



REPLACING THE ANALYST? THE ROLE OF AI IN AUTOMATED INCIDENT RESPONSE

Nishant Gottumukkala

Independent Student Researcher

Email: gnishant2009@gmail.com

Abstract—When a security alert pops up, analysts usually have to dig through identity logs and threat intel to figure out what to do. Everyone says AI will speed this up and handle responses automatically, but we still have to figure out what we can actually trust it to handle on its own. I looked through several studies and industry reports available through October 2026. The exact numbers are tough to compare directly, and the studies report different kinds of workload reductions. For example, NoDoze reported a drop in false alarms of over 84%; a study using real CSOC data reported a 60.16% reduction in shift backlog; and AACT reported 61% fewer alerts shown to analysts. What caught my eye is what happened when researchers tracked how staff used these bots on the job: people mainly treated them as extra hands for deciphering raw telemetry and polishing shift summaries. They did not commonly hand over high-stakes decisions. IBM and Ponemon reports associate heavy security-AI use with cheaper breaches and shorter breach lifecycles, but that is a correlation, not a magic bullet. AI can be a useful assistant for initial triage, but it should not run the whole show. Automate the easy, reversible tasks and keep humans in charge of decisions with broad consequences. Judge these tools by whether they help contain threats and keep the business running—not by how many tickets they close.

Index Terms—Artificial intelligence, Security Operations Center (SOC), automated threat response, human-in-the-loop, security telemetry, alert fatigue.

I. Introduction

Modern Security Operations Centers (SOCs) process high volumes of telemetry generated across endpoints, cloud infrastructures, and application services. While log aggregation has largely been solved by enterprise security platforms, isolating actionable threats within pervasive noise remains a core challenge. Persistent exposure to repetitive false positives induces analyst fatigue, reducing cognitive capacity for complex investigations, a vulnerability well-documented in operational security literature [1], [2]. Machine learning (ML) models offer mechanisms to correlate disparate event streams and prioritize high-risk signals [3], [4]. However, an operational boundary exists between automated recommendations and autonomous execution. Responses carrying significant operational risks such as critical system isolation necessitate human authorization.

II. The Anatomy of an AI-Driven SOC

A. Data Ingestion and SIEM Integration

Security Information and Event Management (SIEM) systems serve as centralized repositories, ingesting telemetry from firewalls, identity providers, endpoint detection platforms, and servers to enable centralized querying and correlation. AI layers leverage these centralized data structures to detect anomalous trends and aggregate related indicators. Nonetheless, model accuracy remains constrained by ingested data quality and visibility gaps; unannounced sensor dropouts or aggressive telemetry truncation can produce silent blind spots or obscure execution context. Consequently, AI-generated alerts must be treated as probabilistic indicators rather than definitive conclusions, requiring transparent lineage back to raw, unaggregated log data [3], [4].



B. Behavioral Triage (UEBA)

Unlike traditional signature-based detection mechanisms, User and Entity Behavior Analytics (UEBA) establishes baseline profiles of normal operational activity to flag statistical anomalies. For example, an authentication event occurring at non-standard hours from a novel geographic region, accompanied by anomalous data egress, triggers an elevated risk score. However, anomalous behavior does not inherently signify compromise; legitimate explanations include remote travel, active VPN sessions, or scheduled maintenance scripts. High risk scores signify baseline deviation rather than intent. Therefore, administrative containment actions such as account suspension require cross-verification against organizational contexts, including travel logs, change management records, and active IT tickets [3], [4].

C. Threat Contextualization and Response Orchestration

Following initial triage, AI models accelerate enrichment by aggregating historical incident data, threat intelligence feeds, and asset ownership metadata. Large Language Models (LLMs) further assist by converting unstructured command-line artifacts or raw obfuscated scripts into human-readable summaries. Security Orchestration, Automation, and Response (SOAR) platforms subsequently execute programmatic playbooks, such as blocking external IP addresses or revoking session tokens. Frameworks like AI4SOAR illustrate the viability of AI-recommended playbooks in specific attack vectors, such as SSH brute-force incidents [5].

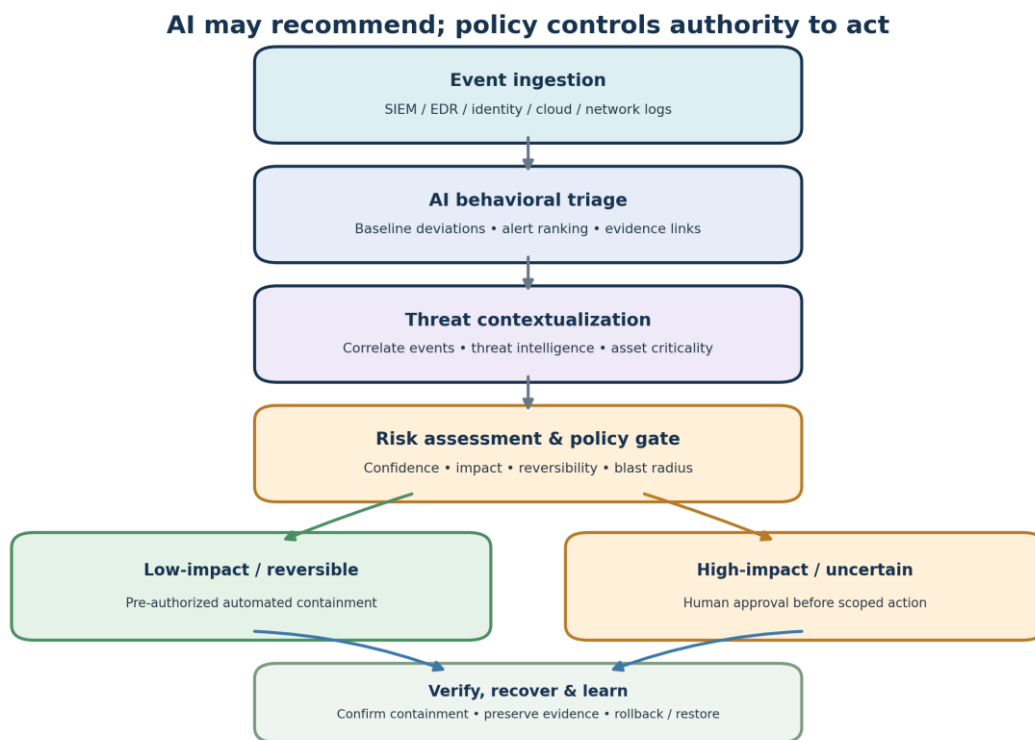


Fig. 1. Proposed AI incident-response pipeline.

To mitigate systemic risk, analytical recommendations must be strictly decoupled from execution mechanisms. Low-stakes, reversible actions may be automated under pre-approved policies, whereas high-consequence remediations must require human oversight and mandate deterministic rollback pathways. Fig. 1 illustrates this operational decision topology [6].



III. The Operational Impact of Automated Triage

While vendor literature often highlights rapid response latency, raw API execution speed does not directly equate to threat containment. Empirical evaluations demonstrate that security automation primarily optimizes queue management and administrative throughput rather than enabling autonomous threat mitigation.

Evaluating individual alert-triage benchmarks reveals substantial variance in metrics across deployment models [7]–[9]:

- **False Alarm Reduction:** Evaluation of *NoDoze* in an enterprise environment demonstrated an 84% reduction in false positives. The platform's reported savings of over 90 analyst hours weekly relied on an estimated average resolution time of 20 minutes per false alarm rather than directly empirical measurement [7].
- **Shift Backlog Mitigation:** A study analyzing operational Security Operations Center (SOC) data recorded a 60.16% drop in shift backlog volume. While representing significant workload relief, this reduction reflects queue optimization rather than a 60.16% increase in threat mitigation velocity [8].
- **Alert Suppression:** Evaluation of *AACT* (Automated Alert Classification and Triage) indicated a 61% reduction in analyst-facing alert volume. False-negative estimates were derived using stratified sampling over suppressed alert sets rather than exhaustive re-investigation across all suppressed instances [9].

Because these frameworks measure distinct operational parameters, direct head-to-head performance comparisons remain Methodologically invalid (Fig. 2).

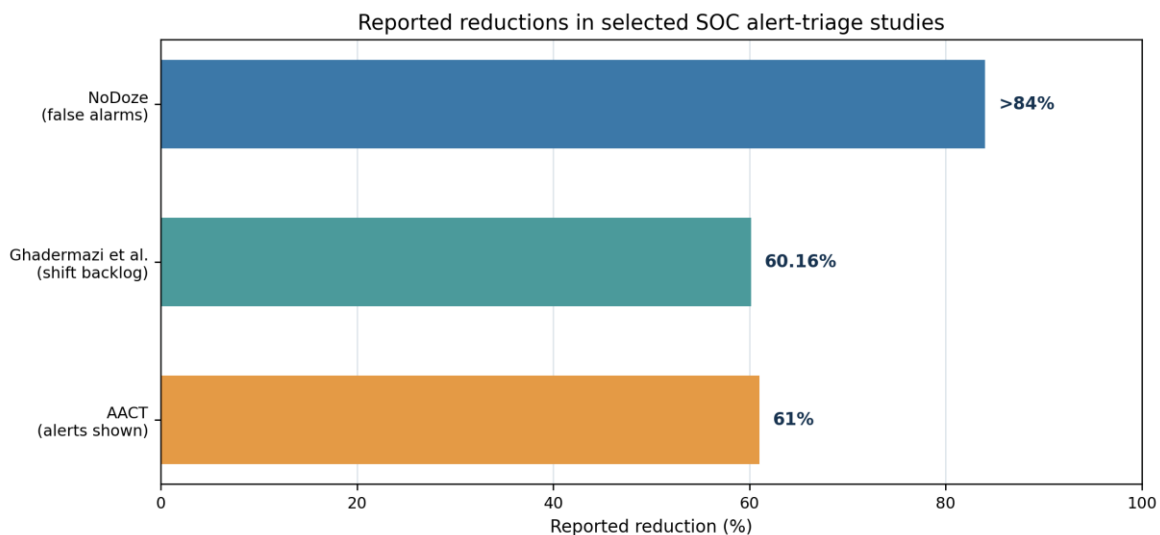


Fig. 2. Reported performance metrics across selected alert-triage frameworks. Differences in baseline parameters and measurement methodologies preclude direct comparative ranking.

A. Human-AI Interaction and Task Allocation

Observational studies tracking real-world operational workflows reveal that analysts deploy Large Language Models (LLMs) primarily as cognitive support tools rather than autonomous decisionmakers [10]. Across an evaluation of 3,090 natural language queries generated by 45 SOC analysts over a ten-month period, primary usage focused on parsing raw terminal telemetry, writing scripts, and refining



shift handovers (Fig. 3). Direct yes-or-no triage determinations accounted for approximately 4% of total queries, indicating that operational staff treat AI models as consultative assistants rather than decision authorities [10].

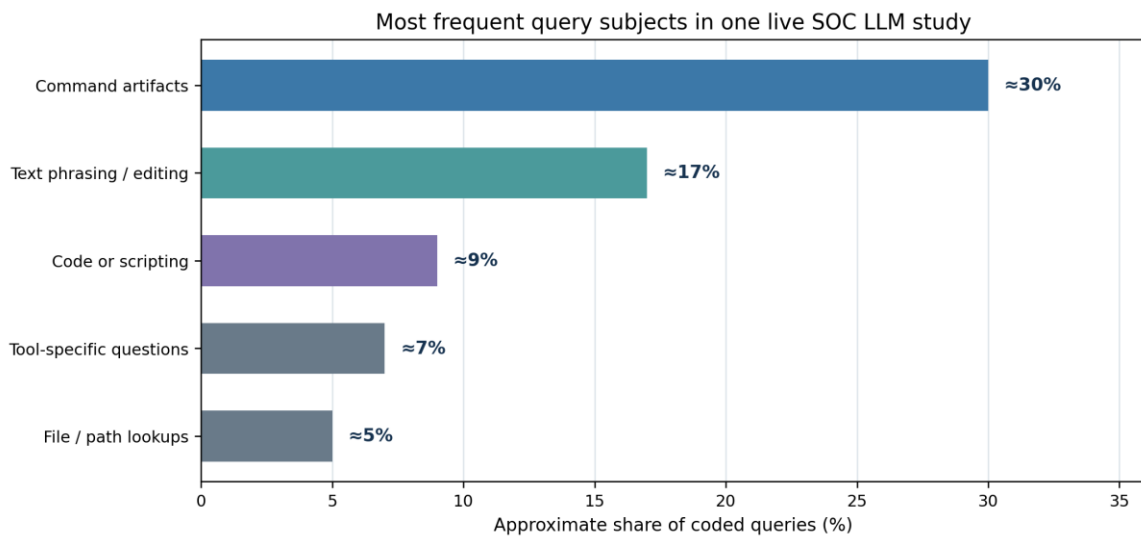


Fig. 3. Primary interaction categories from a 10-month observational study of LLM queries within an active SOC [10].

Interactive benchmarking of commercial Security Orchestration, Automation, and Response (SOAR) platforms reveals trade-offs between speed and investigation thoroughness [11]. In a controlled study involving 24 practitioners across six commercial SOAR tools, platform orchestration reduced context-switching and accelerated ticket resolution times; however, manually completed tickets consistently contained more comprehensive contextual analysis than those populated automatically by SOAR integrations [11].

B. Financial and Breach Lifecycle Correlations

Industry analyses associate extensive security AI deployment with lower financial breach impacts and shorter containment lifecycles [12]–[14].

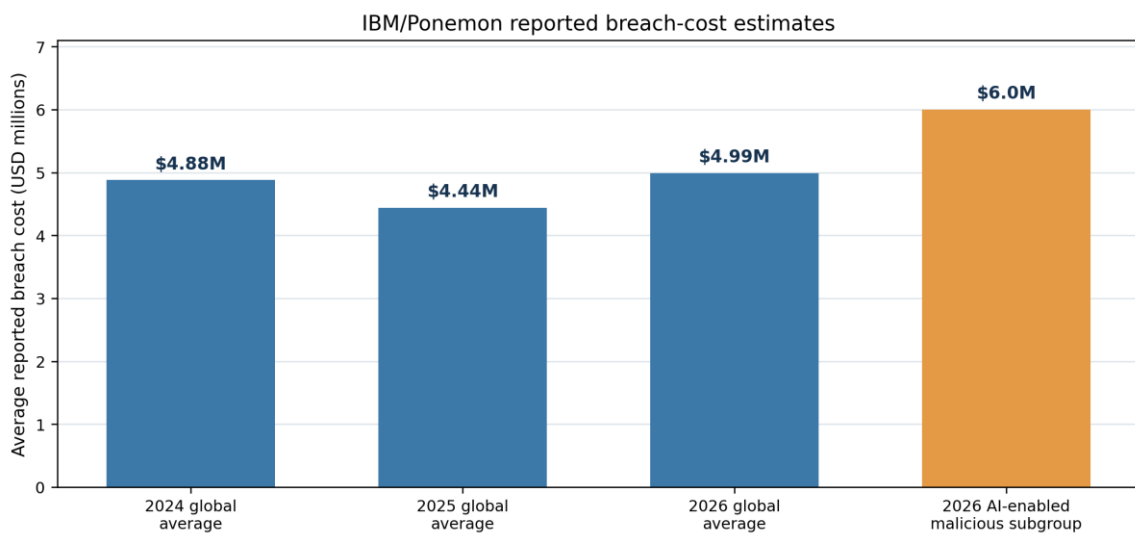


Fig. 4. Reported average breach cost comparisons across security AI adoption levels (2024–2026 data synthesis) [12]–[14]. These values represent observational correlation rather than direct causal effect.



Benchmark reporting established the global average breach cost at USD 4.99 million, with AI-targeted or AI-enabled malicious breaches rising to USD 6.00 million [12], [13]. Conversely, organizations deploying mature security AI implementations recorded lower average breach costs (a difference approaching USD 2 million) and compressed breach lifecycles compared to unautomated environments [12], [14]. These metrics reflect observational correlation: enterprises investing in advanced security AI architectures typically possess larger security budgets, higher overall infrastructure maturity, and specialized operational staff. Consequently, lower breach costs cannot be attributed exclusively to AI tooling [12].

IV. Mathematical Limits and Security Vulnerabilities

While an erroneous alert merely consumes analyst time during manual review, an automated system acting on false positives introduces operational risk—such as locking out executive accounts or disconnecting critical production databases. Security Operations Centers (SOCs) must balance the risk of uncontained attacks against the self-inflicted business disruption caused by false-positive automated containment.

A. The Base-Rate Effect on Automated Precision

A primary limitation in automated threat detection stems from Bayesian probability and the base-rate fallacy. Even a high-performing model exhibiting 95% sensitivity (TPR = 0.95) and 95% specificity (TNR = 0.95) produces a low Positive Predictive Value (PPV) when true malicious events are extremely rare within total network telemetry.

Mathematically, given threat prevalence $P(A)$, PPV is defined as:

$$PPV = \frac{\text{Sensitivity} \cdot P(A)}{\text{Sensitivity} \cdot P(A) + (1 - \text{Specificity}) \cdot (1 - P(A))} \quad (1)$$

At a baseline threat prevalence of 1% ($P(A) = 0.01$), the resulting precision is:

$$PPV = \frac{0.95 \cdot 0.01}{(0.95 \cdot 0.01) + (0.05 \cdot 0.99)} \approx 16.1\% \quad (2)$$

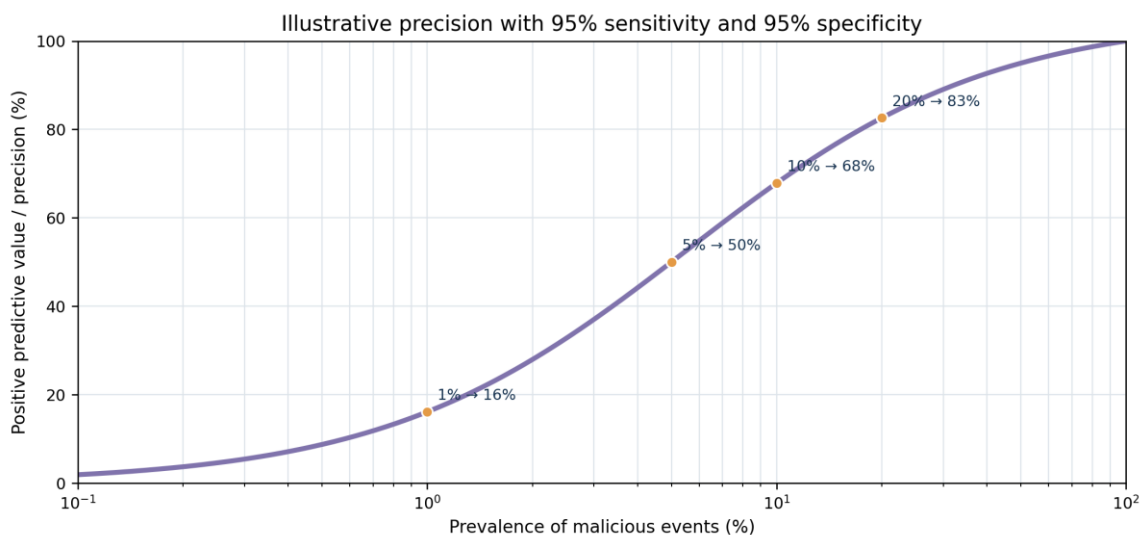


Fig. 5. Mathematical relationship between event prevalence and Positive Predictive Value at fixed 95% sensitivity and specificity. Low threat prevalence inherent to enterprise networks mathematically guarantees high false-positive ratios during automated triage.



B. Adversarial Exploitation and Indirect Prompt Injection

Beyond statistical limitations, automated models present unique attack surfaces:

1. **Adversarial Evasion:** Malicious actors deliberately structure payload characteristics or telemetry bursts to alter feature vectors and bypass model detection thresholds [15].
2. **Indirect Prompt Injection:** When Large Language Models (LLMs) ingest unstructured logs, external ticketing payloads, or raw telemetry containing attacker-crafted text, malicious commands can manipulate LLM tool-use parameters [16].

To mitigate these risks, AI agents must be restricted from possessing administrative privileges. Model permissions must adhere to the Principle of Least Privilege (PoLP), requiring explicit human authorization for high-impact actions and enforcing deterministic rollback capabilities [6], [16].

V. Operational Control Frameworks: Human-in-the-Loop vs. Human-on-the-Loop

A. Human-in-the-Loop (HITL) Architectures

In a Human-in-the-Loop model, the AI agent performs autonomous enrichment, parses telemetry, and formulates remediation strategies, but execution remains strictly blocked until an analyst validates and approves the command. HITL is mandatory whenever a proposed action carries non-reversible consequences, such as taking production services offline, revoking domain administrator access, or modifying perimeter routing tables.

B. Human-on-the-Loop (HOTL) Architectures

In a Human-on-the-Loop setup, the system executes low-stakes, pre-approved actions autonomously while human analysts monitor system state in real time with override authority. HOTL deployment is appropriate for narrowly scoped, reversible playbooks—such as revoking an active single-user session or blocking a verified malicious hash. Prior to live deployment, HOTL policies should operate in "shadow mode" to benchmark model recommendations against human decision matrices [6], [16].

VI. Conclusion

Artificial intelligence provides substantial utility in modern security operations by automating administrative log parsing, contextualizing complex command line scripts, and streamlining low-level incident triage. However, performance metrics from cited literature emphasize workload queue reduction rather than universal operational autonomy. Furthermore, correlation between high security AI adoption and reduced breach costs must not be conflated with direct causation. Security AI functions most effectively as a decision-support capability rather than an autonomous decision authority. Unconstrained automation risks self-inflicted operational downtime comparable to active security breaches. SOC architectures must maintain a clear boundary: automate low-risk, reversible tasks, mandate human oversight for high-stake remediations, and evaluate security AI efficacy by business continuity and threat containment rather than ticket resolution volume.

REFERENCES

- [1] S. Tariq, M. B. Chhetri, S. Nepal, and C. Paris, "Alert fatigue in security operations centres: Research challenges and opportunities," *ACM Computing Surveys*, vol. 57, no. 9, Art. no. 224, 2025, doi: 10.1145/3723158.
- [2] 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.*, 2022, pp. 2783–2800.



- [3] A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.
- [4] Y. Xin et al., “Machine learning and deep learning methods for cybersecurity,” *IEEE Access*, vol. 6, pp. 35365–35381, 2018, doi: 10.1109/ACCESS.2018.2836950.
- [5] M.-D. Nguyen, W. Mallouli, A. R. Cavalli, and E. Montes de Oca, “AI4SOAR: A security intelligence tool for automated incident response,” in *Proc. 19th Int. Conf. Availability, Reliability and Security (ARES '24)*, 2024, Art. no. 170, doi: 10.1145/3664476.3670450.
- [6] A. Nelson, S. Rekhi, M. Souppaya, and K. Scarfone, *Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile*, NIST Special Publication 800-61 Rev. 3, 2025, doi: 10.6028/NIST.SP.800-61r3.
- [7] W. Ul 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. Network and Distributed System Security Symp. (NDSS)*, 2019, doi: 10.14722/ndss.2019.23349.
- [8] J. Ghadermazi, A. Shah, and S. Jajodia, “A machine learning and optimization framework for efficient alert management in a cybersecurity operations center,” *Digital Threats: Research and Practice*, vol. 5, no. 2, Art. no. 19, 2024, doi: 10.1145/3644393.
- [9] M. Turcotte, F. Labrèche, and S.-O. Paquette, “Automated alert classification and triage (AACT): An intelligent system for the prioritisation of cybersecurity alerts,” *arXiv preprint arXiv:2505.09843*, 2025, doi: 10.48550/arXiv.2505.09843.
- [10] R. Singh, S. Tariq, F. Jalalvand, M. B. Chhetri, S. Nepal, C. Paris, and M. Lochner, “LLMs in the SOC: An empirical study of human-AI collaboration in security operations centres,” in *Proc. 2026 IEEE Symp. Security and Privacy*, 2026, doi: 10.1109/SP63933.2026.00111.
- [11] R. A. Bridges et al., “Testing SOAR tools in use,” *Computers & Security*, vol. 129, Art. no. 103201, 2023, doi: 10.1016/j.cose.2023.103201.
- [12] IBM, *Cost of a Data Breach Report 2026*, 2026. [Online]. Available: <https://www.ibm.com/reports/data-breach>
- [13] IBM, “IBM study: One in four malicious breaches are AI-enabled, costing companies \$6 million on average,” Jul. 29, 2026. [Online]. Available: <https://newsroom.ibm.com/2026-07-29-ibm-study-one-in-four-malicious-breaches-are-ai-enabled,-costing-companies-6-million-on-average>
- [14] IBM, “IBM report: 13% of organizations reported breaches of AI models or applications, 97% of which reported lacking proper AI access controls,” Jul. 30, 2025. [Online]. Available: <https://newsroom.ibm.com/2025-07-30-ibm-report-13-of-organizations-reported-breaches-of-ai-models-or-applications,-97-of-which-reported-lacking-proper-ai-access-controls>
- [15] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2e2023, 2024, doi: 10.6028/NIST.AI.100-2e2023.
- [16] E. Tabassi, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.