE Symp. Security and Privacy (S&P), San Francisco, CA, USA, 2019, pp. 1137–1152.
14. V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 6769–6781.
15. S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.
16. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2023.
17. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Hong Kong, China, 2019, pp. 3982–3992.
18. Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 4, pp. 824–836, Apr. 2020.
19. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-augmented language model pre-training," in Proc. 37th Int. Conf. Mach. Learn. (ICML), 2020, pp. 3929–3938.
20. W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W. Yih, "REPLUG: Retrieval-augmented black-box language models," arXiv preprint arXiv:2301.12652, 2023.

21. A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in Proc. Int. Conf. Learn. Representations (ICLR), 2024.
22. S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL): System Demonstrations, 2024, pp. 150–158.
23. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 24824–24837.
24. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. 31st Conf. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.
25. Bisht, K. S. (2025). Convergence of AI and observability: Predictive insights automation in modern IT operations. Journal of Computer Science and Technology Studies, 7(4), 446–454. https://doi.org/10.32996/jcsts.2025.7.4.53
26. Open Cybersecurity Schema Framework, "OCSF schema specification." [Online]. Available: https://schema.ocsf.io/