Quantifying AI Impact in Post-Compromise Incident Response

Muntathar Abid
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
montii.abid@uts.edu.au

Priyadarsi Nanda
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
priyadarsi.nanda@uts.edu.au

***Abstract*— This paper presents metrics for measuring the operational impact of artificial intelligence in cybersecurity incidents. It introduces five practical metrics that assess how AI influences the speed, scale, and visibility of post-compromise activity. The metrics are evaluated through two simulated scenarios representing different levels of organisational maturity and security capability. Results show that artificial intelligence can accelerate attacker actions and challenge traditional detection and response workflows. This highlights the need for measurable indicators of AI's impact in cyber-attacks. The metrics support readiness assessments in enterprise environments with increasing integration of AI systems.**

Keywords*— *Incident response, post-compromise analysis, AI security, defence metrics.

I. Introduction

The adoption of artificial intelligence (AI) technologies particularly large language models (LLMs) is rapidly transforming how businesses operate across various sectors. Organisations are increasingly leveraging AI-powered tools such as Microsoft Copilot, ChatGPT Enterprise, and GitHub Copilot to streamline workflows, automate content creation, assist with software development, and support strategic decision-making. These platforms are becoming integral to modern enterprise operations, offering enhanced efficiency and new capabilities in day-to-day business functions [1]. AI-enabled threats differ fundamentally from conventional cybersecurity threats. These attacks can operate at machine speed, exploit trusted enterprise platforms post-compromise without triggering standard security alerts, and utilise advanced obfuscation techniques that evade signature-based detection systems [2]. The dual role of AI both as a productivity enabler and as a potential attack vector creates complex scenarios where, malicious activity can occur within legitimate business applications, complicating detection and attribution. The threat landscape now includes both AI-enhanced variants of traditional attacks and entirely new categories of AI-specific threats. Adversaries exploit AI systems through techniques such as prompt injection, model poisoning, and adversarial inputs [3] while also using AI to scale conventional attack methods.

This evolution enables adversaries to conduct reconnaissance, generate malware, personalise social-engineering campaigns, and execute lateral movement at machine speed and scale. Established incident-response (IR) frameworks such as NIST SP 800-61 Rev. 2 [4] and ISO/IEC 27035 [5] remain foundational, but they were designed for traditional threat models. While these models are comprehensive, they do not yet provide detailed provisions for AI-augmented threats including autonomous adaptability, rapid operational scaling, and the dual-use nature of AI systems [6]. Classical IR metrics such as mean time to detection (MTTD) and mean time to recovery (MTTR) are likewise ill-suited when objectives can be achieved in minutes, if not seconds. Compounding this, AI system architectures introduce new forensic challenges; evidence may be distributed across multiple service providers, model versions, and data pipelines.

Several initiatives have begun to address parts of this research gap. MITRE's ATLAS catalogues adversarial threats to AI systems provides taxonomies for AI-specific attack techniques [7]. Carnegie Mellon's AI Security Incident Response Team (AISIRT) investigates AI-specific vulnerabilities and develops incident-handling guidance [8]. In parallel, the AI incident sharing initiative launched by MITRE and industry partners facilitates rapid exchange on emerging AI-enabled threats [9]. However, these efforts primarily emphasise threat documentation and general guidance, rather than quantitative frameworks for evaluating IR effectiveness in AI-integrated environments [7–9].

This paper addresses this critical gap by introducing five novel AI-specific incident response metrics: Time-to-AI-Leverage (TAL), AI Amplification Ratio (AAR), Detection Coverage (DC), Containment Latency (CL), and Forensic Coverage Score (FCS). These metrics are designed to quantify response effectiveness in scenarios where AI systems are either weaponised by attackers or compromised as targets. Each metric captures a distinct dimension of AI-enhanced incident response, from exploitation speed to forensic comprehensiveness. Through simulation-based validation using realistic attack scenarios, we demonstrate how these metrics can expose deficiencies in existing IR processes and guide the development of AI-aware response strategies. The proposed framework provides organisations with quantitative tools to assess their preparedness for AI-driven threats enabling more effective incident response in increasingly AI-integrated enterprise environments.

II. Literature Review

Recent years have witnessed a surge in cyber threats augmented by artificial intelligence (AI), particularly through the use of generative models such as the large language models (LLMs). Research shows generative AI can significantly enhance the sophistication and scale of cyberattacks [10]. One study demonstrated that GPT-4 agents could autonomously exploit one-day vulnerabilities in real-world systems, achieving an 87% success rate when provided with CVE description [11]. These capabilities enable AI agents to conduct machine-speed attacks executing thousands of actions such as phishing, credential testing, and vulnerability scanning within seconds far exceeding the pace of human attackers. This acceleration compresses the window available for defenders to detect and respond. Simultaneously, AI amplifies the scale of attacks, allowing autonomous agents to target vast numbers of systems concurrently, thereby increasing the reach and impact of campaigns such as ransomware or data exfiltration [12]. This dual use of AI enhancing both speed and scale marks a fundamental shift in the cybersecurity threat landscape.

Generative AI has lowered barriers for sophisticated social engineering. Deepfakes and advanced text generation enable highly convincing impersonation and disinformation attacks [13]. For instance, in 2025 fraudsters used an AI-generated deepfake video of a CEO to trick employees into transferring $25 million at a UK firm [14]. Similar AI-driven deception was reported in Hong Kong, where scammers on a video call used deepfake avatars of executives to authorise a fraudulent $25 million transfer [15]. These cases underscore how AI can weaponize trust via synthetic media, resulting in severe financial and reputational damage. Attackers have also employed AI to generate convincing phishing content at scale. A Russian group, 'Midnight Blizzard,' in 2023 conducted phishing attacks via Microsoft Teams chats. Vendor reporting suggests these campaigns were AI-assisted [10], illustrating how generative models may be leveraged to bypass user suspicion. Likewise, AI chatbots have been used maliciously to engage victims in real-time, imitating customer support and harvesting credentials, all while handling many targets in parallel [16]. The net effect is that AI-enabled threats are more persuasive, pervasive, and rapid than traditional attacks. Beyond using AI as a tool, attackers are also targeting AI systems themselves. The rise of adversarial machine learning and prompt injection attacks represents a new front for incident responders. The MITRE ATLAS knowledge base (Adversarial Threat Landscape for AI) documents real-world cases of attacks on AI-enabled systems, such as data poisoning of Microsoft's Tay chatbot (which was manipulated into generating toxic outputs) and a camera "image classifier" being hijacked to misidentify inputs [17].

These novel techniques can cause AI systems to malfunction or divulge sensitive data, creating incidents that traditional cybersecurity tools might not recognise. Literature published between 2019 and 2025 indicates that AI has emerged both as a force-multiplier for attackers and as a new target for exploitation, challenging defenders on multiple fronts. The advent of AI-driven threats has also exposed gaps in traditional incident response (IR) frameworks and practices. Established frameworks like NIST SP 800-61 and ISO/IEC 27035 provide generalised guidance for handling cyber incidents, but they lack explicit considerations for AI-centric scenarios. Analysts have noted that cybersecurity processes have not been fully integrated into the AI development and deployment lifecycle [18]. This means organisations deploying AI often do so without equivalent investment in security monitoring or incident planning for those AI systems. In practice, AI models and data pipelines introduce complex, layered components that are not addressed in conventional IR playbooks. As a result, incidents involving AI (e.g. a compromised machine-learning model or an abused AI service) can catch response teams unprepared. For example, detecting a prompt injection or model poisoning attack may require monitoring AI-specific telemetry (like model outputs or confidence metrics), which falls outside the scope of traditional logging and intrusion detection [19]. This gap leaves a blind spot where breaches can go unnoticed or unprioritized until damage is done. Another challenge is the speed and adaptability of AI-driven attacks versus the relatively static nature of many IR processes.

Traditional incident response is often reactive, predicated on alerts and indicators after an attack has begun. AI-augmented threats, however, can morph and propagate faster than human-led analysis. As various literatures report, autonomous AI malware can continually rewrite itself to evade signature-based defences, and AI-directed botnets can change tactics in real time to dodge detection methods [12]. Current IR teams may struggle to investigate such fast-moving threats within the typical cycle of detect → analyse → contain. Recent research indicates that autonomous AI agents are capable of executing complex, multi-step attacks in seconds to minutes, thereby compressing the response window available to defenders [11]. This means the window for effective response is much narrower, yet many organisations have not adjusted their monitoring or staffing to 24/7, AI-speed operations. The result is that by the time an IR team convenes and reacts, an AI-propelled attack may have already achieved its objectives.

Despite extensive research on AI-driven threats, there remains a lack of structured, quantitative approaches to assess their operational impact during incident response. Various initiatives by MITRE ATLAS and AISIRT provide valuable taxonomies and guidance but, do not offer measurable indicators for detection, containment, or forensic completeness in AI-integrated environments. As organisations increasingly adopt AI, the absence of metrics to evaluate their influence or exploitation during incidents is a significant oversight. This paper addresses these gaps with a practical framework for assessing incident response performance in the context of AI-enhanced intrusions.

III. Metrics Overview

To operationalise the evaluation of AI-augmented incidents, we introduce five impacted metrics: Time-to-AI-Leverage (TAL), AI Amplification Ratio (AAR), Detection Coverage (DC), Containment Latency (CL), and Forensic Coverage Score (FCS). These metrics quantify on how adversaries exploit AI and how effectively organisations detect, contain, and investigate such activity. While conceptually aligned with different phases of the incident response lifecycle from initial access to containment and post-incident analysis, the metrics

are only loosely coupled to those phases such as, AI-driven threats blur boundaries and often surface asynchronously.

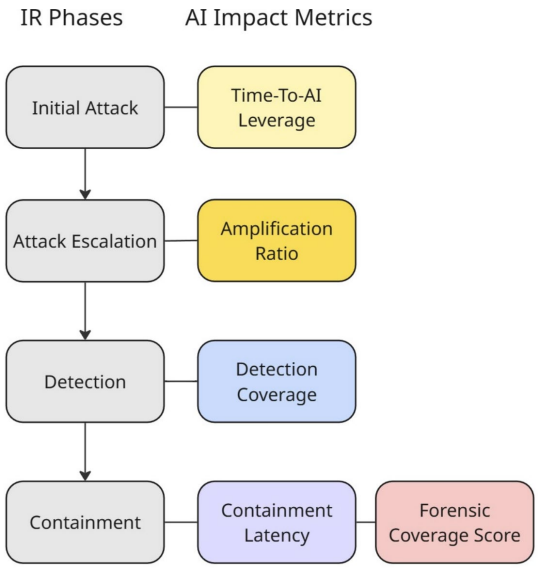

Figure 1. AI-specific impact metrics mapped to incident response phases.

A. Time-to-AI-Leverage (TAL)

TAL captures the elapsed time between the initial compromise and the first observed adversarial use of an AI system in minutes. This metric focuses on the critical inflection point where traditional compromise transitions into AI-enhanced escalation, such as the launch of LLM-generated phishing, automatic script execution, or AI-guided data triage. TAL is conceptually analogous to dwell time but scoped specifically to AI activation events. Equation (1) demonstrates how it can be calculated.

$$TAL = T_{\text{first-prompt}} - T_{\text{initial-access}} \quad (1)$$

Where:

$T_{\text{first-prompt}}$: Timestamp of the first malicious AI prompt
$T_{\text{initial-access}}$: Timestamp when attacker first gained access

TABLE I. TAL METRIC CLASSIFICATION

Classification	***Risk Level***	***Defender Implication***
Good (< 30 minutes)	Low	Fast, noisy attacks are more detectable. Automated systems must respond within this window.
Moderate (30–60 minutes)	Medium	Typical attack pace. Some detection opportunity, but attackers can still succeed quickly.
Poor (> 60 minutes)	High	Stealthy, patient attackers. Hard to detect, often leading to deep compromise.

This classification framework is based on data from leading cybersecurity vendors and reflects the operational tempo of modern cyber intrusions. The proposed time intervals; Good (<30 minutes), Moderate (30–60 minutes), and Poor (>60 minutes) are designed to benchmark defender response capabilities against attacker escalation speed. The Good category captures the upper bound of rapid but, detectable intrusions. CrowdStrike reports the fastest observed breakout time as 51 seconds, while ReliaQuest observed 27-minute escalations, both indicating that elite or automated attackers can act within this window [20-21]. The Moderate range aligns with the average breakout time of 48 minutes, consistently reported by both vendors, and represents the most common escalation window where defenders still have a realistic opportunity to detect and contain threats [20-21]. The Poor category encompasses intrusions that exceed 60 minutes before the first malicious action. Verizon's 2025 DBIR emphasises that the breach timeline is compressing and that delayed detection significantly increases breach severity [22]. Similarly, Palo Alto Networks' Unit 42 report found that 20% of breaches involved data theft within 1 hour, and 25% within 5 hours, underscoring the critical importance of early detection [23]. These findings justify poor (>60 minute) threshold as the point at which defenders are most likely to lose containment and attackers achieve critical objectives.

B. AI Amplification Ratio (AAR)

AAR measures the proportion of attacker actions that were AI-enhanced, counted as discrete operational actions (e.g., send phishing email, execute exfiltration query, create persistence key), not recipients, downstream effects, or data volume.

$$AAR = \frac{n_{AI}}{n_{\text{total}}} \quad (2)$$

Where

n_{AI}: Number of malicious actions conducted via AI

n_{total}: Total number of attacker actions.

Actions are counted, not outcomes or recipients. If a single AI invocation triggers a batch process, one action is recorded per send or execution, not per recipient or row. For chained prompts or tool calls, each distinct prompt or call that results in an action is counted as a separate AI-enhanced action. Retries of the same action within a single sequence are counted once; a new run at a later time is counted again. Where AI involvement cannot be

confirmed, the action is treated as conventional; if uncertainty remains, an AAR range should be reported in an appendix.

TABLE II. AAR METRIC CLASSIFICATION

Rating	*AAR Threshold*	*Interpretation*
Good	⩽1.0	AI had minimal or no influence; attacker actions were mostly conventional.
Moderate	1.01–2.0	AI contributed meaningfully to attack reach or efficiency.
Poor	> 2.0	AI significantly amplified the adversary's operational impact.

The AI Amplification Ratio (AAR) is a metric that may be used to quantify the proportional impact of AI-enhanced attacker actions relative to the total set of malicious activities observed during a cyber intrusion. It provides a scalable indicator of AI's operational influence by comparing the number of actions augmented or enabled by AI to the total number of attacker actions. Unlike normalised metrics constrained to a fixed range, AAR is designed as a ratio that may exceed 1.0, reflecting scenarios where, AI-enhanced actions disproportionately amplify attacker effectiveness such as through automation, decision acceleration, or multi-step chaining. The rating thresholds (e.g., ≤1.0 for Good, 1.01–2.0 for Moderate, and >2.0 for Poor) are not derived from existing literature, as no standardised benchmarks currently exist. Instead, they are proposed based on logical stratification of AI impact severity. This metric enables organisations to assess not only the presence of AI in adversarial operations but, also its relative contribution to breach escalation and impact.

C. Detection Coverage (DC)

DC measures the proportion of AI-related attacker behaviours—such as prompt injection, AI-assisted evasion and automated content generation—that are detected and logged by the organisation's security infrastructure. Unlike general detection metrics, DC focuses on telemetry and behavioural indicators specific to AI use and misuse. Equation (3) defines DC.

$$\mathrm{DC} = \frac{L_{\mathrm{detected}}}{L_{\mathrm{total}}} \tag{3}$$

Where:

L_{detected}: Logged or flagged LLM activities

L_{total}: Total number of AI-involved activities by the attacker

DC captures AI-specific observability and complements general detection metrics. Because no standard exists, we treat at least 80 per cent as Good, 60 to 79 per cent as Moderate, and less than 60 per cent as Poor; these bands indicate maturity rather than compliance.

TABLE III. DC METRIC CLASSIFICATION

Rating	*Threshold*	*Interpretation*
Good	⩾80%	Monitoring captures majority of AI-related activities.
Moderate	60–79%	Partial visibility; likely blind spots in AI agent logs or behaviours.
Poor	<60%	Major gaps; AI misuse largely undetected.

Detection Coverage (DC) evaluates an organisation's capacity to identify and log AI-specific adversarial behaviours, including prompt injection, automated evasion, and AI-generated content misuse. Unlike generic detection measures, DC focuses on telemetry and behavioural indicators that directly signal AI involvement, providing a sharper view of emerging threat vectors. Thresholds—Good (⩾80%), Moderate (60–79%), Poor (<60%)—are proposed in the absence of established benchmarks, enabling organisations to gauge both the presence of AI in attacks and the fidelity of their monitoring capabilities. By isolating AI-related detection performance, DC strengthens broader incident response metrics and informs the development of AI-aware defence strategies.

D. Containment Latency (CL)

CL is defined as the time taken from the detection of malicious activity to the successful restriction or revocation of the adversary's access to AI tools. As LLMs are often accessed via cloud services using federated identity, containment must address not only devices and user accounts, but, also the AI interfaces through which attackers may operate. This metric emphasises the need for AI-specific containment strategies. Equation (4) demonstrates how it can be calculated

$$\mathrm{CL} = T_{\mathrm{contain}} - T_{\mathrm{detection}} \tag{4}$$

Where:

T_{contain}: Time containment action was successfully executed

$T_{\mathrm{detection}}$: Time AI abuse was first detected

TABLE IV. CL METRIC CLASSIFICATION

Rating	*Threshold*	*Interpretation*
Good	⩽10 minutes	AI tools isolated quickly, limiting their use in propagation or evasion.
Moderate	11–20 minutes	Containment takes some time; AI may already be exploited.
Poor	>20 minutes	Significant delay in cutting AI access post-compromise.

Containment Latency (CL) is introduced as a critical metric to quantify the time elapsed between the detection of an AI-involved cyber intrusion and the successful containment of the threat. It captures the responsiveness of an organisation's incident response process, particularly in scenarios where AI

systems are either leveraged by attackers or compromised as targets. Unlike metrics that focus on detection or impact, CL emphasises operational agility and the effectiveness of containment protocols. This metric is justified by its ability to highlight delays in isolating compromised systems, which can significantly influence the overall damage and recovery trajectory. By measuring containment latency, organisations gain insight into the efficiency of their response mechanisms and can identify bottlenecks in decision-making or execution.

CL complements other AI-specific metrics by focusing on the temporal dimension of incident response. Its structured classification and practical relevance make it a valuable tool for evaluating and improving containment strategies in the face of AI-driven threats.

E. Forensic Coverage Score (FCS)

FCS evaluates the completeness and analysability of forensic artefacts associated with AI interactions during a breach. This includes prompt/response logs, AI decision outputs, usage timelines, and related metadata. A high FCS indicates that investigators were able to reconstruct the role of AI in the attack, which is crucial for root cause analysis, attribution, and future prevention. Equation (5) demonstrates how this can be calculated.

$$FCS = \frac{F_{retrievable}}{F_{expected}} \quad (5)$$

Where:

$F_{\text{retrievable}}$: Number of forensically accessible artefacts

F_{expected} : Total number of artefacts required for full reconstruction of AI abuse

Required artefact set. At minimum: (i) prompt and response logs; (ii) model identifiers/version and tool-use metadata; (iii) user/session identifiers (including federated identity claims); (iv) timestamps and request/response IDs; (v) output artefacts (e.g., files, summaries) and associated storage pointers.

Threshold rationale. With no public benchmark, ⩾ 80% denotes near-complete reconstruction, 60 – 79% partial, and < 60% significant blind spots intended as practical readiness bands.

TABLE V. FCS METRIC CLASSIFICATION

Rating	***Threshold***	***Interpretation***
Good	⩾80%	Post-incident review has strong evidence base, including AI-specific logs.
Moderate	60–79%	Key events and artefacts captured, but some gaps remain.
Poor	<60%	Limited forensic visibility into AI-specific behaviors.

Forensic Coverage Score (FCS) is introduced as a metric to assess the completeness and utility of forensic artefacts related to AI interactions during a cyber incident. It captures the extent to which investigators can reconstruct the role of AI in the attack, including access to prompt/response logs, AI-generated outputs, usage timelines, and associated metadata. Unlike metrics focused on detection or response speed, FCS emphasises post-incident visibility and the quality of evidence available for root cause analysis, attribution, and long-term mitigation. The classification thresholds; Good (⩾80%), Moderate (60 – 79%), and Poor (<60%) are proposed based on logical stratification of forensic completeness. This metric is justified by its ability to highlight gaps in forensic readiness, particularly in environments where AI systems are integrated into operational workflows. A high FCS indicates strong evidentiary support for understanding AI's involvement in the breach, while a low score signals critical blind spots that may hinder investigation and recovery. FCS complements the broader metric framework by focusing on the post-compromise phase of incident response. Its structured classification and emphasis on forensic depth make it a vital tool for evaluating organisational preparedness in the face of AI-augmented threats.

IV. SCENARIO EVALUATION

We construct and demonstrate application of the newly proposed AI-specific incident response metrics. Given the current scarcity of publicly available real-world data on AI-augmented cyberattacks, simulated scenarios provide a valuable framework. These simulations are carefully designed to reflect realistic threat vectors and operational outcomes, enabling a structured exploration of how AI can influence both attacker behaviour and defensive response. Importantly, these scenarios do not represent actual recorded incidents. Instead, they are constructed as plausible narratives developed to guide comprehension and practical application of the metrics in environments where empirical data is limited. Each scenario is evaluated using the proposed metrics; Time-to-AI-Leverage (TAL), AI Amplification Ratio (AAR), Detection Coverage (DC), Containment Latency (CL), and Forensic Coverage Score (FCS) to illustrate their diagnostic value and operational relevance in AI-integrated security contexts.

A. Scenario One

In this scenario, a financially motivated threat actor gains unauthorised access to a financial services firm's Microsoft 365 environment using valid corporate credentials acquired via the dark web. Within minutes of initial access, the attacker activates Microsoft Copilot an embedded AI assistant to orchestrate a targeted phishing campaign. The attacker uses Copilot to generate 45 high-fidelity phishing emails impersonating the firm's Chief Financial Officer (CFO). These emails are contextually enriched using data from shared files, calendar events, and recent email threads, significantly increasing their credibility and likelihood of success. The attacker's use of AI is both immediate and strategic, indicating a premeditated playbook with scripted prompts designed to exploit enterprise AI capabilities. Detection occurs 40 minutes after AI activity begins, triggered not by AI-specific telemetry but, by a geo-anomalous login event. Despite this, the attacker continues to

operate for over an hour before containment actions such as account suspension and API token revocation are executed. Post-incident analysis reveals that Copilot interactions were not logged, severely limiting forensic reconstruction.

TABLE VI. METRIC VALUES FOR SCENARIO ONE

Metric	*Value*	*Classification*
Time-to-AI-Leverage (TAL)	18 minutes	Good
AI Amplification Ratio (AAR)	0.75	Good
Detection Coverage (DC)	44.4%	Poor
Containment Latency (CL)	65 minutes	Poor
Forensic Coverage Score (FCS)	30%	Poor

In Scenario 1, the metrics evaluation reveals several key insights into the AI-augmented phishing escalation. The Time-to-AI-Leverage (TAL) is measured at 18 minutes, which is categorised as Good. This short timeframe indicates a rapid transition from initial access to AI exploitation, suggesting a high degree of automation and readiness in the attack process. The AI Amplification Ratio (AAR) stands at 0.75, also rated Good, implying that while AI played a significant role contributing to 75% of the attacker's actions in phishing generation; it remained within a manageable scope. The Detection Coverage (DC) is 44.4%, considered Poor, thus highlighting a substantial gap in the organisation's ability to detect AI-specific behaviours, such as those facilitated by Copilot prompt and response telemetry. Containment Latency (CL) registers at 65 minutes, again rated as Poor, underscoring an extended delay in isolation and revocation actions. This delay allowed continued leveraging of AI tools post-detection. Additionally, the Forensic Coverage Score (FCS) is 30%, Poor as well, pointing to significant deficiencies in logging and the forensic investigation of AI interactions. Together, these metrics underline strategic implications: they not only emphasise operational inefficiencies and vulnerabilities in detection and response but also suggest a crucial need for AI-aware telemetry, enhanced containment protocols, and comprehensive forensic logging to bolster resilience against sophisticated AI-driven threats. This scenario exemplifies the operational risks posed by embedded AI systems in enterprise environments. While the attacker's use of AI was tactically effective, the organisation's inability to detect, contain, and investigate AI-specific behaviours reveals systemic weaknesses. The scenario validates the need for AI-aware telemetry, faster containment protocols, and comprehensive forensic logging to support post-incident learning and resilience.

B. Scenario Two

This scenario explores a case of insider misuse of AI capabilities through persistent federated access. A contractor at a global law firm retains access to the firm's Google Workspace environment following offboarding, due to misconfigured identity federation. Over a six-day period, the contractor leverages Gemini AI to process over 200 sensitive legal documents. The first AI prompt is issued 210 minutes after the initial login, indicating a delayed but deliberate use of AI tools. The contractor uses Gemini to summarise and translate documents and further employs AI to redact client identifiers intentionally obscuring attribution and complicating post-incident analysis. The activity remains undetected for 18 hours, during which time the contractor continues to interact with sensitive content. Detection is eventually triggered by abnormal document access patterns, prompting the revocation of both identity and AI service access within 22 minutes. While Google Workspace audit logs provide partial visibility, Gemini session outputs are ephemeral and not retained, significantly limiting the organisation's ability to determine the full scope of data exposure or manipulation.

TABLE VII. METRIC VALUES FOR SCENARIO TWO

Metric	*Value*	*Classification*
Time-to-AI-Leverage (TAL)	210 minutes	Poor
AI Amplification Ratio (AAR)	0.89	Good
Detection Coverage (DC)	50.0%	Poor
Containment Latency (CL)	22 minutes	Moderate
Forensic Coverage Score (FCS)	75.3%	Moderate

In Scenario 2, the metrics evaluation sheds light on an insider threat exploiting federated access and AI tools. The Time-to-AI-Leverage (TAL) is measured at 210 minutes, classified as Poor. This long interval reflects the insider's strategic pacing, exploiting persistent access before direct AI engagement, which demonstrates value in distinguishing between internal and external threat timelines. The AI Amplification Ratio (AAR) is 0.89, rated as Good, indicating effective use of AI for summarisation, redaction, and translation tasks, without an overwhelming impact on propagation. The Detection Coverage (DC) is measured at 50.0%, classified as Poor, highlighting that half of the AI interactions remained undetected due to limitations in telemetry ingestion. The Containment Latency (CL) stands at 22 minutes, classified as Moderate, which suggests a relatively responsive containment that nonetheless has areas for improvement in integration between identity and AI session controls. The Forensic Coverage Score (FCS) is 75.3%, rated as Moderate, evidencing partial forensic success, aided by some logging yet hindered by ephemeral AI outputs. Together, these metrics illustrate both the complexity of insider AI misuse and the operational challenges in detection, containment, and forensic reconstruction.

This scenario illustrates the complex dynamics of insider misuse of AI in sensitive environments. While the contractor effectively exploited federated access, the law firm's failure to promptly detect, contain, and investigate AI-specific behaviours exposes systemic vulnerabilities. It highlights the need for AI-aware audit controls, integrated containment protocols, and comprehensive forensic logging to support post-incident learning and organisational resilience.

V. Comparative Analysis of Metrics

Comparing the two scenarios highlights recurring gaps in AI observability (DC), containment agility (CL), and forensic readiness (FCS), and shows how these dimensions interact. The external compromise weaponised AI within 18 minutes (TAL: Good), indicating premeditation and automation; by contrast, the insider misuse delayed first AI use to 210 minutes (TAL: Poor), underscoring TAL's value for distinguishing threat type and intent and suggesting that conventional dwell-time indicators may underestimate machine-speed escalation.

Although both scenarios yield a Good AAR (Scenario 1: 0.75; Scenario 2: 0.89), the form of impact differs materially. In Scenario 1, AI accelerated phishing content creation at scale, enhancing credibility and reach; in Scenario 2, AI was applied to document manipulation (summarising, redacting, translating) to obfuscate attribution. Hence AAR should be interpreted with operational context, not as a standalone scalar—similar ratios may reflect very different defensive implications.

Detection Coverage (DC) was Poor in both cases (Scenario 1: 44.4%; Scenario 2: 50.0%). In financial services, Copilot interactions were not logged, rendering AI activity effectively invisible; in legal, limited telemetry obscured Gemini outputs. These results indicate a broad absence of AI-specific observability and position DC as a direct proxy for AI logging maturity. Containment Latency (CL) separated organisational agility: 65 minutes in Scenario 1 (Poor) versus 22 minutes in Scenario 2 (Moderate). The contrast suggests containment efficiency in AI incidents depends on mature access orchestration and cross-domain revocation (identity ↔ cloud AI services), not merely on detecting the anomaly.

Forensic Coverage Score (FCS) was the most scenario-sensitive measure: 30% in Scenario 1 due to absent prompt/output logs versus 75.3% in Scenario 2 with partial Workspace logging. This evidences a growing gap driven by ephemeral AI interactions and cloud APIs, where traditional techniques are insufficient without AI-specific retention policies or native logging. Considered together, the metrics surface systemic patterns: low DC often precedes prolonged CL, and weak FCS hinders learning and eradication. In particular, the combination of short TAL and poor DC (Scenario 1) generates blind spots that extend adversary access to AI systems before defensive controls can be engaged. These dynamics reinforce reading the metrics collectively as interdependent signals of both technical and procedural readiness.

VI. Discussion and Implications

The integration of artificial intelligence into enterprise systems is reshaping the threat landscape, adversarial tactics, and incident response workflows. This paper proposes five AI-specific metrics; TAL, AAR, DC, CL, and FCS to quantify and evaluate an organisation's ability to detect, contain, and recover from AI-augmented incidents. Applied to two realistic but simulated case studies, these metrics provide a structured framework that exposes critical weaknesses in current practices and offers a reproducible readiness assessment methodology. A key strength of the framework lies in its ability to assign quantitative meaning to aspects of response that were previously treated qualitatively or overlooked. Existing incident response models, while foundational, were not designed for AI-enhanced operations, and traditional metrics such as mean time to detection or containment lack the specificity to capture AI-driven speed and obfuscation. In contrast, the proposed metrics offer tailored indicators that capture AI-specific patterns, including the timing of activation (TAL), operational amplification (AAR), and visibility of misuse within enterprise monitoring (DC). The framework also has limitations. Reliance on simulated data, while necessary due to the lack of public AI breach disclosures, limits empirical depth and may not fully reflect real-world dynamics. Variability in AI platform telemetry affects applicability; for instance, some services do not retain prompt data, making FCS calculation infeasible. A limitation is that classification thresholds are derived from expert judgement and scenario studies rather than longitudinal benchmarks; future empirical validation is essential. Despite these constraints, the metrics hold promise for both operational application and academic exploration. Future work should focus on empirical validation through controlled simulations and operational studies, refining thresholds and exploring interdependencies. Embedding researchers within security operations could provide valuable real-world telemetry and inform platform-specific extensions, such as adapting DC and FCS to tools like Microsoft Copilot, Google Gemini, or Salesforce AI. Integration with SIEM, SOAR, or XDR platforms would enable automated scoring and monitoring, embedding the framework as a diagnostic layer in the security stack. The metrics could also serve as indicators of SOC maturity and guide investment in AI observability.

Finally, the structured approach lends itself to governance and policy adoption. As regulatory bodies develop AI oversight, metrics such as Detection Coverage and Forensic Coverage could serve as auditable indicators of assurance. Collaboration with standards bodies such as NIST, ISO, or ENISA could translate the framework into policy-ready language, enabling broad adoption. In summary, this work introduces a foundational approach to quantifying AI's impact on post-compromise incident response. While early in development, the framework bridges a critical gap in practice. It enables defenders to move beyond anecdotal assessments toward evidence-based evaluations of readiness in AI-integrated environments. Continued refinement, validation, and integration will be essential to realising its full potential.

VII. Conclusion

This paper presents a structured framework of AI-specific incident response metrics to address the growing complexity of cyber incidents in environments where artificial intelligence plays an active role. By introducing and applying Time-to-AI-Leverage (TAL), AI Amplification Ratio (AAR), Detection Coverage (DC), Containment Latency (CL), and Forensic Coverage Score (FCS), the study provides a quantifiable lens for evaluating the speed, scope, and recoverability of AI-augmented threats. Simulated scenarios demonstrate the metrics' diagnostic value across both external automation-driven attacks and deliberate insider misuse, exposing systemic shortcomings in detection, containment, and forensic preparedness. These

findings affirm that traditional response models are insufficient alone and must evolve to address the machine-speed dynamics and obfuscation introduced by enterprise AI systems. Beyond their analytical utility, the proposed metrics offer a foundation for standardising AI-specific assessments, integrating into SOC dashboards, informing policy benchmarks, and supporting regulatory assurance as governance frameworks mature. While the scenarios validate the concept, further work is needed to refine thresholds through empirical data, extend platform-specific applicability, and embed the measures into operational and policy frameworks.

This paper addresses a critical gap in the literature by offering a structured framework to measure AI's impact during incident response, enabling defenders to lay the foundation for standardised, repeatable readiness assessments. The framework provides essential scaffolding for operationalising AI-aware readiness, guiding strategic investments, and shaping future standards. In an era where AI is both a tool and a target, these metrics represent a crucial step toward measurable, adaptive, and resilient incident response grounded in evidence and repeatability.

REFERENCES

[1] Forbes Advisor, "How businesses are using artificial intelligence," Forbes, Apr. 24, 2023. [Online]. Available: https://www.forbes.com/advisor/business/software/ai-in-business/

[2] A. K. Sood and S. Zeadally, "Malicious AI models undermine software supply-chain security," Commun. ACM, vol. 68, no. 5, pp. 42–49, May 2025. [Online]. Available: https://cacm.acm.org/research/malicious-ai-models-undermine-software-supply-chain-security/

[3] H. Chen and F. Koushanfar, "Toward robust deep learning against poisoning attacks," ACM Trans. Embedd. Comput. Syst., vol. 22, no. 3, Apr. 2023. [Online]. Available: https://dl.acm.org/doi/fullHtml/10.1145/3574159

[4] P. Cichonski, T. Millar, T. Grance, and K. Scarfone, Computer Security Incident Handling Guide, NIST Special Publication 800-61 Rev. 2, Aug. 2012. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r2.pdf

[5] International Organization for Standardization, ISO/IEC 27035-1:2023 Information Security Incident Management – Part 1: Principles of Incident Management, ISO, 2023.

[6] A. Nelson, S. Rekhi, M. Souppaya, and K. Scarfone, "Incident response recommendations and considerations for cybersecurity risk management," NIST SP 800-61 Rev. 3, Apr. 2025. [Online]. Available: https://csrc.nist.gov/pubs/sp/800/61/r3/final

[7] MITRE Corporation, "Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS)," 2023. [Online]. Available: https://atlas.mitre.org

[8] Carnegie Mellon University Software Engineering Institute, "AI Security Incident Response Team (AISIRT)," 2023. [Online]. Available: https://insights.sei.cmu.edu/news/aisirt-ai-security-incident-response/

[9] MITRE Corporation, "MITRE launches AI incident sharing initiative," Mar. 2024. [Online]. Available: https://www.mitre.org/news-insights/news-release/mitre-launches-ai-incident-sharing-initiative

[10] A. Talapatra, "The rise of AI-driven cyber attacks: Implications for modern security," Radware, 2025. [Online]. Available: https://www.radware.com/blog/application-protection/the-rise-of-ai-driven-cyber-attacks-implications-for-modern-security/

[11] R. Fang, R. Bindu, A. Gupta, Q. Zhan, and D. Kang, "LLM agents can autonomously hack websites," arXiv preprint, Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.06664

[12] MixMode Threat Research, "The rise of AI-driven cyberattacks: Accelerated threats demand predictive and real-time defenses," MixMode, May 2025. [Online]. Available: https://mixmode.ai/blog/the-rise-of-ai-driven-cyberattacks-accelerated-threats-demand-predictive-and-real-time-defenses/

[13] C. F. Ross, "AI in cybersecurity: How AI is impacting the fight against cybercrime," Akamai, May 22, 2025. [Online]. Available: https://www.akamai.com/blog/security/ai-cybersecurity-how-impacting-fight-against-cybercrime

[14] D. Elliott, "Cybercrime: Lessons learned from a $25m deepfake attack," World Economic Forum, Feb. 4, 2025. [Online]. Available: https://www.weforum.org/stories/2025/02/deepfake-ai-cybercrime-arup/

[15] K. Meda, "Identity theft is being fueled by AI & cyber-attacks," Thomson Reuters Institute, May 3, 2024. [Online]. Available: https://www.thomsonreuters.com/en-us/posts/government/identity-theft-drivers/

[16] L. Stanham, "Most common AI-powered cyberattacks," CrowdStrike, Jan. 16, 2025. [Online]. Available: https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/ai-powered-cyberattacks/

[17] L. Mohan, "MITRE ATLAS: Where real-world case studies bring cyber threats to life," LinkedIn, Mar. 22, 2025. [Online]. Available: https://www.linkedin.com/pulse/mitre-atlas-where-real-world-case-studies-bring-cyber-lija-mohan-taqlc/

[18] Carnegie Mellon University, "Leading AI security incident response," Software Engineering Institute, 2024. [Online]. Available: https://insights.sei.cmu.edu/annual-reviews/2024-year-in-review/leading-ai-security-incident-response/

[19] J. Zhang et al., "When LLMs meet cybersecurity: a systematic literature review," Cybersecurity, vol. 8, no. 1, p. 55, 2025, doi: 10.1186/s42400-2500361w.

[20] CrowdStrike, 2025 Global Threat Report. CrowdStrike, 2025. [Online]. Available: https://www.crowdstrike.com/global-threat-report

[21] ReliaQuest, Racing the Clock: Outpacing Accelerating Attacks. ReliaQuest,2025.[Online].Available: https://www.reliaquest.com/resources/reports/racing-the-clock/

[22] Verizon, "Threats are faster, smarter, and more personal," 2025 DBIR, SecureWorld, 2025. [Online]. Available: https://www.secureworld.io/industry-news/verizon-2025-data-breach-report

[23] Palo Alto Networks, "2025 Unit 42 Incident Response Report," 2025. [Online]. Available: https://live.paloaltonetworks.com/t5/community-blogs/2025-unit-42-incident-response-report-attacks-shift-to/ba-p/1225750