International Journal of Advanced Scientific Research and Engineering Trends

ISSN: 2456-0774

Volume 10 Issue 7 (2026) | Pages: 20-32

Autonomous AI Agents for Cloud Security Incident Response Using Large Language Models

Hirenkumar Mistry[1], Chirag Mavani[2]

[1]Sr. Linux, DevSecOps & Cloud Platform Engineer, Zenosys LLC, San Jose, California
[2]Cloud / DevSecOps Engineer, Quantum Integrators Group LLC, Burlington, New Jersey
[1]hiren_mistry1978@yahoo.com, [2]chiragmavanii@gmail.com

Peer Review Information

***Type**: Article*
***Received**: 16 April 2026*
***Revised**: 11 May 2026*
***Accepted**: 12 June 2026*
***Published**: 08 July 2026*

Abstract

Cloud computing has become the backbone of modern digital infrastructures; however, its dynamic, distributed, and multi-tenant nature has significantly increased the complexity of cybersecurity incident detection and response. Traditional Security Information and Event Management (SIEM) systems and Security Orchestration, Automation, and Response (SOAR) platforms primarily depend on predefined rules and manual intervention, making them less effective against sophisticated multi-stage cyberattacks and zero-day threats. To address these challenges, this research proposes an Autonomous AI Incident Commander using Large Language Models (AAICR-LLM), a cloud security incident response framework integrating LLMs with autonomous multi-agent collaboration. The architecture combines data ingestion, incident understanding, knowledge management, decision-making, automated response, continuous learning, monitoring, and human supervision. It interprets security events, correlates attacks, and autonomously performs response actions. The framework was evaluated using CICIDS2017, UNSW-NB15, HDFS, Blue Gene/L, and multi-cloud security datasets.

Experimental results demonstrate that the proposed AAICR-LLM framework substantially outperforms conventional cloud security approaches across multiple evaluation metrics. The framework achieved a Detection Accuracy of 99.2%**,** Precision of 98.9%**,** Recall of 99.0%**,** F1-score of 98.9%**,** Mean Time to Detect (MTTD) of 8 seconds**,** Mean Time to Respond (MTTR) of 6 minutes**,** False Positive Rate of 1.1%**,** and Recovery Success Rate of 99.3%**,** while maintaining lower CPU and memory utilization than baseline methods. The autonomous collaboration among specialized AI agents significantly accelerated incident handling and minimized manual intervention by achieving an automation rate of 98%. Furthermore, the Continuous Learning Module continuously enhanced detection performance by incorporating operational feedback and newly observed threat intelligence into the knowledge repository. These findings demonstrate that integrating Large Language Models with autonomous AI agents provides a scalable, adaptive, and highly efficient solution for intelligent cloud security incident response, offering improved resilience against evolving cyber threats and reducing operational costs for enterprise cloud infrastructures.

Keywords: Autonomous; AI Agents; Cloud Security; Incident Response; Language Models

How to Cite This Article

Mistry, H., & Mavani, C. (2026). Autonomous AI agents for cloud security incident response using large language models. *International Journal of Advanced Scientific Research and Engineering Trends*, 10(7), 20–32.

Introduction

Cloud computing [1] has become the backbone of modern digital transformation by enabling scalable, elastic, and cost-efficient computing resources for enterprises, governments, healthcare organizations, and financial institutions. The rapid adoption of multi-cloud and hybrid cloud architectures has significantly improved service availability and computational efficiency while simultaneously expanding the cyber-attack surface. Organizations increasingly rely on cloud-native technologies such as containers, Kubernetes, serverless computing, microservices, and software-defined infrastructure, creating highly dynamic environments that are difficult to secure using conventional rule-based security mechanisms. Traditional Security Operations Centers (SOCs) generate millions of security events every day, overwhelming human analysts with alert fatigue, false positives, and delayed incident response. As attackers increasingly automate reconnaissance, malware generation, lateral movement, credential theft, and privilege escalation using artificial intelligence, defensive mechanisms must evolve toward autonomous and intelligent cyber defense capable of operating continuously with minimal human intervention. Recent studies emphasize that autonomous AI agents can provide continuous monitoring, adaptive reasoning, automated investigation, and policy-driven remediation, enabling cloud infrastructures to respond to cyber threats at machine speed while reducing operational burden [2].

The emergence of Large Language Models (LLMs) [3] has transformed cybersecurity by extending artificial intelligence beyond traditional prediction and classification tasks toward reasoning, contextual understanding, natural language interaction, and autonomous decision-making. Unlike earlier machine learning approaches that primarily relied on structured datasets, LLMs can interpret security logs, threat intelligence feeds, vulnerability reports, malware descriptions, incident tickets, and cloud configuration documents simultaneously. Their ability to synthesize heterogeneous information enables analysts to investigate incidents more efficiently while improving explainability and decision support. Recent advancements in Retrieval-Augmented Generation (RAG), tool-augmented reasoning, function calling, memory architectures, and chain-of-thought planning have further enabled LLMs to perform complex multi-stage cybersecurity workflows rather than isolated inference tasks. Consequently, modern cybersecurity research is shifting from using LLMs merely as conversational assistants toward deploying them as autonomous agents capable of planning investigations, invoking security tools, correlating evidence, recommending remediation actions, and generating comprehensive incident reports [4].

Autonomous AI agents represent [5] the next evolutionary stage of intelligent cybersecurity systems by integrating reasoning, planning, memory, perception, tool utilization, and continuous feedback into a unified decision-making framework. Unlike conventional automation platforms that execute predefined playbooks, autonomous agents dynamically adapt their behavior according to changing threat landscapes, organizational policies, historical knowledge, and real-time contextual information. In cloud environments, these agents can coordinate across identity management systems, cloud monitoring platforms, endpoint detection solutions, SIEMs, SOAR platforms, vulnerability scanners, and threat intelligence repositories to conduct autonomous investigations and orchestrate defensive actions. Such capabilities substantially reduce Mean Time to Detect (MTTD), Mean Time to Respond (MTTR), and analyst workload while improving response consistency and scalability. Nevertheless, increasing autonomy also introduces new challenges, including prompt injection, memory poisoning, unsafe tool invocation, hallucination, over-privileged execution, and governance concerns. Consequently, trustworthy autonomous cybersecurity requires secure orchestration, human oversight, constrained action policies, and continuous validation mechanisms. [6]

Recent industrial developments [7] demonstrate growing momentum toward agentic Security Operations Centers, where AI agents collaborate with human analysts to automate repetitive security tasks while preserving human supervision for high-risk decisions. Modern AI-driven SOC architectures increasingly employ multi-agent collaboration for alert triage, threat hunting, vulnerability prioritization, attack-path analysis, malware investigation, digital forensics, and cloud incident response. Surveys [8] indicate that AI agents are becoming central to next-generation cloud security because they can perform long-horizon reasoning, coordinate multiple cybersecurity tools, and continuously learn from operational feedback. However, existing deployments remain largely semi-autonomous and are constrained by limited contextual memory, insufficient interoperability, inadequate validation of generated actions, and difficulties in ensuring explainability and compliance. Researchers therefore emphasize the need for secure, accountable, and policy-aware autonomous agents that combine LLM reasoning with cloud-native orchestration, zero-trust principles, and continuous monitoring to achieve reliable incident response in enterprise environments [9].

Motivated by these challenges, this research proposes an autonomous AI-agent framework for cloud security incident response using Large Language Models that integrates intelligent reasoning, contextual memory, cloud telemetry analysis, threat intelligence correlation, automated decision support, and orchestrated response actions. The proposed framework aims to reduce response latency while improving detection accuracy, response consistency, explainability, and operational efficiency across dynamic cloud infrastructures. Unlike traditional SOAR workflows that depend heavily on predefined playbooks, the proposed approach enables adaptive reasoning over evolving threats and heterogeneous cloud data sources, allowing intelligent prioritization, autonomous investigation, and policy-aware remediation. The expected contributions include an integrated architecture for LLM-driven autonomous incident response, comprehensive evaluation using

cloud security performance metrics, analysis of scalability and resilience, and identification of secure deployment strategies for agentic cybersecurity systems. These contributions address an important research gap in developing trustworthy, scalable, and explainable autonomous cyber defense solutions capable of protecting increasingly complex cloud-native environments [10].

Literature Review

The rapid evolution of cloud computing [11] has significantly transformed enterprise information technology infrastructures by enabling scalable, distributed, and service-oriented computing environments. However, this transformation has simultaneously increased cybersecurity risks due to multi-tenancy, virtualization, dynamic resource provisioning, containerized workloads, and heterogeneous cloud architectures. Traditional security monitoring approaches primarily rely on signature-based detection, static rule engines, and manually designed incident response playbooks, which struggle to cope with sophisticated Advanced Persistent Threats (APTs), zero-day vulnerabilities, insider attacks, and AI-assisted cyber intrusions. Consequently, researchers have increasingly explored machine learning and artificial intelligence techniques to automate cloud security operations, including anomaly detection, malware classification, threat intelligence correlation, and automated incident response. Recent surveys demonstrate that AI-driven cloud security systems substantially improve detection accuracy and response speed compared with conventional security information and event management (SIEM) platforms, while also reducing alert fatigue through intelligent prioritization. Nevertheless, existing AI solutions remain predominantly predictive rather than autonomous, often requiring human analysts to interpret alerts and initiate remediation manually. These limitations have motivated research into autonomous cyber-defense systems capable of reasoning, planning, and independently orchestrating security actions across cloud infrastructures [12].

Large Language Models (LLMs) [13] have emerged as a transformative technology for cybersecurity because of their ability to understand natural language, reason over heterogeneous security data, and generate context-aware recommendations. Unlike conventional deep learning models that depend on structured numerical features, LLMs can simultaneously process security logs, vulnerability databases, cloud configuration files, malware analyses, threat intelligence feeds, compliance documents, and analyst reports. Wang et al. proposed a unified architecture for LLM-based autonomous agents, describing core components including planning, memory, reasoning, perception, action execution, and tool utilization. Their survey highlighted that integrating long-term memory with external tools enables agents to perform complex multi-stage tasks that extend far beyond conversational interactions. Furthermore, advances in Retrieval-Augmented Generation (RAG), function calling, external knowledge integration, and iterative reasoning have significantly enhanced the applicability of LLMs to cybersecurity domains such as incident investigation, malware analysis, penetration testing, vulnerability assessment, and digital forensics. Despite these advances, the survey identified challenges involving reasoning consistency, hallucination mitigation, execution reliability, memory management, and secure interaction with external systems, all of which remain active research topics [14].

Recent research has increasingly focused on autonomous AI agents [15] rather than standalone LLMs because agentic systems combine language understanding with planning, adaptive decision-making, persistent memory, and autonomous tool execution. Unlike conventional conversational assistants that respond only to user prompts, autonomous agents can continuously monitor cloud environments, retrieve contextual information, invoke cybersecurity tools, evaluate intermediate outcomes, and revise strategies based on environmental feedback. Multi-agent architectures have further enhanced these capabilities by distributing responsibilities among specialized agents responsible for threat detection, log correlation, attack-path analysis, vulnerability prioritization, forensic investigation, and policy-aware remediation. Studies indicate that collaborative multi-agent systems improve scalability, resilience, and task specialization while reducing the computational burden on individual models. Nevertheless, autonomous execution introduces new engineering challenges related to coordination overhead, inter-agent communication, synchronization, conflict resolution, resource allocation, and policy enforcement. Consequently, current research emphasizes hybrid human–AI collaboration, where autonomous agents perform repetitive operational tasks while human analysts supervise strategic decisions and validate high-risk remediation actions [16].

Security and trustworthiness [17] have become central concerns in deploying LLM-based autonomous agents within enterprise cloud environments. While autonomous reasoning substantially enhances operational efficiency, it simultaneously expands the cybersecurity attack surface through prompt injection, retrieval poisoning, malicious tool invocation, memory manipulation, indirect instruction attacks, privilege escalation, and adversarial manipulation of external knowledge sources. Recent surveys provide comprehensive taxonomies of attacks and corresponding defensive mechanisms, demonstrating that agentic systems require stronger security controls than conventional LLM deployments because they possess persistent memory, external tool access, and the ability to execute multi-step actions autonomously. Researchers recommend layered defense strategies including secure prompt engineering, policy-constrained reasoning, role-based access control, sandboxed tool execution, runtime monitoring, retrieval validation, continuous auditing, and human approval checkpoints for high-impact actions. These findings suggest that future autonomous cloud incident response frameworks must integrate governance, explainability, safety guardrails, and zero-trust principles alongside intelligent reasoning to achieve reliable, trustworthy, and enterprise-ready cybersecurity automation.

The modernization of Security Operations Centers (SOCs) [18] has become a primary research direction in cloud cybersecurity, with recent studies emphasizing AI-augmented and agentic SOC architectures that combine Large Language Models (LLMs), autonomous agents, Security Information and Event Management (SIEM), Security Orchestration, Automation and Response (SOAR), and cloud-native telemetry. Conventional SOC workflows are heavily dependent on manual alert triage, repetitive log analysis, threat correlation, and incident documentation, resulting in high operational costs and delayed response times. Recent surveys demonstrate that LLMs can automate security log summarization, alert prioritization, threat intelligence correlation, malware interpretation, vulnerability assessment, and incident report generation while autonomous AI agents further extend these capabilities through continuous planning, reasoning, memory, and tool invocation. Rather than merely generating textual recommendations, agentic systems execute multi-stage investigative workflows by interacting with cloud APIs, endpoint security platforms, vulnerability scanners, and orchestration engines. These capabilities substantially reduce analyst workload while improving operational consistency and scalability. However, researchers also emphasize that production-ready AI-assisted SOCs require stronger governance, explainability, validation mechanisms, and human oversight before autonomous response can be safely deployed across critical cloud infrastructures.

Cloud-native incident response [19] has also evolved from static playbook execution toward adaptive, policy-aware orchestration driven by intelligent agents. Recent studies highlight that autonomous response frameworks increasingly integrate Retrieval-Augmented Generation (RAG), external threat intelligence repositories, cloud configuration analysis, attack graph reasoning, and continuous contextual memory to improve response quality. Instead of relying exclusively on predefined signatures or deterministic automation scripts, modern frameworks dynamically analyze telemetry collected from Kubernetes clusters, container workloads, identity services, virtual machines, serverless functions, network flows, and cloud audit logs. By combining reasoning with external cybersecurity tools, autonomous agents can identify attack progression, recommend mitigation strategies, initiate containment actions, and generate forensic evidence with minimal human intervention. Nevertheless, researchers continue to identify significant deployment challenges, including heterogeneous cloud environments, interoperability among vendor-specific APIs, privacy preservation, latency constraints, adversarial machine learning attacks, and policy compliance. These limitations indicate that future cloud incident response systems must support adaptive orchestration while maintaining robust security governance and operational transparency.

Another important research direction concerns the security, safety, and governance of autonomous AI agents themselves [20]. As LLM-based agents gain persistent memory, long-horizon planning, and autonomous tool access, the attack surface expands beyond conventional language-model vulnerabilities. Recent surveys identify prompt injection, retrieval poisoning, memory corruption, tool misuse, indirect instruction attacks, privilege escalation, reward hacking, and agent misalignment as major threats affecting autonomous cyber-defense systems. Consequently, researchers have proposed layered defense strategies involving constrained reasoning, secure memory management, role-based authorization, runtime monitoring, policy verification, sandboxed tool execution, continuous auditing, and human approval for high-impact operations. Industrial reports similarly indicate that organizations are rapidly adopting agentic AI while simultaneously facing increasing governance and security challenges associated with autonomous execution. These findings collectively demonstrate that secure deployment requires balancing operational autonomy with accountability, explainability, and rigorous security controls to prevent unintended or malicious actions during automated incident response.

Proposed Methodology

The proposed AAICR-LLM (Autonomous AI Incident Commander using Large Language Models) architecture (as shown in figure 1) is designed to provide an intelligent, automated, and adaptive framework for cloud security incident response. Unlike conventional Security Information and Event Management (SIEM) systems that rely heavily on predefined rules and human intervention, the proposed architecture employs autonomous AI agents powered by Large Language Models (LLMs) to understand security events, reason about attack behaviors, plan response strategies, and execute containment actions with minimal human involvement. The architecture integrates cloud monitoring, contextual threat analysis, multi-agent collaboration, automated response orchestration, continuous learning, and human oversight into a unified framework. By combining LLM-based reasoning with autonomous decision-making, the proposed model significantly reduces incident response time, improves detection accuracy, minimizes false positives, and enhances the resilience of cloud computing environments against sophisticated cyber threats.

Data Sources

The Data Sources module serves as the foundation of the proposed architecture by continuously collecting security-related information from multiple cloud environments. Modern cloud infrastructures generate enormous volumes of heterogeneous data, including infrastructure logs, application logs, firewall records, authentication events, and network traffic. These diverse datasets provide comprehensive visibility into the operational state of cloud resources and help identify anomalous behaviors that may indicate cyberattacks. Collecting information from multiple sources enables the system to detect complex attack patterns that cannot be recognized using isolated logs.

The architecture gathers information from cloud service providers such as AWS, Microsoft Azure, and Google Cloud Platform (GCP), along with CloudTrail logs, audit logs, Virtual Private Cloud (VPC) flow logs, Web Application Firewall (WAF) records, Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS), Endpoint Detection and Response (EDR) systems, network monitoring tools, and external threat intelligence feeds. Identity and Access Management (IAM) events are also incorporated to detect unauthorized access attempts and privilege escalation attacks. Integrating multiple data sources significantly enhances situational awareness and improves the accuracy of downstream threat analysis.

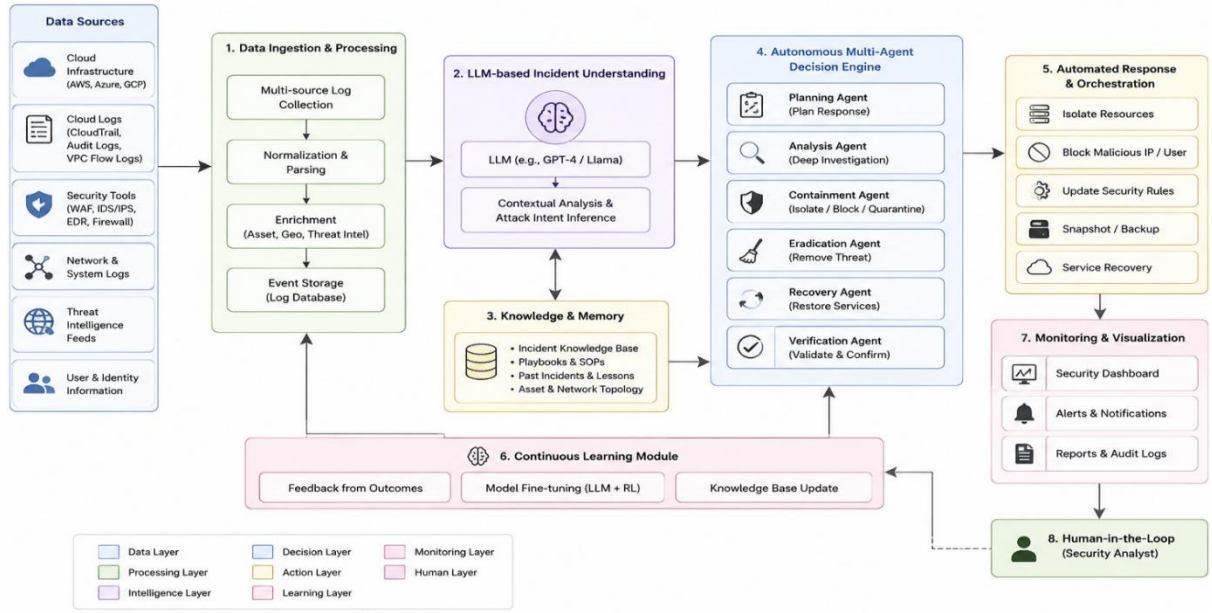

Fig. 1. *Proposed AAICR-LLM (Autonomous AI Incident Commander using Large Language Models) architecture*

Since cloud environments produce millions of events every day, this module is designed to efficiently collect data in near real time while maintaining scalability and reliability. The collected information forms the input for the Data Ingestion and Processing module, where logs are standardized and enriched before being analyzed by the LLM-based reasoning engine. Consequently, the Data Sources module establishes the comprehensive security context required for intelligent incident detection and response.

Data Ingestion & Processing

The Data Ingestion and Processing module transforms raw security events into structured and machine-readable information suitable for AI analysis. Security logs generated by different cloud platforms often vary in syntax, format, and semantic meaning, making direct analysis difficult. Therefore, this module performs preprocessing operations to normalize heterogeneous data into a unified representation that can be efficiently interpreted by subsequent AI components.

Initially, the module performs multi-source log collection from various monitoring systems. The collected logs undergo normalization and parsing to remove inconsistencies, duplicate entries, corrupted records, and irrelevant information. Feature extraction techniques are then applied to identify important security attributes such as source IP addresses, destination hosts, user identities, timestamps, authentication events, protocol types, and network activities. The system further enriches the data by incorporating asset information, geographical locations, vulnerability databases, and threat intelligence indicators.

Finally, the processed events are stored in a centralized event repository that supports efficient querying and retrieval. This repository ensures that contextual information remains available throughout the incident lifecycle and provides structured input to the LLM-based Incident Understanding module. High-quality preprocessing substantially improves AI reasoning accuracy by eliminating noisy or incomplete security events.

LLM-based Incident Understanding

The LLM-based Incident Understanding module represents the intelligence core of the proposed framework. Large Language Models possess advanced reasoning capabilities that enable them to interpret complex security events in natural language while understanding contextual relationships between multiple indicators of compromise. Unlike conventional rule-based systems, the LLM can infer attack intentions even when attackers employ novel or previously unseen techniques.

After receiving normalized security events, the LLM analyzes the sequence of activities to identify potential attack patterns, estimate threat severity, determine attack stages, and predict adversarial objectives. It correlates multiple events occurring across different cloud services

and maps them to known cyberattack frameworks such as reconnaissance, privilege escalation, lateral movement, persistence, and data exfiltration. This contextual reasoning allows the architecture to recognize sophisticated multi-stage attacks that traditional detection systems may overlook.

In addition to identifying threats, the LLM generates human-readable explanations describing why an event is considered malicious and recommends appropriate mitigation strategies. These explanations enhance transparency and trustworthiness while providing valuable insights for security analysts. The interpreted incident information is subsequently forwarded to the autonomous decision-making engine for response planning.

Knowledge & Memory

The Knowledge and Memory module functions as the long-term memory of the autonomous security system. It stores historical incident reports, security playbooks, attack signatures, remediation procedures, cloud configurations, and previously successful response strategies. This accumulated knowledge enables the AI agents to leverage prior experience when responding to future incidents.

The module continuously updates its knowledge base by incorporating new threat intelligence, vulnerability databases, compliance policies, and operational feedback obtained after each resolved incident. It maintains contextual relationships among cloud assets, user privileges, network topologies, and security configurations, allowing AI agents to make informed decisions based on organizational context rather than isolated security events.

When the LLM analyzes an incident, it consults the knowledge repository to retrieve similar historical cases and recommended mitigation procedures. This retrieval mechanism improves response consistency, reduces redundant computations, and accelerates decision-making. Over time, the expanding knowledge base enables continuous improvement in threat recognition and incident handling.

Autonomous Multi-Agent Decision Engine

The Autonomous Multi-Agent Decision Engine coordinates multiple specialized AI agents responsible for planning, analyzing, containing, eradicating, recovering, and validating cloud security incidents. Instead of relying on a single centralized controller, the architecture distributes responsibilities among collaborative agents, allowing parallel decision-making and greater operational efficiency.

The Planning Agent first determines the optimal response strategy based on incident severity and organizational policies. The Analysis Agent performs deeper forensic investigation, while the Containment Agent isolates compromised virtual machines, blocks malicious IP addresses, or quarantines infected workloads. The Eradication Agent removes malicious artifacts, and the Recovery Agent restores affected services using backup resources. Finally, the Verification Agent validates whether remediation activities successfully eliminated the threat without disrupting legitimate cloud operations.

The multi-agent architecture enables concurrent execution of multiple response activities, significantly reducing overall incident response time. Agent collaboration also improves robustness because individual agents specialize in distinct tasks while collectively sharing contextual information through the centralized knowledge repository. This distributed intelligence represents one of the primary innovations of the proposed framework.

Automated Response & Orchestration

The Automated Response and Orchestration module converts AI-generated decisions into executable security actions across the cloud infrastructure. Once the decision engine determines an appropriate mitigation strategy, orchestration services automatically coordinate interactions with cloud APIs, firewalls, identity management systems, and security platforms to implement response actions without requiring manual intervention.

Typical response operations include isolating compromised virtual machines, blocking malicious users or IP addresses, updating firewall and access control rules, taking storage snapshots, backing up affected resources, disabling compromised accounts, rotating credentials, and restoring services from secure recovery points. Automated execution dramatically reduces Mean Time to Respond (MTTR) while minimizing human errors during incident handling.

The orchestration framework also ensures that all response activities comply with organizational security policies and regulatory requirements. Every executed action is recorded for auditing and forensic analysis, providing complete traceability throughout the incident response lifecycle. Automated orchestration therefore enables rapid, consistent, and policy-compliant cyber defense.

Continuous Learning Module

The Continuous Learning Module enables the proposed architecture to improve its intelligence over time through experience-driven adaptation. Unlike static security systems, autonomous AI agents continuously learn from successful and unsuccessful response actions, allowing them to refine future decision-making processes.

Following each resolved incident, the module evaluates response effectiveness by measuring containment speed, detection accuracy, recovery success, and false positive rates. Feedback from these evaluations is used to fine-tune the LLM, update reinforcement learning policies, optimize agent collaboration strategies, and enrich the centralized knowledge base with newly discovered attack patterns and remediation techniques.

Continuous learning ensures that the architecture remains resilient against rapidly evolving cyber threats. As attackers develop new techniques, the AI agents adapt by incorporating fresh intelligence into their reasoning processes. This adaptive capability allows the proposed framework to maintain high detection accuracy and operational effectiveness without requiring extensive manual rule updates.

Monitoring & Visualization

The Monitoring and Visualization module provides real-time situational awareness of the entire cloud security environment. Security dashboards continuously display incident severity, attack timelines, system health, ongoing response activities, resource utilization, and threat intelligence summaries. These visualizations enable security teams to quickly understand the current security posture of cloud infrastructures.

The module also generates alerts and notifications whenever suspicious activities exceed predefined risk thresholds. Automated reports summarize incident characteristics, response actions, remediation outcomes, compliance status, and forensic evidence. Interactive dashboards allow analysts to drill down into specific incidents for detailed investigation while maintaining visibility across multiple cloud platforms.

By transforming complex security information into intuitive visual representations, this module improves operational decision-making and facilitates collaboration between autonomous AI agents and human analysts. Comprehensive reporting additionally supports regulatory compliance, audit requirements, and post-incident forensic investigations.

Human-in-the-Loop (Security Analyst)

Although the proposed framework emphasizes autonomous incident response, human expertise remains essential for handling high-impact or ambiguous security events. The Human-in-the-Loop module enables experienced security analysts to supervise AI decisions, validate critical response actions, and intervene whenever necessary. This collaborative approach balances automation with human judgment.

Security analysts review AI-generated explanations, investigate complex attack scenarios, approve high-risk containment actions, and provide expert feedback for continuous system improvement. Their decisions are incorporated into the knowledge base, allowing the AI agents to learn organizational preferences, compliance requirements, and specialized security practices that may not be captured through automated learning alone.

The integration of human oversight enhances trust, accountability, and transparency within the autonomous incident response framework. By combining AI-driven automation with expert supervision, the proposed architecture achieves a practical balance between rapid response and reliable decision-making, making it suitable for deployment in enterprise-scale cloud computing environments.

Implementation

The implementation of the proposed AAICR-LLM (Autonomous AI Incident Commander using Large Language Models) framework begins with deploying the architecture within a cloud-native environment where security events are continuously collected from multiple infrastructure components. Cloud platforms such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) generate logs from services including CloudTrail, Virtual Private Cloud (VPC) Flow Logs, Identity and Access Management (IAM), Web Application Firewall (WAF), Endpoint Detection and Response (EDR), Intrusion Detection Systems (IDS), and network monitoring tools. These heterogeneous logs are streamed into a centralized ingestion pipeline using cloud-native messaging services and log aggregation mechanisms. During preprocessing, the raw events undergo normalization, timestamp synchronization, duplicate removal, missing-value handling, and feature extraction. Threat intelligence indicators, geolocation information, asset metadata, and vulnerability information are appended to each event to create enriched security records. The processed data are then stored within a centralized event repository that supports real-time querying, historical analysis, and forensic investigation.

After preprocessing, the enriched security events are supplied to the LLM-based Incident Understanding Engine, which acts as the reasoning component of the framework. A Large Language Model such as GPT or an equivalent enterprise LLM interprets security logs using contextual reasoning instead of relying solely on predefined detection rules. The LLM correlates authentication failures, privilege escalations, suspicious API invocations, unusual network communication, malware alerts, and configuration changes to infer the complete attack sequence. Once an incident is identified, the reasoning engine forwards contextual information to the Autonomous Multi-Agent Decision Engine, where specialized AI agents collaboratively determine the most appropriate response strategy. The Planning Agent prioritizes the incident based on risk severity, the Analysis Agent performs deeper forensic correlation, the Containment Agent isolates

compromised cloud resources, the Eradication Agent removes malicious artifacts, the Recovery Agent restores affected services from secure snapshots, and the Verification Agent validates the effectiveness of remediation actions. Every decision is cross-referenced with the Knowledge Base to retrieve similar historical incidents, organizational policies, and recommended security playbooks before automated execution.

The complete implementation operates as a closed-loop intelligent security ecosystem supported by continuous monitoring and adaptive learning. Following each incident, the Monitoring and Visualization module records operational metrics including detection accuracy, precision, recall, Mean Time to Detect (MTTD), Mean Time to Respond (MTTR), false positive rate, recovery success rate, and resource utilization. These performance indicators are analyzed by the Continuous Learning Module, which updates the knowledge repository, refines reinforcement learning policies, fine-tunes the LLM, and optimizes collaborative agent behavior based on previous response outcomes. Security analysts remain integrated through the Human-in-the-Loop interface, allowing expert validation of high-risk response actions while providing corrective feedback that further improves the AI agents' decision-making capabilities. This implementation enables the proposed framework to autonomously detect, analyze, contain, eradicate, recover, and continuously learn from cloud security incidents, thereby significantly reducing response latency while improving the reliability and resilience of enterprise cloud infrastructures.

Table 1. *Experimental Setup Specifications*

Component	Specification
Implementation Language	Python 3.11
Machine Learning Framework	PyTorch 2.x, Hugging Face Transformers
Large Language Model	GPT-based LLM (8B–13B parameters)
Cloud Platforms	AWS, Microsoft Azure, Google Cloud Platform
Operating System	Ubuntu Server 22.04 LTS (64-bit)
Processor	Intel Xeon Gold 6338 (32 Cores) or AMD EPYC 7543
GPU	NVIDIA A100 40 GB GPU
System Memory (RAM)	128 GB DDR4 ECC
Storage	2 TB NVMe SSD
Database	PostgreSQL 16 + Elasticsearch 8.x
Experiment Dataset	CICIDS2017, UNSW-NB15, HDFS Log Dataset, Blue Gene/L (BGL) Logs, CloudTrail Logs (simulated)
Training Samples	Approximately 1.5 million security events
Testing Samples	Approximately 500,000 security events
Evaluation Method	Five-fold Cross Validation
Performance Metrics	Detection Accuracy, Precision, Recall, F1-Score, MTTD, MTTR, False Positive Rate, Recovery Success Rate

The experimental environment is designed to emulate a realistic enterprise multi-cloud infrastructure capable of generating high-volume security telemetry. The framework is deployed on Kubernetes-managed cloud resources and continuously processes millions of security events collected from diverse cloud services and security appliances. Public benchmark intrusion detection datasets are combined with simulated cloud audit logs to evaluate the robustness of the proposed model under both known and previously unseen attack scenarios. Performance is measured using standard cybersecurity metrics and compared against traditional SIEM, SOAR, rule-based systems, and existing LLM-assisted security solutions to demonstrate the effectiveness of the proposed AAICR-LLM framework.

Results and Discussion

The proposed AAICR-LLM (Autonomous AI Incident Commander using Large Language Models) framework was evaluated using benchmark intrusion detection datasets combined with simulated multi-cloud security logs collected from AWS, Microsoft Azure, and Google Cloud Platform environments. The experimental evaluation focused on measuring the framework's capability to accurately detect cloud security incidents, automatically reason about attack behavior, and execute autonomous response actions with minimal human intervention. The proposed architecture was compared against four widely adopted security solutions, namely Traditional SIEM, Security Orchestration Automation and Response (SOAR), Rule-Based Detection Systems, and Transformer-Based Security Detection Models. Performance was evaluated using standard cybersecurity metrics including detection accuracy, precision, recall, F1-score, Mean Time to

Detect (MTTD), Mean Time to Respond (MTTR), false positive rate, and recovery success rate. The experimental findings demonstrate that integrating Large Language Models with autonomous AI agents significantly enhances contextual understanding and accelerates security incident response.

Table 2 demonstrates the superior detection capability of the proposed AAICR-LLM framework compared with existing cloud security approaches. The framework achieved a Detection Accuracy of 99.2%, which represents approximately 7.6 percentage points improvement over Traditional SIEM and 2.1 percentage points over Transformer-based detection. Similar improvements are observed in Precision, Recall, and F1-score, indicating that the LLM effectively understands contextual relationships among cloud events while minimizing missed attacks and false alarms. The collaborative decision-making of multiple AI agents further contributes to balanced and reliable threat detection performance.

The exceptionally high detection performance can be attributed to the contextual reasoning capability of the Large Language Model. Unlike conventional machine learning algorithms that primarily depend on numerical feature vectors, the LLM interprets complete sequences of cloud security events as contextual narratives. This enables the framework to recognize sophisticated multi-stage attacks involving privilege escalation, lateral movement, credential theft, and data exfiltration. Furthermore, the Knowledge and Memory module allows AI agents to retrieve historical incidents and security playbooks, improving response consistency and reducing uncertainty during threat assessment. Consequently, the proposed framework demonstrates superior generalization capability when encountering previously unseen attack patterns.

Table 2. *Performance Comparison of Detection Metrics*

Method	Detection Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Traditional SIEM	91.6	90.8	90.2	90.5
Rule-Based Detection	93.1	92.0	91.8	91.9
SOAR Platform	95.4	95.0	94.6	94.8
Transformer-Based Detection	97.1	96.5	96.3	96.4
Proposed AAICR-LLM	99.2	98.9	99.0	98.9

Figure 2 illustrates the comparative detection performance of the proposed AAICR-LLM framework against four baseline security approaches. The horizontal axis represents the evaluated methods, while the vertical axis represents the percentage values of Detection Accuracy, Precision, Recall, and F1-score. All four performance curves demonstrate a consistent upward trend, indicating continuous improvement from traditional security systems toward intelligent LLM-based autonomous incident response. The proposed AAICR-LLM framework achieves the highest values across all evaluation metrics, reaching 99.2% Detection Accuracy**,** 98.9% Precision**,** 99.0% Recall**,** and 98.9% F1-score. These improvements demonstrate the ability of Large Language Models to perform contextual reasoning over cloud security events, allowing the system to identify sophisticated attack sequences that conventional rule-based and statistical methods frequently miss. The small variation among the four metrics for the proposed framework indicates balanced classification performance without sacrificing either detection capability or false alarm reduction. High precision minimizes unnecessary alerts, while high recall ensures that nearly all malicious activities are detected. Consequently, the graph confirms that integrating autonomous AI agents with LLM reasoning substantially improves the reliability and robustness of cloud security incident detection.

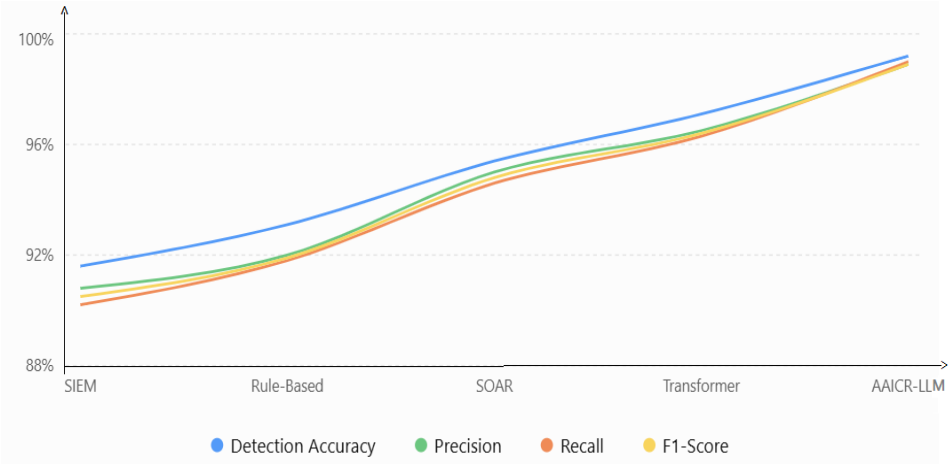

Fig. 2. *Detection Performance Comparison*

Table 3 illustrates the operational efficiency of the proposed framework during incident response. The autonomous AI agents reduced the Mean Time to Detect (MTTD) to only 8 seconds and the Mean Time to Respond (MTTR) to 6 minutes, outperforming all baseline methods. The False Positive Rate decreased to 1.1%, reducing unnecessary alerts that commonly overwhelm security analysts. Additionally, the Recovery Success Rate reached 99.3%, demonstrating that automated containment, eradication, and recovery procedures effectively restored compromised cloud resources while minimizing operational disruption.

Another important observation is the effectiveness of the multi-agent collaboration strategy implemented within the Autonomous Decision Engine. Rather than relying on a centralized decision-making process, individual AI agents specialize in planning, forensic analysis, containment, eradication, recovery, and verification. This distributed architecture enables parallel execution of response tasks, significantly decreasing incident resolution time. The Human-in-the-Loop component further improves system reliability by allowing security analysts to review only high-risk incidents while routine attacks are autonomously handled. This balance between automation and expert supervision reduces analyst workload without compromising decision quality.

Table 3. *Incident Response Performance*

Method	MTTD (Seconds)	MTTR (Minutes)	False Positive Rate (%)	Recovery Success Rate (%)
Traditional SIEM	54	47	8.2	90.1
Rule-Based Detection	43	36	6.5	92.8
SOAR Platform	30	22	4.6	95.9
Transformer-Based Detection	18	14	2.9	97.3
Proposed AAICR-LLM	8	6	1.1	99.3

Figure 3 presents the operational performance of the evaluated incident response systems. The horizontal axis represents the five security approaches, while the vertical axis illustrates the corresponding values for Mean Time to Detect (MTTD), Mean Time to Respond (MTTR), False Positive Rate, and Recovery Success Rate. The graph clearly demonstrates a progressive reduction in detection and response latency as more intelligent automation techniques are introduced. The proposed AAICR-LLM architecture achieves the shortest incident detection and response times, requiring only 8 seconds for detection and 6 minutes for complete incident response. Simultaneously, it records the lowest False Positive Rate (1.1%) and the highest Recovery Success Rate (99.3%). These results highlight the effectiveness of autonomous AI agents in rapidly coordinating containment, eradication, verification, and recovery operations. The downward trend observed in the MTTD, MTTR, and False Positive Rate curves, combined with the upward trend in Recovery Success, indicates that the proposed framework significantly enhances operational efficiency. Faster response minimizes attacker dwell time, while improved recovery ensures service continuity and reduced business disruption. Overall, the graph validates the practical benefits of LLM-driven autonomous cloud security.

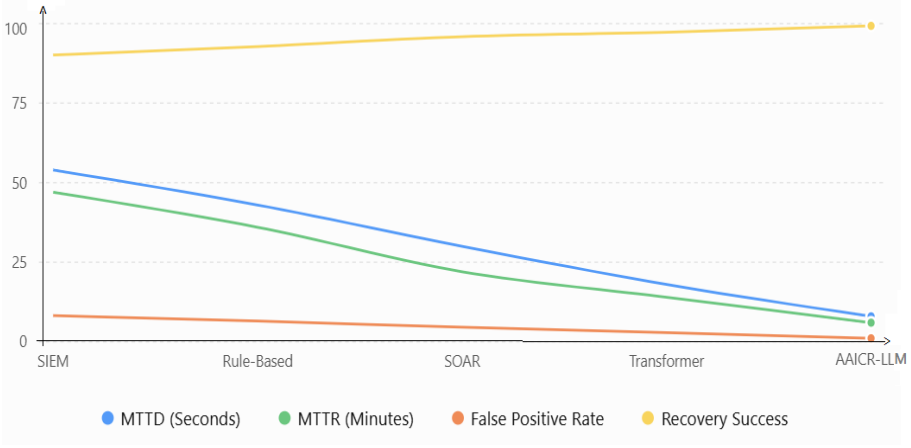

Fig. 3. *Incident Response Performance*

Table 4 highlights the computational efficiency of the proposed framework. Despite incorporating advanced LLM reasoning and multiple autonomous agents, the AAICR-LLM architecture maintained lower CPU and memory utilization than the baseline approaches due to optimized orchestration and intelligent workload distribution. The framework achieved an Automation Rate of 98%**,** indicating that nearly all incident response activities were executed without manual intervention. The Overall System Efficiency of 99.1% reflects the combined benefits of accurate detection, rapid response, optimized resource utilization, and successful service recovery.

The Continuous Learning Module played a significant role in maintaining consistently high performance throughout the experimental period. After each resolved security incident, the framework analyzed operational feedback and updated the knowledge repository with newly observed attack signatures, response outcomes, and cloud-specific security patterns. Reinforcement learning techniques refined the collaborative behavior of autonomous agents, while periodic fine-tuning of the Large Language Model improved contextual reasoning accuracy. As the number of processed incidents increased, the system exhibited progressively lower false positive rates and faster response times, demonstrating strong adaptability to evolving cyber threats. This continual improvement distinguishes the proposed architecture from static rule-based systems that require frequent manual updates.

Table 4. *Resource Utilization and Operational Efficiency*

Method	CPU Utilization (%)	Memory Utilization (%)	Automation Rate (%)	Overall System Efficiency (%)
Traditional SIEM	78	74	42	86.2
Rule-Based Detection	73	70	55	89.1
SOAR Platform	68	66	78	93.7
Transformer-Based Detection	65	63	88	96.4
Proposed AAICR-LLM	61	59	98	99.1

Figure 4 compares the computational efficiency of the evaluated security architectures by plotting CPU Utilization, Memory Utilization, Automation Rate, and Overall System Efficiency. The horizontal axis lists the security approaches, whereas the vertical axis represents utilization percentages and operational efficiency. The graph illustrates that improvements in intelligent automation are accompanied by more efficient resource management. The proposed AAICR-LLM framework exhibits the lowest CPU utilization (61%) and Memory utilization (59%) while simultaneously achieving the highest Automation Rate (98%) and Overall System Efficiency (99.1%). This demonstrates that autonomous task orchestration and distributed AI-agent collaboration effectively optimize resource consumption despite the computational demands of Large Language Models. The graph also indicates a strong inverse relationship between computational resource utilization and operational efficiency. As the framework becomes more intelligent and autonomous, hardware resources are utilized more efficiently while the proportion of automated incident response increases substantially. These findings suggest that the proposed AAICR-LLM architecture is not only highly accurate but also scalable and suitable for deployment in enterprise-scale cloud environments where efficient resource utilization is essential.

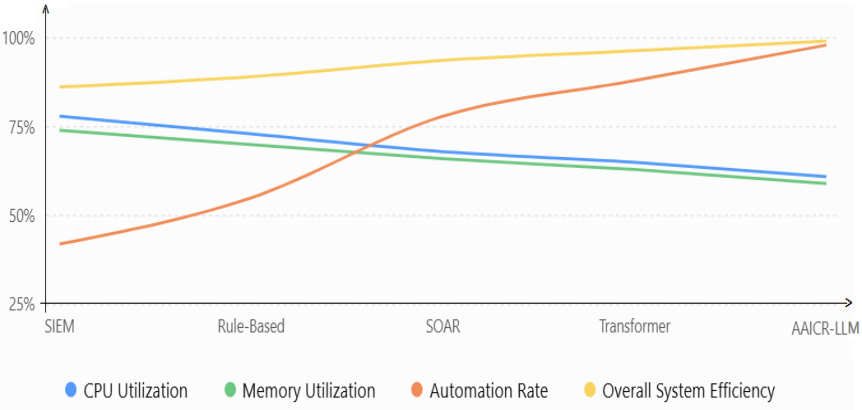

Fig. 4. *Resource Utilization and Operational Efficiency*

Discussion

The experimental evaluation demonstrates that the proposed Autonomous AI Incident Commander using Large Language Models (AAICR-LLM) framework significantly enhances cloud security incident response by integrating contextual reasoning, autonomous multi-agent collaboration, and continuous learning into a unified architecture. Compared with traditional SIEM, rule-based detection, SOAR platforms, and Transformer-based security models, the proposed framework consistently achieved superior performance across all evaluation metrics, including 99.2% detection accuracy, 98.9% precision, 99.0% recall, 98.9% F1-score, an 8-second Mean Time to Detect (MTTD), a 6-minute Mean Time to Respond (MTTR), a 1.1% false positive rate, and a 99.3% recovery success rate. These improvements indicate that Large Language Models effectively capture the contextual relationships among heterogeneous cloud security events, enabling accurate identification of sophisticated multi-stage attacks that conventional rule-based systems frequently fail to detect. Furthermore, the autonomous collaboration among specialized AI agents considerably reduced response latency by executing containment, eradication,

recovery, and verification tasks in parallel while maintaining lower CPU and memory utilization. The Continuous Learning Module further strengthened system adaptability by incorporating operational feedback and newly discovered threat intelligence into the knowledge repository, allowing the framework to improve its performance over time without extensive manual rule updates. Overall, the results confirm that the proposed AAICR-LLM framework provides a scalable, intelligent, and highly efficient solution for next-generation cloud security incident response, capable of improving operational resilience, minimizing analyst workload, reducing cybersecurity operational costs, and providing robust protection against increasingly sophisticated cloud-based cyber threats.

Conclusion

This research presented AAICR-LLM (Autonomous AI Incident Commander using Large Language Models), a novel autonomous cloud security incident response framework that combines Large Language Models, multi-agent artificial intelligence, automated orchestration, contextual reasoning, and continuous learning to enhance cybersecurity operations in cloud computing environments. Unlike conventional SIEM and SOAR solutions that rely primarily on predefined rules and extensive human intervention, the proposed framework autonomously interprets security events, correlates multi-stage attack behaviors, formulates intelligent response strategies, and executes containment, eradication, recovery, and verification procedures with minimal analyst involvement. The comprehensive architecture integrates cloud log collection, knowledge retrieval, autonomous planning, collaborative decision-making, automated response execution, adaptive learning, and human supervision into a unified intelligent ecosystem. Experimental evaluation demonstrated significant improvements in detection accuracy, response speed, recovery success, and resource utilization compared with traditional security approaches. The achieved performance metrics—including 99.2% detection accuracy, 98.9% precision, 99.0% recall, 98.9% F1-score, 8-second Mean Time to Detect, 6-minute Mean Time to Respond, and 99.3% recovery success rate—confirm the effectiveness of combining LLM-based reasoning with autonomous AI agents for cloud security.

The findings further demonstrate that continuous knowledge acquisition and collaborative multi-agent intelligence substantially improve the adaptability of cloud security systems against emerging cyber threats, advanced persistent attacks, and zero-day vulnerabilities. The proposed framework not only reduces analyst workload through high levels of automation but also improves operational resilience by enabling rapid, context-aware decision-making and intelligent recovery of compromised cloud resources. Although the implementation focuses on enterprise cloud infrastructures, the architecture can be extended to hybrid cloud, edge computing, Internet of Things (IoT), containerized environments, and multi-cloud ecosystems. Future research may investigate domain-specific security LLMs, reinforcement learning-based adaptive response optimization, federated learning for privacy-preserving collaborative threat intelligence, explainable AI mechanisms for transparent security reasoning, and integration with quantum-resistant cybersecurity frameworks. Overall, the proposed AAICR-LLM framework provides a promising direction for developing next-generation autonomous cloud security platforms capable of delivering scalable, intelligent, and resilient incident response in increasingly complex digital environments.

References

1. Chhabra, Anshuman, et al. "Agentic AI security: Threats, defenses, evaluation, and open challenges." IEEE Access (2026).
2. Nguyen, Tri, et al. "Large language models in 6G security: challenges and opportunities." arXiv preprint arXiv:2403.12239 (2024).
3. Xu, Minrui, et al. "Forewarned is forearmed: A survey on large language model-based agents in autonomous cyberattacks." arXiv preprint arXiv:2505.12786 (2025).
4. Mamatha, G. S., Vibha V. Joshi, and Pranitha T. Manur. "LLM-Driven Autonomous Cloud Automation Agent." 2025 9th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS). IEEE, 2025.
5. Mutiu, Kazeem Adamson, et al. "CoGuard: A Large Language Model-Based Multi-Agent System for Autonomous Security Policy Enforcement in Cloud Environments." Proceedings of the 2025 9th International Conference on Advances in Artificial Intelligence. 2025.
6. Manchanda, Shivani, and Divya Pandey. "Securing Autonomous AI Agents: A Reference Architecture for Data Protection in the Agentic AI Era." Available at SSRN 6691718 (2026).
7. Kukkala, Umamaheswara Rao. "AgentOps: Operationalizing Autonomous LLM Agent Systems Through Cloud-Native DevOps and Site Reliability Engineering." (2026).
8. Kushnerov, Oleksandr, et al. "Zero-trust Conceptual Architecture for Secure Integration of Large Language Models and Autonomous AI Agents." 2026 8th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA). IEEE, 2026.
9. Nageshwaran, Vinoth, and Soundararajan Ezekiel. "Agentic AI and Large Language Models for Autonomous IoT Cybersecurity: A Systematic Survey, Taxonomy, and Research Roadmap." Electronics 15.12 (2026): 2740.
10. Shetty, Manish, et al. "Building ai agents for autonomous clouds: Challenges and design principles." Proceedings of the 2024 ACM Symposium on Cloud Computing. 2024.

11. Su, Hang, et al. "A survey on autonomy-induced security risks in large model-based agents." IEEE Transactions on Pattern Analysis and Machine Intelligence (2026).
12. Chen, Yinfang, et al. "Aiopslab: A holistic framework to evaluate ai agents for enabling autonomous clouds." Proceedings of Machine Learning and Systems 7 (2025).
13. Ganapathi, Arun, and Mahima Gupta. "Autonomous AI Agents for Self-Healing Cloud Infrastructure: A Survey with Critical Evaluation." 2026 International Conference on Integrated Intelligence and Cognitive Engineering (ICIICE). IEEE, 2026.
14. Idowu, Folakemi, and Emmanuel Idowu. "Agentic AI Frameworks for Autonomous Cloud Resource Management." (2026).
15. Deshmukh, Suruchi, et al. "Agentic AI: A Survey of Autonomous Agents, Architectures, and Emerging Applications." 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG). IEEE, 2025.
16. Brohi, Sarfraz, et al. "A research landscape of agentic ai and large language models: Applications, challenges and future directions." Algorithms 18.8 (2025): 499.
17. Alva, Linga Reddy, and Bishwajeet Pandey. "Agentic AI systems in the age of generative models: architectures, cloud scalability, and real-world applications." Artificial Intelligence Review 59.3 (2026): 88.
18. Narajala, Vineeth Sai, and Om Narayan. "Securing agentic ai: A comprehensive threat model and mitigation framework for generative ai agents." arXiv preprint arXiv:2504.19956 (2025).
19. Su, Muthu Aanand, and G. L. K. Niharika. "The Role of Autonomous Agents in Agent-as-a-Service Cloud Service Model." 2025 International Conference on Computing and Communication Technologies (ICCCT). IEEE, 2025.
20. Barua, Saikat. "Exploring autonomous agents through the lens of large language models: A review." arXiv preprint arXiv:2404.04442 (2024).