

Uncertainty-Aware RAG for Cybersecurity Threat Intelligence and Automated Incident Response

Eraldo Olivetti

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL,
USA.

eraldo.work@uab.edu

Abstract

Retrieval-augmented generation has emerged as a practical architecture for grounding large language models in external knowledge, yet its application to cybersecurity threat intelligence and automated incident response remains constrained by uncertainty, adversarial ambiguity, and high-stakes governance requirements. This paper examines uncertainty-aware retrieval-augmented generation as a socio-technical infrastructure for security operations centers, where generated recommendations must be accurate, traceable, calibrated, and actionable under incomplete and deliberately deceptive evidence. The discussion develops a system-level perspective that connects retrieval design, uncertainty estimation, policy enforcement, human oversight, deployment sustainability, and regulatory accountability. It argues that uncertainty-aware retrieval-augmented generation should not be treated merely as a model enhancement but as an institutional control layer that mediates between heterogeneous threat feeds, incident data, automated response mechanisms, and human decision-makers. The paper analyzes architectural trade-offs involving latency, coverage, calibration, explainability, and resilience. It also considers adversarial risks such as corpus poisoning, prompt injection, and evasion through fabricated indicators. Governance implications include auditability, role-based authority, escalation thresholds, fairness in triage, and continuous validation. The conclusion emphasizes that trustworthy automated incident response requires uncertainty to be rendered visible, contestable, and operationally meaningful rather than hidden behind fluent natural language output.

Keywords

cybersecurity threat intelligence, uncertainty-aware retrieval, automated incident response, trustworthy artificial intelligence, security operations, governance.

1. Introduction

Modern security operations centers face a scale and speed problem that exceeds purely manual analysis. Threat intelligence feeds, endpoint telemetry, network logs, vulnerability disclosures, and incident tickets produce heterogeneous evidence streams that must be interpreted within minutes. Retrieval-augmented generation offers a promising approach because it can combine the generative flexibility of large language models with external knowledge retrieval, allowing responses to be conditioned on current advisories, internal playbooks, and historical incident records [1]. The same architectural pattern has been extended through large-scale retrieval and memory mechanisms that improve factual grounding and domain adaptation [2]. However, cybersecurity is not a conventional question-answering domain. Adversaries act strategically, evidence is often incomplete or intentionally misleading, and incorrect automated actions can disable critical services or violate legal obligations. Consequently, the central design problem is not simply whether a model can

retrieve relevant documents, but whether the resulting system can represent, communicate, and govern uncertainty across the entire incident response lifecycle.

The integration of foundation models into security workflows also raises concerns that extend beyond technical accuracy. Risk taxonomies for language models identify harms related to misinformation, privacy leakage, automation bias, and inappropriate reliance on generated content [19]. In incident response, these harms become operationally concrete. A confidently generated but incorrect containment recommendation may isolate a production database, disrupt clinical systems, or escalate a minor anomaly into a major outage. A retrieval system that misses a critical advisory because of embedding limitations may produce a plausible but obsolete mitigation. An automated agent that acts without calibrated uncertainty may violate zero trust principles by granting excessive authority to a probabilistic component [11]. This paper therefore treats uncertainty-aware retrieval-augmented generation as a governance and infrastructure problem as much as a machine learning problem.

The argument developed here is that uncertainty-aware retrieval-augmented generation can improve cybersecurity threat intelligence and automated incident response only when uncertainty is structurally represented in the architecture. This means uncertainty must influence retrieval, generation, scoring, escalation, and response authorization. It must be visible to analysts, auditable by regulators, and bounded by policy. The paper proceeds by reviewing foundations of retrieval-augmented generation, analyzing sources of uncertainty in threat intelligence, proposing architectural considerations for uncertainty-aware incident response, examining governance and human oversight, discussing deployment and sustainability trade-offs, assessing robustness and fairness, outlining evaluation strategies, and considering policy implications.

2. Foundations of Retrieval-Augmented Generation in Security Operations

Retrieval-augmented generation combines parametric language model knowledge with non-parametric retrieval from an external corpus. Early formulations demonstrated that dense retrieval over a document index can substantially improve knowledge-intensive tasks by grounding generation in retrieved passages [1]. Dense passage retrieval introduced efficient dual-encoder architectures that map queries and documents into a shared semantic space, enabling retrieval beyond lexical matching [3]. Subsequent work integrated retrieval during pre-training and fine-tuning, showing that language models can learn to use retrieved evidence more effectively when retrieval is part of the training objective [4]. Other approaches demonstrated that retrieved passages can be reranked and synthesized by generative models for open-domain question answering [5]. More recent surveys describe retrieval-augmented generation as a broad family of architectures with variations in retrieval timing, index structure, and generation control [6].

For cybersecurity, these foundations are necessary but insufficient. Security knowledge is temporally volatile, hierarchically structured, and often expressed in specialized formats such as indicators of compromise, tactics, techniques, and procedures, and machine-readable threat intelligence. Transformer architectures provide the representational capacity to model long-range dependencies across logs and reports [18], while sentence embedding methods allow semantic similarity search over advisories and playbooks [17]. Yet semantic similarity alone does not capture operational relevance. A retrieved passage may be semantically related to a malware family but irrelevant to the current environment because it describes a different platform, sector, or attack stage. It may also be outdated, superseded by a vendor correction, or contradicted by internal telemetry. A security-oriented retrieval-augmented generation

system must therefore combine semantic retrieval with structured filtering, temporal validity, source reputation, and environmental context.

The foundation model literature further complicates the picture because large models exhibit emergent capabilities alongside unpredictable failure modes [19]. In security operations, this unpredictability is dangerous when automated response is authorized. Retrieval can reduce hallucination, but it does not eliminate it. A model may still misinterpret retrieved evidence, overgeneralize from a single incident, or fabricate a causal chain that links unrelated indicators. Uncertainty-aware retrieval-augmented generation therefore needs mechanisms that estimate confidence at multiple levels: retrieval confidence, source reliability, evidence consistency, generation certainty, and action risk. These estimates must then be translated into operational decisions such as whether to recommend, request confirmation, or execute a response.

3. Uncertainty in Cyber Threat Intelligence

Uncertainty in cyber threat intelligence arises from several distinct sources. Epistemic uncertainty reflects incomplete knowledge about an adversary, infrastructure, or vulnerability. Aleatoric uncertainty reflects inherent randomness in observed signals, such as false positives in intrusion detection or noisy endpoint telemetry. Linguistic uncertainty arises because threat reports use vague, hedged, or inconsistent language. Temporal uncertainty occurs when intelligence is valid only within a narrow window. Adversarial uncertainty is deliberate: attackers may plant false indicators, poison public feeds, or craft prompts that manipulate retrieval systems. These categories are not merely academic; they determine what kind of uncertainty quantification is useful. A system that reports a single confidence score without distinguishing among these sources may create false precision. Work on hallucination in natural language generation shows that fluent outputs can be unsupported by evidence, and that evaluation must separate factuality from fluency [7]. In cybersecurity, this distinction is existential because fluent but unsupported attribution can trigger diplomatic, legal, or military consequences.

Risk taxonomies for language models emphasize that harms emerge from the interaction of model capability, deployment context, and user behavior [8]. In security operations, automation bias is particularly salient. Analysts may defer to a generated recommendation because it appears authoritative, especially under time pressure. If the system does not expose uncertainty, the human oversight layer becomes ceremonial rather than substantive. Research on artificial intelligence safety has long warned about specification gaming, negative side effects, and reward hacking [9]. An automated incident response system optimized only for containment speed may take actions that satisfy a narrow metric while degrading service availability or violating privacy constraints. Interpretability research argues that high-stakes decisions should favor inherently interpretable models or, at minimum, provide explanations that support contestation [10]. Uncertainty-aware retrieval-augmented generation can contribute by making the evidential basis of a recommendation inspectable, including which sources were retrieved, how consistent they were, and where the model's confidence declined.

Uncertainty-aware retrieval-augmented generation frameworks attempt to control trustworthiness by estimating uncertainty during retrieval and generation and by applying risk controls when uncertainty exceeds acceptable bounds [13]. In the security domain, this idea must be extended into an operational policy. For example, high uncertainty about an indicator's maliciousness may permit monitoring but not blocking. Moderate uncertainty about lateral movement may justify endpoint isolation only after human approval. Low

uncertainty combined with high severity may authorize predefined containment playbooks, but only within explicitly scoped systems. Risk assessment methodologies provide structured ways to reason about likelihood, impact, and control effectiveness [14]. When integrated with retrieval-augmented generation, they allow uncertainty to be mapped onto response authority rather than treated as an abstract model property.

4. Architecture of Uncertainty-Aware RAG for Incident Response

An uncertainty-aware retrieval-augmented generation architecture for incident response should be understood as a layered socio-technical system. The ingestion layer collects threat intelligence feeds, vulnerability databases, internal playbooks, case histories, and telemetry. The normalization layer maps heterogeneous data into shared schemas and preserves provenance. The retrieval layer combines lexical, dense, and structured search. The uncertainty layer estimates reliability and consistency across retrieved evidence. The generation layer produces candidate explanations and response recommendations. The policy layer maps uncertainty and severity to permissible actions. The orchestration layer executes approved actions or routes them to human analysts. The audit layer records evidence, model outputs, confidence estimates, policy decisions, and human interventions. This architecture is not a single model but a control system. Its purpose is to maintain traceability from raw evidence to automated action.

Zero trust principles are relevant because they reject implicit trust based on network location and instead require continuous verification [11]. Applied to retrieval-augmented generation, this means that retrieved content should not be trusted merely because it appears in a corpus. Sources must be authenticated, integrity-checked, and evaluated for freshness. Generated recommendations should not be trusted merely because they are fluent. They must be verified against evidence and policy. Adversary behavior frameworks such as MITRE ATT&CK provide structured vocabularies for tactics and techniques that can improve retrieval precision and enable cross-incident comparison [12]. When a retrieved passage mentions credential dumping, for example, the system can map it to a technique identifier and assess whether current telemetry contains corresponding indicators. This structured mapping reduces reliance on unconstrained semantic similarity.

The policy layer is where uncertainty becomes operationally meaningful. A cybersecurity framework such as the NIST Cybersecurity Framework provides a high-level structure for identify, protect, detect, respond, and recover functions [15]. An information security management system standard such as ISO/IEC 27001:2022 provides requirements for governance, risk treatment, and continual improvement [16]. In an uncertainty-aware retrieval-augmented generation system, these frameworks inform escalation rules. A recommendation with low retrieval confidence and high potential impact should not be auto-executed. A recommendation with conflicting evidence should be routed to a senior analyst. A recommendation with high confidence but novel context should be treated as provisional until validated. The architecture must also support rollback and compensating controls, because automated response is not merely a prediction; it is an intervention in a complex system.

5. Governance, Risk, and Human Oversight

Governance is not an external wrapper around uncertainty-aware retrieval-augmented generation; it is constitutive of trustworthy operation. The system must define who is accountable for automated actions, how authority is delegated, and how disputes are resolved.

Interpretability requirements matter because analysts and auditors need to understand why a recommendation was made [10]. This does not necessarily require full mechanistic interpretability, but it does require evidence trails, source citations, uncertainty annotations, and counterfactual explanations. For example, an analyst should be able to ask which retrieved documents most influenced a containment recommendation, whether any source was contradicted, and what alternative actions were considered. If the system cannot answer such questions, it cannot be safely integrated into incident response.

Human oversight must be designed for realistic operational conditions. Security analysts are often overloaded, and asking them to review every low-level alert defeats automation. The better approach is risk-based escalation. Low-impact and high-confidence actions may be automated within narrow scopes. Moderate uncertainty triggers human confirmation. High uncertainty or high impact requires senior review and possibly legal or privacy consultation. This tiered model aligns with risk assessment practices that consider likelihood, impact, and control effectiveness [14]. It also aligns with cybersecurity governance frameworks that emphasize continuous monitoring, improvement, and documented risk treatment [15]. The key is that uncertainty estimates must be calibrated enough to support thresholds. An uncalibrated confidence score is dangerous because it creates an illusion of control.

Technical debt in machine learning systems compounds governance challenges [20]. Retrieval corpora age, embeddings drift, source formats change, and adversary behavior evolves. A system that performs well in a controlled pilot may degrade silently in production. Governance must therefore include periodic revalidation, red teaming, data quality audits, and incident postmortems focused on model-assisted decisions. The required uncertainty-aware retrieval-augmented generation framework emphasizes trustworthiness and risk control [13], but institutional adoption depends on whether those controls are operationally legible. Policies should specify acceptable false positive and false negative rates for different response classes, maximum acceptable latency, minimum source diversity, and mandatory human review for actions affecting safety-critical systems. Without such specifications, uncertainty-aware retrieval-augmented generation becomes a technical slogan rather than a governance mechanism.

6. Deployment, Sustainability, and Infrastructure Trade-offs

Deployment choices shape both capability and risk. A centralized cloud deployment can aggregate threat intelligence and provide elastic compute for retrieval and generation, but it raises data sovereignty, privacy, and latency concerns. An on-premises deployment can keep sensitive telemetry local, yet it may struggle to maintain large indexes and frequent model updates. A hybrid architecture may separate sensitive retrieval from shared intelligence, using policy-aware federation. Each choice affects sustainability. Large language models and dense retrieval systems consume significant energy and require continuous hardware investment. Security operations are expected to run continuously, so inefficient retrieval can increase operational cost and carbon footprint. The system should use cascaded retrieval, where cheap lexical or structured filters narrow the candidate set before expensive semantic models are invoked. It should also cache stable knowledge, expire stale embeddings, and route only uncertain cases to larger models.

The foundation model literature warns that scale alone does not guarantee robustness [19]. In security, a smaller domain-adapted model with well-curated retrieval may outperform a larger general model on specific tasks while reducing cost and latency. Uncertainty-aware routing can dynamically select model and retrieval depth based on incident severity and ambiguity.

For example, a routine phishing alert may require only a lightweight classifier and curated playbook retrieval. A suspected supply chain compromise may require multi-source retrieval, cross-lingual search, and adversarial verification. This adaptive approach supports sustainability because it avoids applying maximum compute to every event. It also supports resilience because the system can degrade gracefully if a cloud service or feed becomes unavailable. Zero trust principles suggest that fallback mechanisms must not silently trust lower-quality sources [11]. Instead, degraded mode should narrow authority and escalate more decisions to humans.

Sustainability also includes maintainability. A security system that cannot be updated quickly after a new vulnerability disclosure loses value. Deployment pipelines must support rapid corpus updates, index rebuilding, policy changes, and model rollbacks. Yet rapid updates create risk of poisoning or regression. The system should maintain signed source manifests, staging environments, canary deployments, and automated consistency checks. Technical debt accumulates when teams treat retrieval-augmented generation as a model rather than a data-intensive service [20]. Infrastructure investment in provenance, observability, and versioning is therefore not secondary. It is what allows uncertainty-aware behavior to remain trustworthy over time.

7. Robustness, Fairness, and Adversarial Considerations

Adversarial robustness is a first-order requirement. Attackers may poison public threat intelligence, compromise internal documents, or inject malicious instructions into retrieved content. Prompt injection is especially dangerous when a retrieval-augmented generation system reads external text and treats it as evidence. An adversary could embed instructions in a vulnerability report that cause the model to recommend disabling a security control. Defenses include content sanitization, instruction hierarchy, source allowlisting, provenance verification, and separation between retrieved evidence and executable policy. Adversary behavior frameworks help defenders anticipate techniques [12], but they do not eliminate the need for adversarial testing. Red teams should attempt to manipulate retrieval rankings, confidence scores, and policy thresholds. Uncertainty-aware mechanisms can help by flagging contradictory or anomalous evidence, but they must themselves be robust to manipulation. If an attacker can inflate confidence, uncertainty-aware design fails.

Fairness in cybersecurity automation is less discussed but important. Triage and response systems may disproportionately affect certain departments, geographic regions, or user groups. For example, an automated containment policy might isolate systems used by a marginalized language community because threat intelligence coverage is weaker for that language. Or it might over-monitor employees in sensitive roles due to biased historical incident data. Risk taxonomies for language models highlight representational harms and disparate performance [8]. In incident response, fairness requires evaluating whether uncertainty estimates are equally reliable across environments, vendors, and attack types. A system that is well calibrated for Windows enterprise networks but poorly calibrated for industrial control systems may systematically under-protect critical infrastructure. Governance should therefore include disaggregated evaluation and inclusive threat modeling. Interpretability can support fairness by revealing which features or sources drive differential outcomes [10].

Safety research emphasizes that systems should be designed to avoid negative side effects and to remain corrigible under uncertainty [9]. In automated incident response, corrigibility means that human operators can override, pause, or reverse automated actions. It also means that the system should not escalate its own privileges or resist oversight. Uncertainty-aware retrieval-

augmented generation should treat high uncertainty as a signal to reduce autonomy, not to seek more aggressive action. This principle must be encoded in policy, not left to model behavior. The system should also maintain an explicit distinction between observation, inference, and recommendation. Conflating them increases the risk that an uncertain inference is treated as a confirmed fact.

8. Evaluation and Operational Metrics

Evaluating uncertainty-aware retrieval-augmented generation requires metrics beyond answer accuracy. Retrieval quality must be assessed for relevance, coverage, freshness, source diversity, and susceptibility to poisoning. Generation quality must be assessed for factual grounding, citation correctness, and resistance to hallucination [7]. Uncertainty quality must be assessed through calibration, discrimination, and selective prediction. A well-calibrated system should be approximately correct when it claims high confidence and should abstain or escalate when it does not know. Evaluation must also be scenario-based. Tabletop exercises can simulate ransomware, supply chain compromise, or insider threat scenarios and measure whether the system escalates appropriately. Operational metrics should include time to triage, time to containment, percentage of actions reversed, analyst override rate, and incident impact. These metrics must be interpreted alongside uncertainty thresholds, because a system that never acts will have low error but also low value.

The required uncertainty-aware retrieval-augmented generation perspective emphasizes trustworthy generation and risk control [13]. In practice, this means evaluation should be continuous and adversarial. Static benchmarks become obsolete as attackers adapt. The system should monitor for distribution shift in queries, retrieved documents, and model outputs. It should detect when confidence scores become decoupled from outcomes. It should also evaluate human factors. If analysts ignore uncertainty indicators, the interface has failed. If they over-rely on them, calibration has failed. Interpretability research suggests that explanations should be actionable and contestable [10]. An uncertainty display that merely shows a number is insufficient. It should show which evidence is missing, which sources conflict, and what additional data would reduce uncertainty. Safety research further suggests that evaluation should include failure modes under distributional shift and adversarial pressure [9]. A system that performs well on historical logs but fails during a novel attack is not robust.

9. Policy Implications and Cross-Domain Lessons

Policy implications extend beyond organizational security. Automated incident response can affect critical infrastructure, civil liberties, and international norms. If an uncertainty-aware retrieval-augmented generation system recommends countermeasures against an attributed actor, it may influence diplomatic or military decisions. Risk assessment frameworks provide a disciplined language for likelihood and impact [14], while cybersecurity frameworks provide a structure for governance and improvement [15]. Information security management standards require documented controls, responsibilities, and continual improvement [16]. These instruments should be adapted to explicitly address model-assisted decisions. Regulators may require audit logs, human review for high-impact actions, transparency about automated decision logic, and redress mechanisms for erroneous containment. Zero trust architecture offers a useful principle: verify explicitly and assume breach [11]. Applied to artificial intelligence, this means verify model outputs explicitly and assume the corpus may be compromised.

Cross-domain lessons from other safety-critical sectors are instructive. Aviation uses layered redundancy, checklists, and incident reporting. Healthcare uses clinical decision support with clear escalation pathways. Nuclear power uses defense in depth and independent verification. Cybersecurity incident response should adopt similar layers. Uncertainty-aware retrieval-augmented generation can serve as one layer, but it cannot replace independent sensors, deterministic controls, or human judgment. The risk taxonomy literature shows that harms often arise from deployment context rather than model capability alone [8]. Therefore, policy should focus on the socio-technical system: who uses the output, under what time pressure, with what authority, and with what recourse. A system that is technically calibrated but institutionally opaque will not earn trust. A system that is institutionally accountable but technically uncalibrated will produce dangerous automation.

10. Conclusion

Uncertainty-aware retrieval-augmented generation offers a credible path toward more trustworthy cybersecurity threat intelligence and automated incident response, but only if uncertainty is treated as a first-class architectural and governance concern. Retrieval can ground language models in current advisories, playbooks, and historical cases, yet cybersecurity evidence is adversarial, incomplete, and time-sensitive. A single confidence score is insufficient. The system must distinguish source reliability, retrieval relevance, evidence consistency, generation certainty, and action risk. It must map these distinctions onto escalation thresholds, human oversight, and automated authority. It must be deployable in sustainable and resilient infrastructures, maintainable under rapid change, and auditable under regulatory scrutiny. It must also be robust to poisoning and fair across diverse environments. The required uncertainty-aware retrieval-augmented generation framework [13] provides an important foundation for trustworthiness and risk control, but its value depends on integration with security governance, zero trust principles, risk assessment, and continuous evaluation. The future of automated incident response should not be a contest between human analysts and autonomous agents. It should be a layered socio-technical system in which uncertainty is made visible, contestable, and operationally meaningful, allowing organizations to act decisively when confidence is warranted and to refrain or escalate when it is not.

References

1. 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* (Vol. 33, pp. 9459–9474). Curran Associates.
2. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 2206–2240). PMLR.
3. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (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.

4. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 3929–3938). PMLR.
5. 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: Main Volume* (pp. 874–880). Association for Computational Linguistics.
6. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv*. <https://arxiv.org/abs/2312.10997>
7. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. <https://doi.org/10.1145/3571730>
8. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., ... Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). ACM. <https://doi.org/10.1145/3531146.3533088>
9. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv*. <https://arxiv.org/abs/1606.06565>
10. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
11. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture (NIST Special Publication 800-207). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-207>
12. MITRE Corporation. (2024). MITRE ATT&CK. <https://attack.mitre.org/>
13. Zhang, R. (2025, December). Uncertainty-Aware RAG: Trustworthy Generation and Risk Control. In *2025 5th International Symposium on Artificial Intelligence and Big Data (AIBDF)* (pp. 526-531). IEEE.
14. National Institute of Standards and Technology. (2012). Guide for conducting risk assessments (NIST Special Publication 800-30 Rev. 1). <https://doi.org/10.6028/NIST.SP.800-30r1>
15. National Institute of Standards and Technology. (2018). Framework for improving critical infrastructure cybersecurity, version 1.1. <https://doi.org/10.6028/NIST.CSWP.04162018>
16. International Organization for Standardization. (2022). Information security, cybersecurity and privacy protection — Information security management systems — Requirements (ISO/IEC 27001:2022). <https://www.iso.org/standard/27001>
17. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

Language Processing (pp. 3982–3992). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/D19-1410>

18. 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* (Vol. 30, pp. 5998–6008). Curran Associates.
19. 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*.
<https://arxiv.org/abs/2108.07258>
20. 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* (Vol. 28, pp. 2503–2511). Curran Associates.