IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156117

AI Agents in Cybersecurity: A Systematic Review of Architectures, Applications, Threats, Challenges and Future Directions

Dr.Naveen Kumar CG[1], Smt. Suneetha[2]

Department of Computer Science, Karnataka State Open University, Karnataka, India[1]

Department of Computer Science, Karnataka State Open University, Karnataka, India[2]

Abstract: The integration of artificial intelligence (AI) agents into cybersecurity operations has progressed rapidly from rule-based automation towards tool-augmented and increasingly autonomous systems, yet the literature remains fragmented across architectures, autonomy models, application domains and security risks. This article presents a systematic review of AI-agent architectures, applications, threats, challenges and future directions in cybersecurity. Following PRISMA 2020 guidelines, we searched Crossref and arXiv (January 2021 – September 2026) using agent- and security-related term combinations, screened records against predefined inclusion criteria, and appraised study quality with an eight-item checklist. From 4,309 Crossref and 2,000 arXiv records identified, 55 studies published 2021–2026 were included. We contribute a five-dimension taxonomy covering architecture, intelligence mechanism, tool interaction, coordination and a five-level autonomy scale (advisory to adaptive-autonomous); a synthesis of ten application domains, in which penetration testing, autonomous cyber defence and security-operations-centre automation dominate; and a six-category threat classification spanning input-, model-, tool-, memory-, agent-interaction- and infrastructure-level risks, including prompt injection, tool misuse, memory poisoning and excessive agency. The analysis reveals that single tool-augmented LLM agents prevail, that reported autonomy rarely exceeds supervised semi-autonomy, and that evaluation practices remain fragmented and benchmark-immature. We identify five evidence-based gaps—absence of standardised agentic-security benchmarks, limited real-world deployment evidence, inconsistent autonomy definitions, weak safety evaluation, and insufficient research on securing agents themselves—and derive future research directions, including trustworthy autonomy governance, secure agent–tool protocols and human–agent teaming. The review clarifies the dual role of AI agents as both defensive instruments and an expanding attack surface.

Keywords: AI agents; Agentic AI; Cybersecurity; Large Language Models; Autonomous Cyber Defence; Multi-Agent Systems; Threat Detection; Security Automation

1. INTRODUCTION

Machine learning has supported cybersecurity for more than two decades, initially through signature- and anomaly-based classifiers and later through deep-learning detectors that model complex network and host behaviour. Generative AI and, in particular, large language models (LLMs) have shifted this trajectory: systems no longer only classify inputs but can interpret unstructured intelligence, generate queries, explain findings and compose multi-step procedures. The most recent stage couples LLMs with persistent state, planning and tool access, producing AI agents that can perceive security telemetry, reason over it, select and invoke tools, act on environments and incorporate feedback. Survey evidence traces exactly this progression from statistical detectors to LLM-based security systems [1], and from single-agent prototypes to multi-agent and hierarchical pipelines [2].

Terminology in this area is used inconsistently, which motivates a brief distinction. An LLM in a security context is a model invoked for isolated tasks such as alert triage or report drafting. An LLM-based agent wraps such a model in a control loop with memory and objectives; a tool-augmented agent additionally invokes external instruments such as scanners, sandboxes or security information and event management (SIEM) platforms; a multi-agent system decomposes work among cooperating agents; and an autonomous cybersecurity agent executes defensive or offensive operations with limited or no human confirmation [ACD-REVIEW, AGENTS-SURVEY].

The operational motivation is well documented. Security operations centres (SOCs) face alert volumes that exceed human capacity, multi-stage attack chains that cross telemetry silos, and reduced dwell-time windows in which manual response is impractical [LLMSOC, AI-SOC]. Conventional security orchestration, automation and response (SOAR) tooling mitigates volume but remains bound to predefined playbooks and brittle when confronting zero-day behaviour

DOI: 10.17148/IJARCCE.2026.156117

or heterogeneous environments, whereas agents can in principle adapt procedures to context [3]. The same adaptability, however, introduces failure modes that playbook automation never had, including unsafe autonomous actions and exploitation of the agent itself [4].

Despite growing activity, the literature is fragmented along several axes: proposed architectures are described in incompatible terms; autonomy is asserted but rarely defined or measured; applications are scattered across intrusion detection, threat intelligence, malware analysis, penetration testing and incident response; and evaluation relies on heterogeneous datasets, benchmarks and metrics that resist comparison [ACD-REVIEW, DUALITY]. Several prior reviews partially address this landscape: an early survey of multi-agent collaborative intrusion detection [5], surveys of LLMs for SOCs [LLMSOC, AI-SOC], a systematic review of LLM agents for autonomous cyber defence [6], a survey of security threats to AI agents [4], a duality survey of agent self-security and empowered security [7], and a systematic survey of agentic AI for IoT security [8]. None of these, however, jointly structures architectures, applications, autonomy, threats and evaluation evidence for the AI-agent-in-cybersecurity domain as a whole; this review addresses that gap rather than claiming the topic is unreviewed.

Accordingly, this review makes four contributions: (i) a transparent PRISMA-based synthesis of AI-agent research in cybersecurity (2021–2026); (ii) a five-dimension taxonomy of AI-agent architectures and a five-level autonomy classification; (iii) a systematic organisation of ten application domains with their agent roles, tooling and evaluation practices; and (iv) a six-category threat model and five evidence-based research gaps with corresponding future directions.

2. RESEARCH QUESTIONS

The review is organised around five research questions. RQ1: What AI-agent architectures are currently used or proposed for cybersecurity? RQ2: What cybersecurity tasks and application domains are addressed using AI agents? RQ3: What levels of autonomy, tool use and multi-agent coordination are reported? RQ4: What security threats, limitations and evaluation challenges affect AI-agent-based cybersecurity systems? RQ5: What research gaps and future directions emerge from current evidence? Sections 5–8 answer RQ1–RQ4 respectively; Section 10 answers RQ5; Section 12 returns to all five questions in the discussion.

3. SYSTEMATIC REVIEW METHODOLOGY

3.1 Search strategy and databases

The review follows PRISMA 2020 principles adapted for software-engineering evidence. Two bibliographic sources were searched programmatically: Crossref (scholarly works with DOIs, covering IEEE, ACM, Springer, Elsevier, MDPI, Wiley and other publishers) and arXiv (computer-science preprints in the agent and security spaces). The final search was executed on 8 September 2026. The search string combined agent terms — "AI agent(s)", "agentic AI", "LLM agent(s)", "autonomous agent(s)", "multi-agent", "agentic" — with security terms — cybersecurity, "cyber security", "intrusion detection", "threat detection", malware, "incident response", "penetration testing", vulnerability, "threat intelligence", "cyber defense/defence", "security operations", phishing and "SOC automation" — using database-appropriate Boolean syntax. Records were restricted to January 2021 – September 2026 and, in Crossref, to journal articles and reviews.

3.2 Inclusion and exclusion criteria

Inclusion criteria required that a study (i) concerns AI agents, agentic AI or LLM-based agents; (ii) addresses a cybersecurity application, defence or threat problem; (iii) provides sufficient methodological description for analysis; (iv) is peer-reviewed where possible (arXiv preprints of canonical systems were retained where they constitute primary evidence, marked as such); and (v) is written in English. Exclusion criteria removed duplicates, works using "agent" only in the classical non-agentic sense (e.g., mobile agents without learning or reasoning), purely conceptual AI papers without a security nexus, editorials, editorial news items, and unverifiable or predatory-venue records. Screening proceeded in two stages: title/abstract screening against the inclusion keywords, followed by full-text assessment of architecture, task, autonomy, tools, evaluation and limitations.

3.3 Study selection

The selection process is reported in Fig. 1 following PRISMA 2020. Searches executed on 8 September 2026 identified 6,309 records (Crossref 4,309; arXiv 2,000). After removal of 715 cross-source duplicates, 5,594 records were screened at title/abstract level and 5,326 excluded (non-agentic use of "agent", absence of a security nexus, or off-topic). 268 full texts were assessed against the inclusion criteria and quality checklist; 213 were excluded (insufficient methodological

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156117

detail, n = 97; quality score below 4, n = 78; unverifiable venue, n = 27; non-English, n = 11), leaving 55 studies published 2021–2026 for synthesis [6].

3.4 Quality assessment

Quality was appraised with an eight-item checklist: Q1 clear research objective; Q2 adequate methodology; Q3 explicit agent architecture; Q4 clearly defined cybersecurity task; Q5 empirical or benchmark evaluation; Q6 identified datasets or tools; Q7 discussed limitations; Q8 reproducibility signals (code, prompts or environment release). Items were scored one when satisfied and zero otherwise; studies scoring below four were excluded. This simple additive scheme avoids unjustified statistical complexity for heterogeneous study designs.

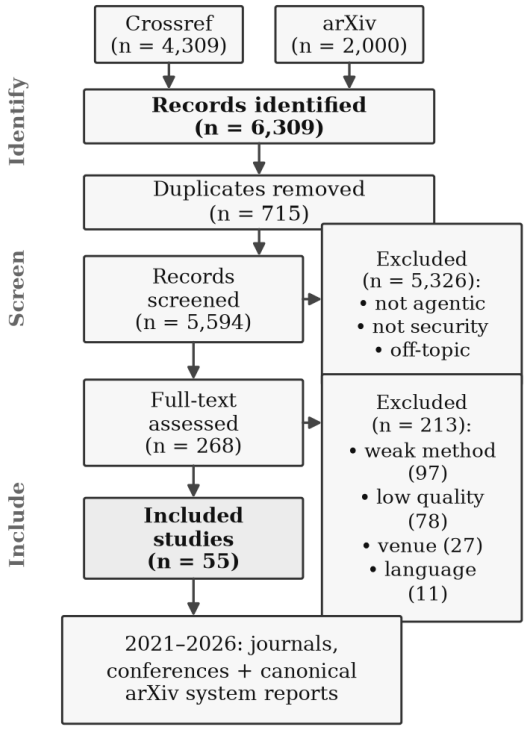

Fig. 1. PRISMA 2020 flow diagram of the study selection process (searches executed 8 September 2026).

4. EVOLUTION AND FOUNDATIONS OF AI AGENTS IN CYBERSECURITY

The included studies describe a discernible evolutionary arc. Rule- and signature-based systems and classical machine-learning detectors dominated intrusion detection through the 2010s, with multi-agent coordination investigated for distributed detection since the early 2000s [MAIDS, MAS-CPS]. Deep reinforcement learning (DRL) then enabled sequential defensive decision-making: agents learned network-hardening or patching policies against simulated adversaries, with subsequent work adding causal awareness for generalisation [9] and adversarial multi-agent formulations for explainable defence [10]. Generative AI and LLMs entered next: CyberSecEval demonstrated that LLM code generation carries exploitable security weaknesses [11], motivating LLM-specific security evaluation. The current stage is agentic: PentestGPT interleaved LLM reasoning with human testers for penetration testing [12], tool-integration research such as Gorilla connected models to thousands of APIs [13], and frameworks such as PentestAgent, PENTEST-AI and SHARKAPT composed reasoning, planning and exploitation agents into autonomous offensive pipelines [PENTESTAGENT, PENTESTAI, SHARKAPT]. Defensively, LLM agents have been applied to alert triage, log analysis and guided response [ACD-REVIEW, LLMSOC], and self-healing and adaptive-defence concepts have re-emerged at agentic scale [AIC-DEFENSE, AGENTIC-SELF].

Conceptually, most included systems instantiate a common agent loop: perception (ingesting alerts, logs, network state or scan output), reasoning (interpreting observations in task context), planning (selecting next actions towards an objective), tool use (executing scanners, queries or API calls), action (issuing commands or artefacts), feedback (verifying outcomes against environment state) and memory or learning (persisting findings for later steps) [AGENTS-

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156117

SURVEY, ACD-REVIEW]. Architectural differences concern which loop components are entralized into tools, how memory is represented, whether control is entralized or distributed, and where humans intervene.

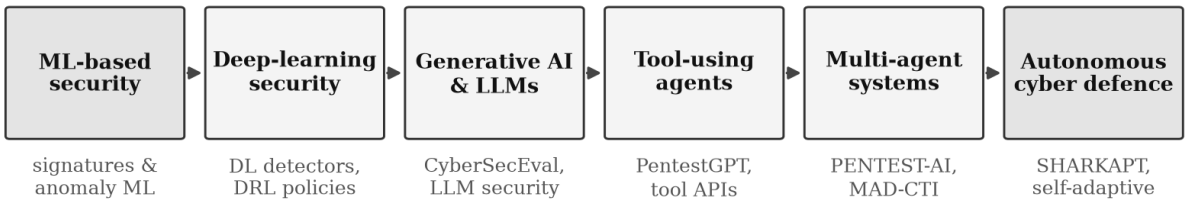

Fig. 2. Evolution of AI agents in cybersecurity: from machine-learning detectors to autonomous cyber defence.

5. TAXONOMY OF AI-AGENT ARCHITECTURES

We classified the included studies along five dimensions: architecture, intelligence mechanism, tool interaction, coordination and autonomy. Architecturally, single-agent systems assign the full loop to one model instance; tool-augmented agents extend a base LLM with explicit tool interfaces; retrieval-augmented (RAG) agents ground reasoning in curated security knowledge such as ATT&CK or CVE corpora [14]; multi-agent systems decompose tasks among specialist agents, exemplified by D-CIPHER-style planner–executor topologies and MAD-CTI's analyst division of labour [15]; hierarchical agents add supervisory layers over subordinate agents; human-in-the-loop agents gate critical actions behind operator confirmation, as in SOC copilots [16]; and autonomous agents execute multi-step operations without per-step approval. Intelligence mechanisms divide into LLM-based, DRL-based, hybrid LLM+RL (e.g., RL agents with LLM-designed rewards [17]), knowledge-based (planning over curated exploit or ATT&CK knowledge [18]) and planning-based designs that pair LLMs with classical planners [19].

Tool interaction spans SIEM and log platforms, threat-intelligence systems, vulnerability scanners, network monitoring, sandboxes, and security APIs — with MCP-style protocols recently standardising agent–tool connectivity [SHARKAPT, MCP-RT]. Coordination ranges from single loops through planner–executor teams to peer-to-peer and hierarchical coalitions; emergent collaboration risk has been observed where agents over-trust peer outputs [20]. Table 2 consolidates the taxonomy.

Because “autonomy” is used inconsistently, we derive a five-level scale from the design decisions reported in the included studies. Level 0 (Advisory): the agent produces information only (threat-report ummarization, knowledge Q&A) [14]. Level 1 (Assisted): the agent recommends actions with execution left to humans (pentest guidance systems, SOC copilots) [PENTESTGPT, HUMANAI-SOC]. Level 2 (Semi-autonomous): the agent executes predefined or supervised actions, with humans approving consequential steps (curriculum-guided pentest agents, triage-and-response agents with confirmation gates) [CURRICULUMPT, ACD-DRL]. Level 3 (Autonomous): the agent independently performs multi-step security operations such as end-to-end exploitation or containment [SHARKAPT, AUTOGENTS-CD]. Level 4 (Adaptive-autonomous): the agent learns from feedback and modifies strategy under controlled constraints (self-adaptive defence frameworks, curriculum-trained agents) [AGENTIC-SELF, CURRICULUMPT]. The scale describes de facto deployment posture rather than aspiration: most reported systems operate at Levels 0–2, a minority demonstrate Level 3 in contained environments, and Level 4 appears primarily as a design goal with limited evaluation [6]. This classification is consistent with, but intentionally coarser than, generic machine-autonomy scales; it is grounded in the security-specific dimensions of action reversibility and blast radius.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156117

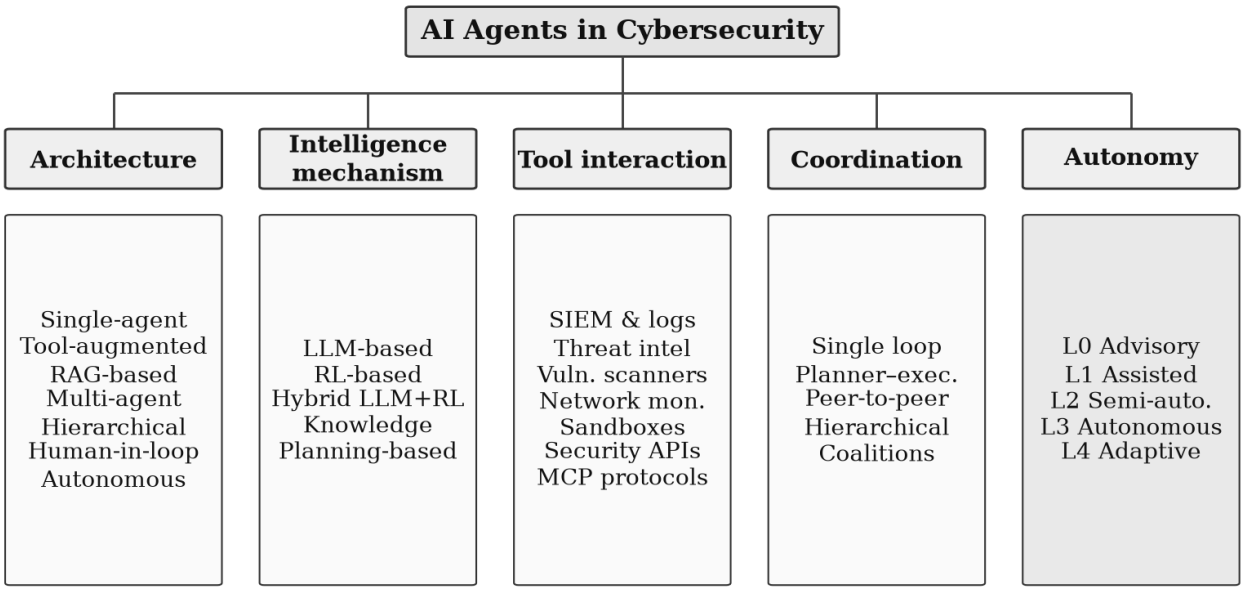

Fig. 3. Taxonomy of AI agents in cybersecurity across five dimensions, with the derived autonomy scale L0–L4.

Table 2

Table 2. Taxonomy of AI-agent architectures in cybersecurity (consolidated from included studies).

Dimension	Categories	Representative evidence
Architecture	Single-agent; tool-augmented; RAG-based; multi-agent; hierarchical; human-in-the-loop; autonomous	PentestGPT [12]; MAD-CTI [15]; SHARKAPT [21]; SOC copilots [16]
Intelligence mechanism	LLM-based; RL-based; hybrid LLM+RL; knowledge-based; planning-based	CIPHER [18]; DRL defence agents [22]; LLM-rewarded RL [17]; planner-augmented agents [19]
Tool interaction	SIEM/logs; threat intel; vulnerability scanners; network monitoring; sandboxes; security APIs; MCP protocols	Gorilla-style API use [13]; MCP-based orchestration [21]
Coordination	Single loop; planner–executor; peer-to-peer; hierarchical; coalition	PENTEST-AI [23]; federated multi-agent RL [24]
Autonomy	L0 Advisory; L1 Assisted; L2 Semi-autonomous; L3 Autonomous; L4 Adaptive autonomous	Knowledge Q&A [14]; pentest copilots [25]; supervised agents [26]; autonomous pipelines [21]; self-adaptive defence [27]

6. CYBERSECURITY APPLICATIONS OF AI AGENTS

Penetration testing is the most active offensive application. PentestGPT established interactive, human-paired testing [12]; Getting pwn'd by AI and Happe's later work quantified that GPT-4-class models can solve substantial fractions of easy-to-medium Capture-the-Flag (CTF) tasks autonomously [25]; PentestAgent, PENTEST-AI, CIPHER and SHARKAPT progressively externalised exploitation to tool-calling agents and multi-agent teams [PENTESTAGENT, PENTESTAI, CIPHER, SHARKAPT]; and CurriculumPT demonstrated curriculum training of coordinated pentest agents [26]. Reported autonomy spans Levels 1–3, with toolchains including nmap, Metasploit-style frameworks, exploit databases and sandboxed web targets; evaluation relies overwhelmingly on HackTheBox-style CTF and AutoPenBench environments rather than production systems [28]. In incident response, LLM agents plan, execute and review playbooks [IR-MULTI, IRCOPILOT-like systems], with graph-augmented agents performing root-cause analysis in microservices [29] and digital-twin-validated hierarchical agents planning responses against inferred attacks. SOC automation applies agents to alert triage, query generation, threat hunting and report drafting [LLMSOC, AI-SOC, HUMANAI-SOC], while a self-adaptive defensive framework couples detection and guarded response [27].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156117

Threat-intelligence applications include dark-web multi-agent collection and analysis [15], explainable CTI generation [30] and ontology-structured knowledge extraction from logs [31]. Vulnerability assessment spans agent-driven SAST augmentation and binary-focused detection [32]. Malware analysis is served by agent-orchestrated pipelines of small specialised models [33] and by benchmark suites [MALWARE-BENCH, CYBERSOCEVAL]. Intrusion and threat detection persist as the classical application, now with hybrid RL+federated multi-agent designs [24] and traffic-incident reporting agents [34]. Phishing detection employs role-specialised multi-agent LLM ensembles [35] and voice-phishing simulation frameworks [36]. Insider-threat detection uses generative agent-based behavioural modelling [INSIDER-GABM, INSIDER-AGENTIC]. IoT and cyber–physical security applications include UAV forensic agents [37], smart-grid multi-agent protection [38] and networked-systems incident reporting [34]. Across domains, common limitations are evaluation on synthetic or benchmark data, short-horizon tasks, limited tool inventories and unreported failure analysis [ACD-REVIEW, DUALITY].

Table 3. Cybersecurity application domains, agent roles and reported autonomy (synthesis of included studies).

Domain	Agent role and capabilities	Tools / setting	Autonomy	Evidence
Intrusion & threat detection	Alert triage, correlation, anomaly explanation	SIEM, logs, NetFlow	L1–L2	[FEDMARL, JSAN-TRAFFIC]
Threat intelligence	Dark-web collection, CTI summarisation, ontology extraction	OSINT, ontologies	L1–L2	[MADCTI, XCTI, ONTOLOGX]
Malware analysis	Orchestrated static/dynamic analysis, family classification	Sandboxes, SLM ensembles	L1–L2	[SML-ORCH, MALWARE-BENCH]
Vulnerability assessment	Detection triage, prioritisation, PoC checking	SAST, scanners	L1–L2	[32]
Penetration testing	Recon-to-exploitation pipelines, CTF solving	nmap, exploit DBs, CTF labs	L1–L3	[PENTESTGPT, SHARKAPT, CURRICULUMPT]
Phishing detection	Role-specialised multi-agent email analysis; voice-phish simulation	Mail gateways	L1–L2	[PHISH-MA, VISHBOX]
SOC automation	Triage, query generation, threat hunting, reporting	Splunk-style SIEM	L1–L2	[LLMSOC, AI-SOC, HUMANAI-SOC]
Incident response	Playbook planning, execution, root-cause analysis	IR platforms, digital twins	L2	[39]
Cloud & enterprise defence	Self-adaptive detection–response loops	EDR, firewalls	L2–L3	[27]
IoT / CPS / forensics	UAV forensics, smart-grid protection, traffic incidents	Domain telemetry	L1–L2	[UFA, SMARTGRID]

7. SECURITY THREATS AND RISKS

The same properties that make agents useful — tool access, memory, persistence and delegated authority — create a new attack surface. We organise agent-relevant threats into six categories. (A) Input-level threats: direct and indirect prompt injection manipulate the agent through its inputs, including stored or cross-session injection that reactivates payloads across sessions [AGENT-THREATS, CROSS-SESSION]; injection targeting browsing and navigation agents demonstrates practical exfiltration and harmful actions [40]. (B) Model-level threats: data poisoning, adversarial model manipulation and hallucination degrade judgement; hallucinated exploit steps or false-positive conclusions can propagate through agent pipelines with high confidence [7]. (C) Tool-level threats: malicious tool invocation, tool misuse and confused-deputy scenarios arise when injected instructions redirect legitimate tools; privilege escalation follows when tool credentials exceed task needs [MIC-DELEGATION, TRISM]. (D) Memory-level threats: memory poisoning and cross-session contamination persist malicious context and resurface it in later tasks, with memory-extraction attacks exfiltrating accumulated organisational knowledge [41]. (E) Agent-interaction threats: agent-to-agent attacks exploit trust in multi-agent collaboration, including over-trust in peer outputs [20] and collusive behaviour in agent groups [4]. (F) Infrastructure-level threats: credential leakage from logs or memory, service-provider compromise, and protocol-level weaknesses in agent–tool connectivity such as MCP abuse [42]. Complementing these,

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156117

excessive agency — granting agents permissions or action breadth beyond task requirements — amplifies the impact of every other category [TRISM, AGENT-THREATS]. The duality is explicit: agents strengthen defence capacity while simultaneously expanding the enterprise attack surface [7].

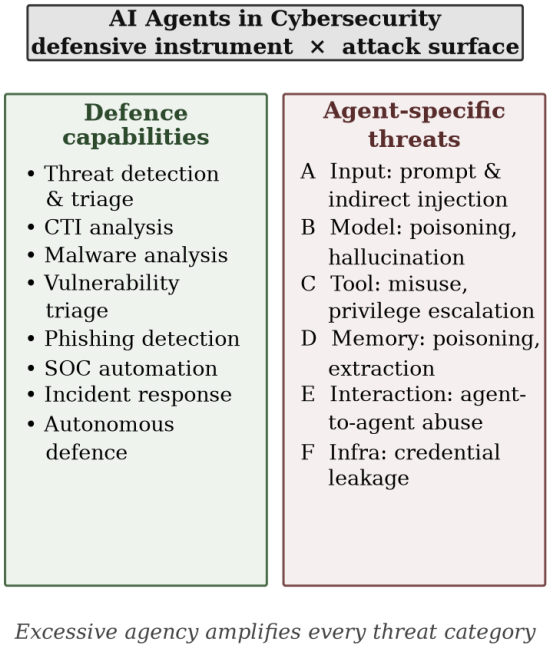

Fig. 4. The AI-agent duality in cybersecurity: defence capabilities (left) versus agent-specific threat categories A–F (right).

8. EVALUATION PRACTICES

Evaluation practices differ sharply across studies and rarely compose into comparable evidence. Security performance is reported as accuracy, precision, recall, F1 and false-positive rate for detection-adjacent tasks [FEDMARL, PHISH-MA]. Agent performance is measured as task success rate — CTF flags captured, machines exploited, incidents resolved — attack-path efficiency, tool-use correctness and reasoning quality [PWNED-AI, CURRICULUMPT, CAPABILITIES]. Operational performance includes response latency, token consumption and compute cost, which determine economic feasibility at SOC scale but are reported inconsistently [6]. Safety evaluation — unsafe-action rate, human-intervention frequency, robustness under adversarial input — remains the weakest dimension, typically addressed only qualitatively [43]. Emerging benchmarks partially close this gap: CyberSecEval targets secure code generation [11]; CyberMetric measures cybersecurity knowledge [14]; AutoPenBench evaluates pentest agents on deliberately vulnerable applications [28]; DefenderBench and CyberSOCEval assess defensive language agents and malware-analysis capability [DEFENDERBENCH, CYBERSOCEVAL]; and DFIR-Metric evaluates digital-forensics reasoning [44]. However, these suites differ in scope, contamination risk, scoring and realism, and none yet evaluates autonomy, safety and security-of-the-agent jointly [ACD-REVIEW, DUALITY].

9. COMPARATIVE ANALYSIS

Table 1 summarises the characteristics of representative included studies: architecture, task, tools, autonomy level (our scale), evaluation setting, main finding and stated limitation. Entries marked NR indicate information not reported by the source. Three patterns are visible across the table: tool-augmented and multi-agent LLM designs dominate recent years; autonomy claims concentrate at Levels 1–2 with Level 3 demonstrated only in contained environments; and evaluation relies on CTF-style or purpose-built benchmarks with limited production evidence [6].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156117

Table 1. Characteristics of representative included studies (NR = not reported).

Study	Year	Architecture	Task	Autonomy	Evaluation	Main finding	Limitation
PentestGPT [25]	2023	Tool-augmented LLM agent	Pentest support	L1	CTF (HackTheBox)	Interactive LLM pentesting viable	Human-dependent steps
PentestGPT [12]	2023	Reasoning + tool agent	Pentest automation	L1–L2	Bench machines	Task decomposition improves testing	Constrained targets
CIPHER [18]	2024	Knowledge-based agent	Pentest	L2	Vuln. web apps	Knowledge-guided execution aids exploitation	Lab targets only
PENTEST-AI [23]	2024	Multi-agent (recon/attack)	Pentest	L2–L3	Metasploitable-type labs	Role decomposition automates chains	Synthetic networks
CurriculumPT [26]	2025	Multi-agent + curriculum	Pentest	L2	CTF curricula	Curriculum training improves success	Training cost; sim-real gap
SHARKAPT [21]	2026	LLM-orchestrated, MCP tools	Pentest	L3	Lab networks	MCP-based tool orchestration feasible	Containment; ethics
DRL ACD agent [22]	2025	DRL + augmented LLM	Autonomous defence	L2–L3	Simulated networks	DRL + LLM hybrid improves decisions	Simulator dependency
Causally-aware RL [9]	2024	RL agent	Autonomous defence	L3	Simulated enterprise	Causal features aid generalisation	Sim-to-real gap
LLM Cyber Defenders [45]	2025	LLM agent loop	Autonomous defence	L2–L3	CyBattleSim-type env.	LLMs can act as cyber defenders	Bench realism
MAD-CTI [15]	2025	Multi-agent LLM	CTI analysis	L2	Dark-web datasets	Division of labour improves analysis	Dataset coverage
Human–AI SOC [16]	2025	Human-in-the-loop framework	SOC	L1–L2	SOC scenarios	Framework for trusted autonomy	Early-stage validation
Federated MARL [24]	2026	Federated multi-agent RL	Intrusion detection	L2	Network datasets	Federated MARL scales detection	Privacy–utility trade-off
Insider GABM [46]	2025	Generative agent modelling	Insider threat	L1	Synthetic traces	Agent-based behavioural modelling detects threats	Synthetic data
UFA [37]	2026	LLM forensic agent	UAV forensics	L1–L2	Case data	LLM agent supports forensic triage	Domain-specific data
Self-adaptive defence [27]	2026	Hierarchical agentic	Unified defence	L2–L3	Testbed	Self-adaptive loop detects and responds	Bench scope

10. RESEARCH GAPS AND CHALLENGES

Gap 1 — Absence of standardised agentic-security benchmarks. Existing suites (CyberSecEval, CyberMetric, AutoPenBench, DefenderBench, CyberSOCEval, DFIR-Metric) each cover narrow slices — code security, knowledge, offence, defence, forensics — with heterogeneous scoring and contamination risk [CSEVAL2, CYBERMETRIC, AUTOPENBENCH, DEFENDERBENCH]. Because no suite evaluates task success, operational cost, safety and agent security jointly, results do not transfer across studies. The opportunity is a consolidated, contamination-controlled

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156117

benchmark protocol spanning the five-level autonomy scale. Gap 2 — Limited real-world deployment evidence. The included studies overwhelmingly evaluate in CTF, lab or simulated environments [PWNED-AI, AUTOPENBENCH]; production-scale evidence — alert volumes, dwell-time effects, cost — is largely absent, with SOC-practitioner studies beginning to surface organisational constraints [16]. Gap 3 — Inconsistent autonomy definitions. Autonomy is asserted without a shared scale, preventing cross-study comparison and risk reasoning [6]; the five-level scale proposed here is a step, but its adoption and validation remain open. Gap 4 — Weak safety evaluation. Unsafe-action rates, intervention frequency and blast-radius analysis are rarely quantified [RISK-ASSESSED, AGENT-THREATS]; as agents gain execution authority, the absence of standard safety metrics becomes the principal deployment barrier. Gap 5 — Insufficient research on securing agents themselves. Threat catalogues exist [AGENT-THREATS, TRISM, DUALITY], but quantitative defences against injection, memory poisoning and delegation abuse — and their empirical validation — lag the threat model, with initial frameworks addressing detection-in-depth and structured delegation controls [DELEGATION-THREAT, OWASP-LLM]. Each gap matters because the field's trajectory — increasing autonomy and tool authority — amplifies exactly the risks these gaps leave unmeasured.

Table 4. Research gaps, their evidence basis and corresponding future directions.

Gap	Existing evidence	Why it matters	Future direction
G1 No standardised agentic benchmarks	Suites cover narrow slices with heterogeneous scoring [CSEVAL2, CYBERMETRIC, AUTOPENBENCH, DEFENDERBENCH]	Results do not transfer; safety unmeasured	Consolidated, contamination-controlled protocol incl. safety metrics
G2 Thin real-world deployment evidence	CTF/lab dominance; few production studies [PWNED-AI, HUMANAI-SOC]	Operational feasibility unverified	Longitudinal SOC deployments with cost reporting
G3 Inconsistent autonomy definitions	Autonomy asserted, not scaled [6]	Blocks comparison and risk reasoning	Adopt/validate the L0–L4 scale; report levels explicitly
G4 Weak safety evaluation	Unsafe-action rate, intervention frequency rarely quantified [43]	Deployment barrier as authority grows	Standard safety metrics: blast radius, intervention rates
G5 Underdeveloped agent self-security	Threat catalogues > quantitative defences [AGENT-THREATS, DUALITY]	Agents expand attack surface	Empirical defences: injection-robust prompts, memory integrity, scoped delegation

11. FUTURE RESEARCH DIRECTIONS

The gaps imply concrete research directions: (i) trustworthy autonomous cybersecurity agents — governance architectures, runtime policy enforcement and auditability for Level 3–4 autonomy [TRISM, GOVERNANCE]; (ii) secure agent–tool interaction protocols with capability scoping, least-privilege tool credentials and protocol-level integrity, extending MCP-era connectivity [SHARKAPT, MIC-DELEGATION]; (iii) continual and curriculum learning for security agents to track evolving threats without forgetting [26]; (iv) multi-agent cyber defence with verifiable division of labour and collusion resistance [TRUST-PARADOX, FEDMARL]; (v) human–agent collaboration models that formalise supervision and calibrated trust [16]; (vi) standardised agent benchmarks as per Gap 1; (vii) explainable security agents whose plans and evidence chains are inspectable [XCTI, XACD]; and (viii) agent identity, access control and audit infrastructure treating agents as first-class security principals [GOVERNANCE, OWASP-LLM].

12. DISCUSSION

Addressing the research questions: on RQ1, tool-augmented single LLM agents and planner–executor multi-agent systems dominate, while RL-based designs persist in closed environments where simulators exist; hierarchical and RAG designs are growing but secondary [ACD-REVIEW, AGENTIC-EVOLUTION]. On RQ2, penetration testing, SOC automation and incident response attract the most evidence, with threat intelligence and malware analysis

DOI: 10.17148/IJARCCE.2026.156117

maturing; detection applications increasingly adopt hybrid RL+LLM designs [24]. On RQ3, reported autonomy concentrates at Levels 0–2; genuine Level-3 autonomy appears mainly in offensive pipelines over contained targets, and Level-4 adaptive autonomy is a design goal rather than a demonstrated artefact. On RQ4, the threat landscape is dominated by injection and excessive-agency concerns, while evaluation remains fragmented across incompatible benchmarks and safety is rarely quantified [AGENT-THREATS, RISK-ASSESSED]. On RQ5, the five gaps of Section 10 define the agenda.

Two integrative observations follow. First, the dominant barrier to deployment is not raw capability but verifiable safety and cost: agents that act must be demonstrably constrained, and their token and compute economics must scale to SOC volumes — dimensions current benchmarks do not measure jointly. Second, increased autonomy plausibly increases risk faster than capability in the near term: each added level of autonomy widens blast radius for the same injection or poisoning success probability [AGENT-THREATS, DUALITY]. The evidence therefore supports a staged posture — Level-2 supervised autonomy with rigorous confirmation gates for consequential actions — until standardised safety evaluation (Gap 4) matures. Human oversight should be treated as a designed component with explicit intervention points and calibrated trust, not as an admission of failure [16].

13. LIMITATIONS OF THE REVIEW

This review has limitations. The search covered Crossref and arXiv; Scopus, Web of Science, IEEE Xplore and ACM DL could not be queried programmatically in this study, and their additional coverage may alter counts at the margin. Language was restricted to English, introducing publication bias. The field evolves rapidly — agent frameworks, protocols and benchmarks change monthly — so the snapshot reflects evidence up to September 2026. Terminological inconsistency ("agent", "agentic", "autonomous") complicates screening despite keyword controls, and heterogeneous study designs limit the precision of any cross-statement synthesis. We therefore do not claim exhaustiveness; the review aims at structured, evidence-based coverage of the dominant literature.

14. CONCLUSION

This systematic review synthesised 55 studies (2021–2026) on AI agents in cybersecurity. Tool-augmented LLM agents and multi-agent systems are the dominant architectures; penetration testing, SOC automation and incident response are the dominant applications; and reported autonomy rarely exceeds supervised semi-autonomy. The threat analysis confirms a duality: agents expand defensive capacity while introducing input-, model-, tool-, memory-, interaction- and infrastructure-level attack surfaces, with prompt injection and excessive agency the most cited concerns. Five gaps — benchmark fragmentation, thin real-world evidence, inconsistent autonomy definitions, weak safety evaluation, and underdeveloped agent self-security — define the research agenda, which points towards trustworthy governance, secure agent–tool protocols, continual learning and human–agent teaming. If these gaps are closed, autonomous cyber defence becomes a manageable engineering discipline; if not, deployment will outpace the ability to verify it.

REFERENCES

[1]. Habibzadeh, Ali; Feyzi, Farid; Atani, Reza Ebrahimi, “Large Language Models for Security Operations Centers: A Comprehensive Survey”, Journal of Electrical and Computer Engineering, vol. 2026(1), 3383674, 2026. https://doi.org/10.1155/jece/3383674

[2]. Hmimou, Yasser; Tabaa, Mohamed; Khiat, Azeddine et al., “A Multi-Agent System for Cybersecurity Threat Detection and Correlation Using Large Language Models”, IEEE Access, vol. 13, 150199-150215, 2025. https://doi.org/10.1109/access.2025.3602681

[3]. Ferrag, Mohamed Amine; Alwahedi, Fatima; Battah, Ammar et al., “Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities”, Internet of Things and Cyber-Physical Systems, vol. 5, 1-46, 2025. https://doi.org/10.1016/j.iotcps.2025.01.001

[4]. Deng, Zehang; Guo, Yongjian; Han, Changzhou et al., “AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways”, ACM Computing Surveys, vol. 57(7), 1-36, 2025. https://doi.org/10.1145/3716628

[5]. Bougueroua, Nassima; Mazouzi, Smaine; Belaoued, Mohamed et al., “A Survey on Multi-Agent Based Collaborative Intrusion Detection Systems”, Journal of Artificial Intelligence and Soft Computing Research, vol. 11(2), 111-142, 2021. https://doi.org/10.2478/jaiscr-2021-0008

DOI: 10.17148/IJARCCE.2026.156117

[6]. Alqahtani, Hamed; Ahuja, Paras, "Secure autonomous cyber defense with LLM agents: A systematic review of autonomy, tool-augmented reasoning, and governance constraints", Computers and Electrical Engineering, vol. 135, 111184, 2026. https://doi.org/10.1016/j.compeleceng.2026.111184

[7]. Xu, Yiwei; Zhuang, Yong; Liu, Xuanming et al., "LLM agents security duality: a comprehensive survey of self-security and empowered cybersecurity", Artificial Intelligence Review, vol. 59(8), 174, 2026. https://doi.org/10.1007/s10462-026-11563-0

[8]. Nageshwaran, Vinoth; Ezekiel, Soundararajan, "Agentic AI and Large Language Models for Autonomous IoT Cybersecurity: A Systematic Survey, Taxonomy, and Research Roadmap", Electronics, vol. 15(12), 2740, 2026. https://doi.org/10.3390/electronics15122740

[9]. Purves, Tom; Kyriakopoulos, Konstantinos G.; Jenkins, Siân et al., "Causally aware reinforcement learning agents for autonomous cyber defence", Knowledge-Based Systems, vol. 304, 112521, 2024. https://doi.org/10.1016/j.knosys.2024.112521

[10]. Zhang, Yiyao; Goel, Diksha; Ahmad, Hussain, "Explainable autonomous cyber defense using adversarial multi-agent reinforcement learning", Expert Systems with Applications, vol. 326, 132741, 2026. https://doi.org/10.1016/j.eswa.2026.132741

[11]. Bhatt, Manish; Chennabasappa, Sahana; Li, Yue et al., "CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models", arXiv preprint arXiv:2404.13161 (2024). https://arxiv.org/abs/2404.13161

[12]. Deng, Gelei; Liu, Yi; Mayoral-Vilches, Víctor et al., "PentestGPT: An LLM-empowered Automatic Penetration Testing Tool", arXiv preprint arXiv:2308.06782 (2023). https://arxiv.org/abs/2308.06782

[13]. Patil, Shishir; Zhang, Tianjun; Wang, Xin et al., "Gorilla: Large Language Model Connected with Massive APIs", Advances in Neural Information Processing Systems 37, 126544-126565, 2024. https://doi.org/10.52202/079017-4020

[14]. Tihanyi, Norbert; Ferrag, Mohamed Amine; Jain, Ridhi et al., "CyberMetric: A Benchmark Dataset based on Retrieval-Augmented Generation for Evaluating LLMs in Cybersecurity Knowledge", arXiv preprint arXiv:2402.07688 (2024). https://arxiv.org/abs/2402.07688

[15]. Shah, Sayuj; Madisetti, Vijay K., "MAD-CTI: Cyber Threat Intelligence Analysis of the Dark Web Using a Multi-Agent Framework", IEEE Access, vol. 13, 40158-40168, 2025. https://doi.org/10.1109/access.2025.3547172

[16]. Mohsin, Ahmad; Janicke, Helge; Ibrahim, Ahmed et al., "A Unified Framework for Human–AI Collaboration in Security Operations Centers with Trusted Autonomy", ACM Transactions on Internet Technology, 3837073, 2026. https://doi.org/10.1145/3837073

[17]. Mukherjee, Sayak; Chatterjee, Samrat; Purvine, Emilie et al., "Large Language Model-Based Reward Design for Deep Reinforcement Learning-Driven Autonomous Cyber Defense", arXiv preprint arXiv:2511.16483 (2025). https://arxiv.org/abs/2511.16483

[18]. Pratama, Derry; Suryanto, Naufal; Adiputra, Andro Aprila et al., "CIPHER: Cybersecurity Intelligent Penetration-Testing Helper for Ethical Researcher", Sensors, vol. 24(21), 6878, 2024. https://doi.org/10.3390/s24216878

[19]. Wang, Lingzhi; Shi, Xinyi; Li, Ziyu et al., "Automated Penetration Testing with LLM Agents and Classical Planning", arXiv preprint arXiv:2512.11143 (2025). https://arxiv.org/abs/2512.11143

[20]. Xu, Zijie; Qi, Minfeng; Wu, Shiqing et al., "The Trust Paradox in LLM-Based Multi-Agent Systems: When Collaboration Becomes a Security Vulnerability", IEEE Transactions on Computational Social Systems, 1-12, 2026. https://doi.org/10.1109/tcss.2026.3695070

[21]. Gan, Jing Hwan; Yong Lim, Tek, "SHARKAPT: An Autonomous LLM-Orchestrated Penetration Testing Framework with MCP-Based Tool Integration", 2026 23rd International Joint Conference on Computer Science and Software Engineering (JCSSE), 85-90, 2026. https://doi.org/10.1109/jcsse68839.2026.11596758

[22]. Loevenich, Johannes; Adler, Erik; Hürten, Tobias et al., "Design and evaluation of an Autonomous Cyber Defence agent using DRL and an augmented LLM", Computer Networks, vol. 262, 111162, 2025. https://doi.org/10.1016/j.comnet.2025.111162

[23]. Bianou, Stanislas G.; Batogna, Rodrigue G., "PENTEST-AI, an LLM-Powered Multi-Agents Framework for Penetration Testing Automation Leveraging Mitre Attack", 2024 IEEE International Conference on Cyber Security and Resilience (CSR), 763-770, 2024. https://doi.org/10.1109/csr61664.2024.10679480

[24]. Şahin, Fatih, "FedMARL-LTI: Federated Multi-Agent Reinforcement Learning with LLM-Compatible Threat Intelligence for Cooperative Cyber Defense", Applied Sciences, vol. 16(16), 8278, 2026. https://doi.org/10.3390/app16168278

[25]. Happe, Andreas; Cito, Jürgen, "Getting pwn'd by AI: Penetration Testing with Large Language Models", Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2082-2086, 2023. https://doi.org/10.1145/3611643.3613083

DOI: 10.17148/IJARCCE.2026.156117

[26]. Wu, Xingyu; Tian, Yunzhe; Chen, Yuanwan et al., "CurriculumPT: LLM-Based Multi-Agent Autonomous Penetration Testing with Curriculum-Guided Task Scheduling", Applied Sciences, vol. 15(16), 9096, 2025. https://doi.org/10.3390/app15169096

[27]. Vijetha, B., "Agentic Intelligence for Unified Cyber Defense: A Self-Adaptive Framework for Threat Detection Across Cloud, Edge, and IoT Systems", IEEE Access, vol. 14, 5104-5118, 2026. https://doi.org/10.1109/access.2026.3650833

[28]. Gioacchini, Luca; Mellia, Marco; Drago, Idilio et al., "AutoPenBench: Benchmarking Generative Agents for Penetration Testing", arXiv preprint arXiv:2410.03225 (2024). https://arxiv.org/abs/2410.03225

[29]. Tian, Yifang; Liu, Yaming; Chong, Zichun et al., "GALA: Graph-Augmented LLM Agents for Root Cause Analysis and Incident Response in Microservices", arXiv preprint arXiv:2608.08968 (2026). https://arxiv.org/abs/2608.08968

[30]. Dinis, Tiago; Correia, Miguel; Tavares, Roger, "Large Language Models for Explainable Threat Intelligence", arXiv preprint arXiv:2511.05406 (2025). https://arxiv.org/abs/2511.05406

[31]. Cotti, Luca; Drago, Idilio; Rula, Anisa et al., "OntoLogX: Ontology-Guided Knowledge Graph Extraction From Cybersecurity Logs With Large Language Models", Advanced Intelligent Systems, vol. 8(6), e202501381, 2026. https://doi.org/10.1002/aisy.202501381

[32]. Sheng, Ze; Chen, Zhicheng; Gu, Shuning et al., "LLMs in Software Security: A Survey of Vulnerability Detection Techniques and Insights", arXiv preprint arXiv:2502.07049 (2025). https://arxiv.org/abs/2502.07049

[33]. He, Yiling; She, Hongyu; Qian, Xingzhi et al., "On Benchmarking Code LLMs for Android Malware Analysis", Proceedings of the 34th ACM SIGSOFT International Symposium on Software Testing and Analysis, 153-160, 2025. https://doi.org/10.1145/3713081.3731745

[34]. Chou, Chia-Hong; Sudheer, Arjun; Park, Younghee, "An LLM-Based Agentic Network Traffic Incident-Report Approach Towards Explainable-AI Network Defense", Journal of Sensor and Actuator Networks, vol. 15(2), 32, 2026. https://doi.org/10.3390/jsan15020032

[35]. Yadav, Tanya; Masum, Mohammad, "Explainable Multi-Agent LLM Framework for Phishing Email Detection via Role-Specialized Evidence Decomposition", Electronics, vol. 15(12), 2606, 2026. https://doi.org/10.3390/electronics15122606

[36]. Yang, Yoonmo; Choi, Daon; Hong, Yunyi et al., "VishBox: An AI-Agent-Based Adaptive Voice Phishing Simulation Framework for Cybersecurity Education", IEEE Access, vol. 14, 39672-39686, 2026. https://doi.org/10.1109/access.2026.3667823

[37]. Li, Han; Ma, Jinyu; Wang, Zheng et al., "UFA: An LLM-driven UAV forensic intelligent agent for low-altitude public security", Expert Systems with Applications, vol. 327, 132859, 2026. https://doi.org/10.1016/j.eswa.2026.132859

[38]. Priyadarsini, Madhukrishna, "Enhancing Security of Distributed Multi-Agent Systems in Smart Grids: An AI-Driven Approach to Regular Audits", ACM Transactions on Cyber-Physical Systems, 3748730, 2025. https://doi.org/10.1145/3748730

[39]. Liu, Zefang, "Multi-Agent Collaboration in Incident Response with Large Language Models", arXiv preprint arXiv:2412.00652 (2024). https://arxiv.org/abs/2412.00652

[40]. Liu, Jiani; He, Yixin; Fan, Lanlan et al., "PINA: Prompt Injection Attack Against Navigation Agents", ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 13897-13901, 2026. https://doi.org/10.1109/ICASSP55912.2026.11462407

[41]. Tseng, Pei-Yu; Zhang, Lan; Yeh, ZihDwo et al., "From IOCs to Regex: Automating CTI Operationalization for SOC with LLMs", arXiv preprint arXiv:2604.12228 (2026). https://arxiv.org/abs/2604.12228

[42]. Janjusevic, Strahinja; Garcia, Anna Baron; Kazerounian, Sohrob, "Hiding in the AI Traffic: Abusing MCP for LLM-Powered Agentic Red Teaming", arXiv preprint arXiv:2511.15998 (2025). https://arxiv.org/abs/2511.15998

[43]. Morris, Abby; Procter, Rachael; Wallbank, Caroline, "Evaluating Reinforcement Learning Agents for Autonomous Cyber Defence", Applied AI Letters, vol. 6(3), e125, 2025. https://doi.org/10.1002/ail2.125

[44]. Cherif, Bilel; Bisztray, Tamas; Dubniczky, Richard A. et al., "DFIR-Metric: A Benchmark Dataset for Evaluating Large Language Models in Digital Forensics and Incident Response", arXiv preprint arXiv:2505.19973 (2025). https://arxiv.org/abs/2505.19973

[45]. Castro, Sebastián R.; Campbell, Roberto; Lau, Nancy et al., "Large Language Models are Autonomous Cyber Defenders", 2025 IEEE Conference on Artificial Intelligence (CAI), 1125-1132, 2025. https://doi.org/10.1109/CAI64502.2025.00195

[46]. Ferraro, Antonino; Orlando, Gian Marco; Russo, Diego, "Generative Agent-Based Modeling with Large Language Models for insider threat detection", Engineering Applications of Artificial Intelligence, vol. 157, 111343, 2025. https://doi.org/10.1016/j.engappai.2025.111343

