

Adaptive Cybersecurity Threat Intelligence Retrieval Through Collaborative Language Model Agents and Risk-Aware Ranking

Mart Woutersz

School of Computing, Clemson University, Clemson, SC, USA.
hellomart@clemson.edu

Benthe Kort

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis,
OR, USA.
benthe.work@oregonstate.edu

Ola Bertilsson

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
olabertilsson@missouri.edu

Abstract

Cybersecurity organizations increasingly operate within an intelligence environment characterized by volume, heterogeneity, and rapid obsolescence. Vulnerability disclosures, vendor advisories, incident reports, malware analyses, and regulatory notices arrive through multiple channels, while analysts must decide which items deserve immediate attention and which can be deferred. Conventional retrieval systems prioritize lexical or semantic relevance, but relevance alone does not express operational risk. This paper examines an adaptive threat intelligence retrieval architecture that combines collaborative language model agents with risk-aware ranking. The proposed system-level perspective treats retrieval as a socio-technical process involving ingestion, normalization, multi-agent verification, ranking, human oversight, and continuous governance. Collaborative agents assume differentiated roles such as collection, triage, validation, synthesis, and audit, while ranking mechanisms incorporate asset criticality, exploitability, exposure, temporal decay, provenance, and confidence. The paper analyzes structural trade-offs concerning latency, cost, explainability, fairness, robustness, and sustainability. It also considers deployment challenges across cloud, edge, and hybrid infrastructures, including technical debt, observability, and model update cycles. Particular attention is given to adversarial manipulation, data poisoning, prompt injection, and bias in prioritization. The discussion concludes that adaptive threat intelligence retrieval should not be framed as a purely algorithmic problem. Instead, it requires layered governance, calibrated uncertainty, auditable provenance, resilient infrastructure, and policy frameworks that align machine-assisted prioritization with institutional accountability.

Keywords

adaptive retrieval; cybersecurity threat intelligence; collaborative language model agents; risk-aware ranking; socio-technical governance; infrastructure sustainability; robustness; fairness.

1. Introduction

Contemporary cybersecurity threat intelligence is no longer scarce. It is abundant, fragmented, and unevenly trustworthy. Security teams receive vulnerability disclosures, vendor patches, government advisories, commercial feeds, open-source intelligence, malware reports, and incident postmortems through systems that rarely share a common representation. The result is not simply information overload but a prioritization crisis. Analysts must decide which threats matter to specific assets, missions, and jurisdictions under severe time constraints. Frameworks such as the NIST Cybersecurity Framework emphasize risk management and continuous improvement, yet operational workflows often remain manual and fragmented [1]. Regional threat landscapes further complicate prioritization because adversaries, regulatory expectations, and infrastructure dependencies vary across sectors and geographies [2]. Vulnerability catalogs such as the Known Exploited Vulnerabilities Catalog provide authoritative signals, but they still require interpretation within local contexts [20].

Large language models have introduced new possibilities for summarization, question answering, and reasoning over unstructured text [3], [4]. Retrieval-augmented generation can ground model outputs in external corpora and reduce reliance on parametric memory [5]. Dense retrieval methods improve semantic matching beyond lexical overlap [6], while generative readers can synthesize answers from retrieved passages [7]. These advances suggest that threat intelligence retrieval could become more adaptive, conversational, and context-sensitive. However, the same capabilities introduce new failure modes. Language models may hallucinate, overgeneralize from outdated sources, amplify biased rankings, or be manipulated by adversarial content. A system that retrieves and ranks threat intelligence must therefore be designed as an accountable infrastructure rather than as a standalone model.

This paper advances a system-level argument: adaptive cybersecurity threat intelligence retrieval should be organized around collaborative language model agents and risk-aware ranking. Collaborative agents can divide cognitive labor, cross-check claims, and preserve provenance, while risk-aware ranking can align retrieval outputs with operational consequence rather than mere topical similarity. The discussion emphasizes architecture, governance, deployment, sustainability, robustness, fairness, and policy. The central claim is that no single model, benchmark, or ranking function can resolve the underlying tension between speed, accuracy, explainability, and accountability. Instead, adaptive retrieval requires layered design choices that make uncertainty visible and allow human institutions to intervene.

2. Background and Related Work

Threat intelligence has evolved from static signatures and perimeter alerts toward continuous, heterogeneous, and often adversarial data streams. Standards and frameworks provide vocabulary for risk, resilience, and controls [1], while threat landscape reports describe shifting attacker tradecraft and sectoral exposure [2]. Vulnerability exploitation signals, such as known exploited vulnerability catalogs, introduce temporal urgency into prioritization [20]. These sources are necessary but insufficient. They must be integrated with internal telemetry, asset inventories, business impact assessments, and regulatory obligations. Retrieval systems therefore sit between external intelligence and internal decision-making.

Information retrieval research offers foundational mechanisms for matching queries to documents. Probabilistic relevance models and their descendants remain important because they are interpretable, efficient, and robust in many operational settings [8]. Dense passage retrieval learns continuous representations that can capture semantic similarity and improve recall for paraphrase-heavy queries [6]. Retrieval-augmented generation combines retrieval

with language model generation, enabling systems to answer questions using retrieved evidence [5]. Later work has refined how retrieved passages condition generative outputs [7]. These approaches are highly relevant to threat intelligence, where analysts often ask natural-language questions about vulnerabilities, actors, tactics, and mitigations.

The emergence of language model agents extends retrieval from a single-step query-response pattern to multi-step reasoning and tool use. Multi-agent frameworks enable specialized roles, conversation, and coordination among language model instances [12]. Reasoning-and-acting approaches interleave thought with external actions, allowing agents to search, inspect, and update their plans [13]. Reflection mechanisms allow agents to critique and revise prior attempts, which can improve reliability in complex tasks [14]. Self-consistency methods sample multiple reasoning paths and select more stable answers, offering a way to estimate confidence and reduce idiosyncratic errors [10]. These techniques are promising for threat intelligence because analytical judgments often require corroboration across sources rather than a single retrieval event.

Ranking and reranking remain central to adaptive retrieval. A recent multi-role collaborative reranking workflow demonstrates how language model agents can assume complementary roles to improve ranking outcomes [11]. Such work highlights that ranking is not merely a final scoring step but a collaborative sense-making process. In cybersecurity, however, ranking must incorporate risk dimensions that general web search does not usually prioritize. An item may be semantically relevant but operationally low risk, or it may be terse but extremely urgent because it concerns an exploited vulnerability in an internet-facing asset. Risk-aware ranking therefore requires a different objective from conventional relevance ranking.

3. System Architecture for Collaborative Language Model Agents

A practical adaptive threat intelligence retrieval architecture can be understood as a layered socio-technical pipeline. The ingestion layer collects advisories, vulnerability records, vendor notifications, research posts, internal incident tickets, and telemetry summaries. The normalization layer resolves entity identities, timestamps, affected products, versions, and geographic or sectoral relevance. The indexing layer combines lexical indexes, dense vector stores, and structured metadata stores. The retrieval layer accepts natural-language queries, analyst profiles, asset context, and organizational policies. The collaborative reasoning layer coordinates multiple language model agents. The ranking layer produces a risk-aware ordering. The governance layer records provenance, access, decisions, and feedback. This architecture is not a linear sequence but a set of feedback loops in which human corrections and operational outcomes reshape future retrieval.

Collaborative agents can assume differentiated responsibilities. A collection agent monitors sources and detects changes. A triage agent classifies incoming items by threat type, urgency, and affected technology. A validation agent cross-checks claims against authoritative sources and flags contradictions. A synthesis agent produces concise briefings with citations. A risk assessment agent estimates potential impact using asset criticality and exposure. An audit agent reviews decisions for policy compliance and bias. Multi-agent conversation frameworks provide a conceptual basis for such coordination [12]. Reasoning-and-acting patterns allow agents to call search tools, query databases, and update their hypotheses [13]. Reflection and self-critique can help agents identify weak evidence before outputs reach analysts [14]. These mechanisms do not eliminate error, but they can make error more visible and correctable.

Retrieval within this architecture should be adaptive rather than fixed. A query about a newly disclosed vulnerability may require immediate retrieval of vendor patches, exploitation evidence, and internal exposure data. A query about a long-running campaign may require broader historical retrieval and actor profiling. Dense retrieval can support semantic matching across heterogeneous reports [6], while generative synthesis can produce readable answers from retrieved evidence [5], [7]. However, the system must distinguish between evidence retrieval and inference. Retrieved passages should remain separately inspectable, and generated conclusions should be traceable to specific sources. This separation is essential for auditability and for preventing fluent but unsupported claims from entering operational decisions.

Ranking in this architecture should be collaborative and risk-aware. A multi-role reranking workflow suggests that multiple language model agents can critique and refine an initial ordering [11]. In cybersecurity, the initial ranking might combine lexical relevance, semantic similarity, source authority, and recency. The collaborative reranking stage might then weigh exploitability, asset criticality, exposure, mitigation availability, and confidence. Self-consistency can be used to estimate whether agents converge on similar risk judgments [10]. Human analysts should retain the ability to override rankings, annotate rationales, and adjust thresholds. The architecture should also support degraded modes, such as returning only authoritative source matches when language model services are unavailable or untrusted.

4. Risk-Aware Ranking and Retrieval Governance

Risk-aware ranking differs from relevance ranking because it asks not only whether an item matches a query but whether it warrants action. In cybersecurity, action depends on consequence, likelihood, timing, and controllability. A vulnerability in an isolated legacy system may be relevant but less urgent than a vulnerability in an externally exposed authentication service. NIST guidance provides a risk management vocabulary that can inform such judgments [1], while threat landscape reports supply contextual likelihood estimates [2]. Known exploited vulnerability signals can raise urgency, but they must be interpreted alongside local exposure and compensating controls [20]. A risk-aware ranker therefore combines multiple dimensions, including asset criticality, exploitability, exposure, temporal decay, provenance strength, and confidence.

Governance is necessary because risk scoring is never purely technical. The choice of weights, thresholds, and features reflects institutional priorities. If the system consistently ranks threats affecting one business unit, vendor ecosystem, or geographic region above others, it may create structural blind spots or inequities. Provenance governance requires that every ranked item can be traced to its sources, retrieval history, model transformations, and human edits. Confidence governance requires that uncertainty be represented in a way analysts can understand, not hidden behind a single score. Calibration is especially difficult when language model agents produce fluent but variable outputs. Self-consistency and cross-agent agreement can provide partial signals [10], but they should not be treated as guarantees of correctness.

Retrieval governance also involves lifecycle management. Threat intelligence changes rapidly, so indexes must be updated, stale documents must be down-ranked or archived, and contradictory sources must be surfaced rather than silently resolved. Benchmarks such as BEIR highlight the difficulty of zero-shot retrieval across heterogeneous domains [9], and threat intelligence is more adversarial and time-sensitive than many benchmark collections. Governance should therefore include periodic evaluation against expert-curated cases, red-team exercises, and operational outcomes. Feedback loops should reward accurate

prioritization and penalize missed critical items, while protecting against feedback loops that simply reinforce existing analyst preferences.

5. Deployment, Infrastructure, and Sustainability

Deploying adaptive retrieval systems requires infrastructure that spans data ingestion, vector indexing, model serving, orchestration, monitoring, and access control. Cloud deployments offer elasticity and managed services, but they raise concerns about data residency, third-party exposure, and cost. On-premises deployments provide greater control but demand specialized hardware and operations. Hybrid architectures may place sensitive telemetry and asset inventories locally while using external services for non-sensitive summarization or retrieval. The choice affects latency, resilience, and sustainability. Foundation models are resource-intensive to train and serve, and their environmental and financial costs must be considered alongside functional benefits [15]. Tiered model usage, caching, and selective invocation of large models can reduce waste.

Technical debt is a central deployment risk. Machine learning systems accumulate hidden dependencies, feedback loops, configuration drift, and entanglement between data, models, and code [18]. In threat intelligence retrieval, technical debt can manifest as undocumented prompt changes, inconsistent entity resolution, stale embeddings, and opaque ranking weights. Deployment research shows that many machine learning failures arise from organizational and infrastructure issues rather than model accuracy alone [19]. A sustainable architecture therefore requires observability, versioning, reproducible pipelines, and clear ownership. Model updates should be staged and evaluated against operational scenarios before full rollout.

Sustainability also has an institutional dimension. Analysts may resist systems that appear to replace judgment rather than augment it. If the system is too slow, too noisy, or too opaque, it will be bypassed. If it is too authoritative, it may induce automation bias. Deployment should therefore include training, transparent explanations, escalation paths, and mechanisms for reporting errors. A system that cannot be maintained, audited, or safely updated is not adaptive in any meaningful sense. It is a temporary technical artifact embedded in a long-lived security organization.

6. Robustness, Fairness, and Security Implications

Adaptive retrieval systems are vulnerable to adversarial manipulation. Attackers may seed false reports, poison public sources, manipulate timestamps, or craft content designed to influence language model agents. Prompt injection can cause agents to ignore instructions, leak context, or prioritize malicious items. Data poisoning can corrupt retrieval indexes or ranking behavior over time. Foundation models introduce broad risks related to misuse, opacity, and downstream harm [15]. Ethical and social risk analyses of language models identify concerns such as misinformation, bias, privacy, and concentration of power [16]. These risks are amplified in cybersecurity, where outputs can trigger urgent operational action.

Fairness concerns arise when ranking systems systematically under-prioritize certain technologies, sectors, languages, or regions. Threat intelligence is unevenly produced. Widely used vendors and English-language sources tend to dominate public reporting, while smaller organizations, local-language communities, and critical infrastructure in underrepresented regions may receive less attention. If a risk-aware ranker learns from historical analyst behavior, it may reproduce these imbalances. The dangers of large language models trained on uncensored data further complicate fairness and representation [17]. Governance should

therefore include bias testing, coverage audits, and deliberate mechanisms for surfacing underrepresented threats.

Security of the retrieval infrastructure itself is equally important. Agents that access internal telemetry, vulnerability databases, and incident systems become high-value targets. They require strong authentication, least-privilege access, segmentation, and comprehensive logging. The system should distinguish between public, internal, and restricted sources, and it should prevent cross-context leakage. Ranking explanations should be available without exposing sensitive asset details to unauthorized users. Standards and frameworks such as NIST provide a starting point for controls [1], while threat landscape analysis highlights the evolving adversary tradecraft that such systems must withstand [2]. Robustness is not a feature added after deployment; it must shape the architecture from the beginning.

7. Evaluation, Metrics, and Operational Validation

Evaluation of adaptive threat intelligence retrieval cannot rely solely on static relevance judgments. Benchmarks such as BEIR provide useful evidence about zero-shot retrieval performance across domains [9], but they do not capture operational urgency, exploitability, or organizational consequence. A system may achieve high semantic retrieval quality while missing the most operationally important item. Evaluation should therefore combine offline retrieval metrics, ranking quality assessments, risk-calibration checks, and human expert review. Collaborative reranking approaches can be evaluated not only by final ordering accuracy but also by the quality of role specialization, evidence use, and disagreement handling [11]. Self-consistency can be evaluated as a confidence signal rather than as a substitute for ground truth [10].

Operational validation should include scenario-based exercises, red-team simulations, and post-incident reviews. Analysts might evaluate whether the system surfaced relevant patches, exploitation evidence, and internal exposure data within a reasonable time. They might also assess whether explanations were understandable and whether overrides were easy to enact. Deployment research suggests that evaluation must include organizational workflows, not only model outputs [19]. Technical debt considerations imply that maintainability, monitoring, and update cost should be treated as evaluation criteria [18]. A system that performs well in a controlled demonstration but fails under degraded network conditions or sudden source changes is not operationally mature.

Metrics should be interpreted with caution. Precision and recall remain useful, but they can obscure asymmetries in consequence. A false positive may waste analyst time; a false negative may leave a critical asset exposed. Ranking metrics can summarize ordering quality, but they may not reflect temporal dynamics or source trust. Human trust is another important but difficult measure. Excessive trust can produce automation bias, while insufficient trust can lead to disuse. Evaluation should therefore examine how the system affects analyst decision-making, escalation behavior, and collective situational awareness.

8. Policy, Governance, and Cross-Domain Lessons

Policy frameworks for cybersecurity increasingly emphasize risk management, information sharing, and resilience [1]. Government catalogs and advisories provide authoritative signals that can anchor ranking [20]. Regional agencies produce threat landscape analyses that help organizations contextualize global trends [2]. However, policy guidance rarely specifies how machine-assisted prioritization should be governed. Questions remain about liability, transparency, appeal, and redress when automated rankings influence security decisions. If a

system systematically deprioritizes a class of vulnerabilities, who is accountable? If a language model agent produces a misleading synthesis, how should that error be recorded and remedied? These questions require governance mechanisms that extend beyond model documentation.

Foundation model governance offers relevant lessons. Broad analyses of foundation models emphasize both opportunity and concentrated risk [15]. Language model harm taxonomies identify concerns such as misinformation, bias, privacy, and malicious use [16]. Critical perspectives on large-scale language models warn against treating scale as a proxy for social benefit [17]. In threat intelligence retrieval, these concerns translate into requirements for provenance, contestability, and human oversight. Governance should define which decisions can be automated, which require human approval, and which must be escalated to incident command. It should also specify retention policies, audit trails, and access controls.

Cross-domain comparisons are instructive. Healthcare prioritization systems must balance urgency, equity, and liability. Financial fraud detection systems must manage adversarial adaptation and false positives. Critical infrastructure monitoring must handle safety, resilience, and regulatory reporting. Adaptive threat intelligence retrieval shares features with all three: it is adversarial, time-sensitive, consequence-laden, and institutionally embedded. The most transferable lesson is that ranking systems should not be optimized in isolation. They must be evaluated within workflows, governed by clear policies, and sustained by infrastructure that can evolve without accumulating hidden risk.

9. Conclusion

Adaptive cybersecurity threat intelligence retrieval through collaborative language model agents and risk-aware ranking offers a promising direction for managing intelligence abundance. Collaborative agents can divide analytical labor, cross-check evidence, and generate traceable syntheses, while risk-aware ranking can align retrieval with operational consequence. Yet the benefits depend on architecture and governance, not merely on model capability. Systems must preserve provenance, represent uncertainty, support human override, and remain robust against adversarial manipulation. They must also be deployable, maintainable, fair, and sustainable across heterogeneous organizations and jurisdictions. The paper has argued that adaptive retrieval should be treated as a socio-technical infrastructure. Its success will be measured not only by ranking accuracy but by whether it helps security institutions make faster, more defensible, and more equitable decisions under uncertainty.

References

1. NIST. (2018). Framework for improving critical infrastructure cybersecurity, Version 1.1. National Institute of Standards and Technology.
2. ENISA. (2023). ENISA threat landscape 2023. European Union Agency for Cybersecurity.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30.
4. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* 33.

5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* 33.
6. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). Association for Computational Linguistics.
7. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 874–880). Association for Computational Linguistics.
8. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.
9. Thakur, N., Reimers, N., Rüclé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.
10. Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.
11. Jin, Can, et al. "Rankflow: A multi-role collaborative reranking workflow utilizing large language models." *Companion Proceedings of the ACM on Web Conference 2025*. 2025.
12. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Awadallah, A., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.
13. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.
14. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems* 36.
15. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
16. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
17. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM*

Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery.

18. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems* 28.
19. Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys*, 55(6), Article 114.
20. CISA. (2024). Known exploited vulnerabilities catalog. Cybersecurity and Infrastructure Security Agency. <https://www.cisa.gov/known-exploited-vulnerabilities-catalog>