

The Evolution of Agentic AI in Cybersecurity: From Single LLM Reasoners to Multi-Agent Systems and Autonomous Pipelines

Vaishali Vinay

vpapneja@microsoft.com

Microsoft Security Research

Redmond, Washington, USA

Abstract—Cybersecurity has become one of the earliest adopters of agentic AI, as security operations centers increasingly rely on multi-step reasoning, tool-driven analysis, and rapid decision-making under pressure. While individual large language models can summarize alerts or interpret unstructured reports, they fall short in real SOC environments that require grounded data access, reproducibility, and accountable workflows. In response, the field has seen a rapid architectural evolution from single-model helpers toward tool-augmented agents, distributed multi-agent systems, schema-bound tool ecosystems, and early explorations of semi-autonomous investigative pipelines.

This survey presents a five-generation taxonomy of agentic AI in cybersecurity. It traces how capabilities and risks change as systems advance from text-only LLM reasoners to multi-agent collaboration frameworks and constrained-autonomy pipelines. We compare these generations across core dimensions – reasoning depth, tool use, memory, reproducibility, and safety. In addition, we also synthesize emerging benchmarks used to evaluate cyber-oriented agents. Finally, we outline the unresolved challenges that accompany this evolution, such as response validation, tool-use correctness, multi-agent coordination, long-horizon reasoning, and safeguards for high-impact actions. Collectively, this work provides a structured perspective on how agentic AI is taking shape within cybersecurity and what is required to ensure its safe and reliable deployment.

Keywords—agentic AI, cybersecurity automation, large language models, multi-agent systems, benchmarking of AI agents, AI safety and verification, security operations center (SOC)

I. INTRODUCTION

Cybersecurity has become one of the earliest and most aggressive adopters of agentic AI. In contrast to most traditional enterprise domains, security operations centers (SOCs) continuously monitor high-volume telemetry, investigate multiple parallel alerts, and come up with threat hypotheses under severe time pressures [1]. Contemporary security workflows often include several interdependent steps: triage, enrichment, threat intelligence lookup, hypothesis building, evidence correlation, escalation, and reporting, so they are indeed multi-stage reasoning problems at heart [2]. The high level of cognitive load on analysts, continued understaffing, and increasing threat complexity have inspired the search for AI agents to help decision-intensive activities by assisting or automating these throughout the SOC lifecycle. Large language models (LLMs) were among the first AI systems shown to have strong usability in cybersecurity, summarizing alerts, extracting indicators from unstructured text, and explaining malware reports, but static LLM-only systems still don't suffice in realistic defensive environments [3]. For instance, LLMs continue to experience hallucinations,

producing plausible but factually incorrect results [4]. More importantly, they lack grounded interaction with operational data sources, as they cannot independently pull telemetry from Security Information and Event Management (SIEM), endpoint detection and response (EDR), and threat-intelligence platforms, and their reasoning is stateless, without persistent memory or audit logs, which leads to inconsistencies over lengthy investigations [5]. To deploy in a SOC context with regulatory/compliance requirements, auditability, and verifiability are critical, but many LLM-only frameworks do not offer them [6]. Furthermore, without native tool integration, they may not be helpful in executing multi-step workflows in live environments.

Study of this kind is timely, as the cybersecurity industry is rapidly accelerating to commercial levels of AI-powered copilots, automated SOC modules, and early-stage autonomous response workflows. Nevertheless, not all vendor and open-source platforms provide a sufficient description of their agentic system architectural underpinnings, and the academic literature is disjointed, with previous reviews having mainly concentrated on LLMs, on the other hand, as an abstract concept (applications, vulnerabilities, defenses) rather than the structured development of agentic AI systems that are intended for defense in cyberspace. As a result, researchers and practitioners do not come up with one common evolutionary perspective to perceive how cyber-agents evolved as agents, how they acquire new competencies, and what risk they pose.

To address this void, this survey is informed by two high-level research questions: (1) What do we know about the architectural evolution of cybersecurity agents throughout time? and (2) As evolution progresses, what capabilities and risks emerge? There are four significant fundamental contributions achieved from this. First, the research presents a five-generation taxonomy of cyber-focused, agentic AI, from early single-LLM reasoners to fully autonomous cybersecurity pipelines. Second, the study presents a cross-generational comparative analysis of capabilities, multi-step reasoning, tool-use, memory, reproducibility, and safety. Third, the study synthesizes evaluation tools to benchmark gaps in the cyber-agent domain. Fourth, we suggest structured research agendas to develop safe, reliable, and verifiable agentic AI systems for cybersecurity.

II. SCOPE AND METHODOLOGY

This survey focuses in particular on agentic AI systems for cybersecurity operations, especially on workflows of the Security Operations Center (SOC), threat-intelligence processing, detection engineering, and incident-response (IR) automation. Here, agentic AI means LLM-based reasoning is combined with tool usage, memory, planning, or verification

capabilities that facilitate multi-step action execution as opposed to static text generation. This scope accords with new studies suggesting that cyber-defense tasks increasingly involve multi-hop reasoning in communication with operational sources of data [7], [8]. The survey clearly excludes adjacent but unrelated concepts, such as classic machine-learning classification models (malware detection, spam filtering), LLM-centric red-team automation, synthetic phishing generation, or generic chat-assistant models not built to structured cyber workflows. There is a lot of prior work on these domains, which do not reflect the architectural features of agentic, tool-based systems. The temporal boundary covering included works extends between 2020 and 2025, the period of the advent of GPT-3-scale models, the maturity of AI copilots with a security focus, and the introduction of structured tool-calling and agent frameworks. This span represents the transition away from independent LLM reasoning towards orchestration, multi-agent, and tool-based architectures. The study focuses on architectural evolution rather than product comparison or vendor assessment. This approach is justified by the recent findings showing that cybersecurity environments impose distinct and high reliability, auditability, and reproducibility requirements because incorrect or unverifiable automatic actions are commonly found [9], [10], [11]. Cybersecurity is chosen as the target domain of analysis given that it is home to an API-rich operational suite, including SIEM, EDR, SOAR, sandboxing, and threat-intelligence platforms, and it naturally delivers agent behavior supported by tools. Furthermore, SOC environments are decision-based and time-critical, so they serve as a good way to assess the limitations and the potential of agentic AI systems to be able to achieve real-world operational challenges

III. BACKGROUND

The proliferation of LLMs is changing the way AI is used in cybersecurity, especially for interpreting unstructured text, reasoning over semi-structured data, and supporting analysts with high-level assessments. According to recent studies, LLMs have already been examined for cyber threat detection, threat-intelligence analysis, malware investigation support, and higher-level security reasoning, no longer confined to traditional classification pipelines or rule-based systems [12], [13]. These works indicate that LLMs provide assistance in activities such as finding IOCs, mapping text descriptions to threat techniques, and summarizing lengthy technical reports, reducing analysts' workload in large volumes [14], [15]. Simultaneously, contemporary security operations center (SOC) workflows are fundamentally tool- and data-based, and investigations are often performed relying on cross-correlating telemetry from SIEM platforms, EDR tools, network sensors, and cloud logs. It is emphasized in surveys of these SIEM technologies that they centralize log collection, real-time event correlation, incident monitoring & compliance reporting, functioning as the analytical backbone of several SOCs [16]. That way, any AI intended to be used for realistic SOC tasks must communicate with these systems rather than work directly on static, stored text.

Nevertheless, standalone LLMs face constraints that prove to be important, particularly in security, and they have all been found to hallucinate (producing fluent but factually incorrect statements), and the literature indicates that a certain degree of hallucination is intrinsic to models that generalize from their training distributions [17], [18], [19]. For long or multi-step investigations, these behaviors can undermine trust, particularly where outputs are not driven by live telemetry or coupled with explicit trace-making of their reasoning. Recent state-of-the-art reviews in the area of cybersecurity argue that LLMs facilitate innovative defensive capabilities; however, their reliability, controllability, and auditability are still key open challenges to mission-critical deployment [20]. These factors cumulatively drive a move from "LLM-as-a-chatbot" to agentic AI, systems in which LLMs serve as planners or controllers, calling tools, managing intermediate state, and working within orchestrated workflows. Current literature on the topic search looks into the application areas and limitations of LLMs in cybersecurity, yet it is not enough to trace the architectural development in the respective agentic systems over generations [10], [21], [22]. This gap paves the way for the current work, which focuses directly on the transition from single-model helpers to multi-agent, tool-integrated, and ultimately autonomous cybersecurity pipelines.

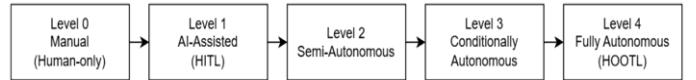


Figure 1 Five levels of autonomy for AI-based SOCs [23]

IV. GENERATION 1- SINGLE LLM REASONERS

AI security tools have built upon standalone LLMs, for instance, GPT-3 and GPT-4, which work only in text space. These Gen-1 AI agents can perform natural language processing efficiently: they summarize long cyber reports in accessible language to summarize unstructured security warnings, help the analyst make sense of them, and respond to analyst queries with contextual reasoning and the like [7], [24]. Their best feature is rapid sense-making, as an LLM can parse a phishing email, extract IOCs, recognize social-engineering clues, and assess intent in a coded way similar to human instinctual code, at an impressive speed. In response, models can condense a long malware analysis into executive summaries or associate alert descriptions with MITRE ATT&CK techniques, thereby relieving cognitive burden for beleaguered SOC [7], [24]. In practical usage, it is the case that more commonly Gen-1 LLMs are being employed as assistive copilots among analysts. In the face of noisy alerts, the model can demystify probable root cause, map relevant TTPs, or regroup jargon-laden intelligence reports into plain language to inform briefings, but even with those beneficial capabilities, Gen-1 systems are still fundamentally restricted, and one of the reason is that they don't run external queries, can't fetch data, run code, or talk to security tools [7].

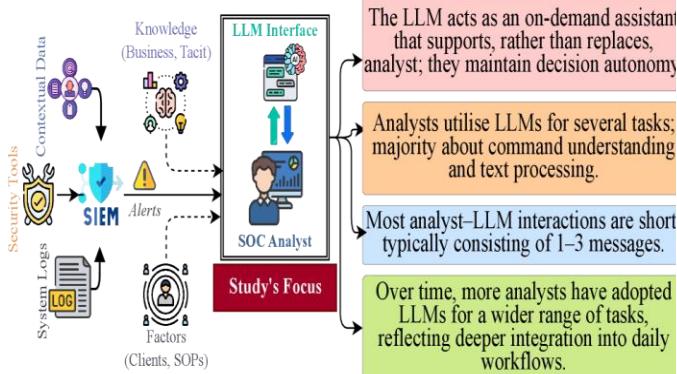


Figure 2 SOC Workflow [24]

Every response is “stateless,” as the model has no persistent memory and cannot pass information from query to query if not given explicit information. These constraints of the architecture bring with them observed vulnerabilities. While models work in probability space and do not use external verification mechanisms, if the prompt is ambiguous or data is found missing, it hallucinates or misses crucial details [23], [25]. The output of an even relatively minor prompt change can be very different, and consequently, the reproducibility will be poor. Further, because it cannot carry out actual acts, whether by reviewing logs, accessing CVE registries, or parsing output from the sandbox, they all rely upon internal pattern recognition, rendering them prone to confident yet false conclusions. So, Gen-1 LLMs can indeed produce some powerful text processing, but they are powerless to self-validate the knowledge they have developed and to act. This instability restricts them to ineffective decision-making in a high-stakes context. They remain suitable for summarizing, explaining, and threat-intel digestion [7], but analysts need to consider their outputs in a more advisory approach than authoritative. Ultimately, Gen-1 provides a valuable but limited base: a strong linguistic intelligence coupled with structural weaknesses that stop reliable automation [23], [25].

V. GENERATION 2 - TOOL-AUGMENTED AGENTS

Gen-2 systems were created to rectify the main shortcoming of Gen-1, which is the lack of action. These agents leverage the reasoning techniques from LLM as well as external tools to execute queries, run filters, access APIs, and generate structured output. Architectures including ReAct and Planner–Executor present a structured loop where the agent reasons over iterated steps, selects an action, executes a call to tools, observes results, and updates its internal plan [23]. This turns the LLM from a passive text generator into an enabler for validating information. Within cybersecurity workflows, Gen-2 agents provide significant enhancements. They have the ability to automate SIEM or data-lake queries by converting the natural language instructions into structured search expressions. Analysts can ask "Show events related to this suspicious IP," and the agent prepares and runs the correct query against the log store. Similarly, agents can supplement alerts by consulting WHOIS services, CVE databases, passive DNS sources, and threat-intelligence platforms for authoritative facts instead of hallucinating [23], [26], [27],

[28]. This tool-enhanced pattern also allows IOC extraction and pivoting, and an LLM reads a threat report, isolates hashes or domains, and forwards them to an external API for supporting evidence. The change from pure language to action-enabled workflows greatly improves reliability.

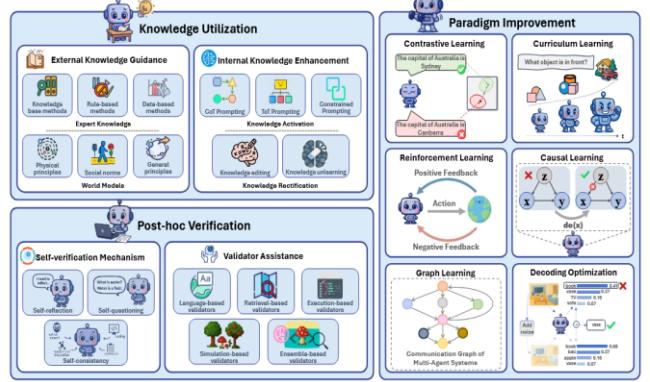


Figure 3 Hallucination mitigation [29]

By the LLM checks assumption, which is whether an indicator occurs or not, for instance, in logs, it can help to reduce unsupported claims. However, introducing multiple steps also brings new failure modes. Mistakes can propagate through the chain: an erroneous early hypothesis can cause the agent to generate a suboptimal query, which gives erroneous output, exacerbating the agent’s error. These cascades are particularly dangerous given that agents frequently construct iterative rationales, which become more and more entrenched with the first error [29]. Tool-based execution also involves the fragility of API availability: if one service goes down or returns incorrect data, the entire logic pipeline can crack. Centralization is another architectural risk. Most Gen-2 systems involve one LLM instance as the planner and executor. If the model outputs a wrong plan or misinterprets the output from the tools, the pipeline is alone with no secondary layer to intervene. Consequently, the system is now one single failure point. A high-risk, high-stakes cybersecurity operation (e.g., incident response) cannot afford such a risk without human oversight. Despite these constraints, Gen-2 agents significantly expand SOC capabilities, including semi-automated alert triage, log investigation, and contextual enrichment [23]. They’re excellent at being analyst amplifiers and not individual operators. Their strength comes in joining natural language and the real data, but they are vulnerable to error drift and ripple-level logic failures that only their persistent monitoring can alleviate [23], [29].

VI. GENERATION 3 — MULTI-AGENT CYBER SYSTEMS

Gen-3 represents a structural move away from single agents to distributed role-based teams of collaborating agents. Rather than having one LLM plan, execute, and evaluate tasks, the systems now orchestrate a series of specialized agents sharing context, critiquing one another, and modularizing the incident-response pipeline [30], [31], [32]. This is similar to the structure in which real SOC teams are defined, composed of triage analysts, investigators, and reviewers. There are generally 3 core agents for a design pattern: a triage agent, an

investigator agent, and a verifier/critic agent. The triage is where the context of incidents is determined, alert severity is classified, and the plan for investigation is developed. The investigator agent sorts through search information, parses logs, enriches IOCs, and links evidence between sources of data. Finally, the verifier actor tests whether conclusions can be deduced from evidence and detects hallucinations, inconsistencies, or otherwise significant and often missing data. This multi-level architecture is more robust and leads to more accountability. Prototypes of real-world research illustrate these principles. The CORTEX system employs a Behavior Analysis agent in its search engine interface to identify relevant workflows, Evidence Acquisition agents to execute SIEM or threat-feed queries to help in generating action plans, and lastly, a Reasoning agent to aggregate this information into an auditable triage decision [33]. IRCopilot is a third example that uses Planner, Generator, Reflector, and Analyst agents to simulate a full range of incident response processes and documentation [34]. Open-source ecosystems such as Microsoft AutoGen also make multiple-agents orchestration libraries available for developers to create similar modular workflows.

There are several advantages of the multi-agent paradigm. It allows the multi-step multi-task complexity (i.e., associating multiple log sources) to be passed between agents, leading to more reliable reasoning [33], [34]. Second, cross-verification among agents lowers the chance of unchallenged hallucinations. Third, multi-agent systems can maintain shared long-horizon context during an incident, which improves memory continuity compared with single agents, but Gen-3 has its very own class of failure modes. Agents may have competing hypotheses, resulting in diverging paths to investigate [33], [34]. Poorly constrained communication protocols may lead to infinite loops wherein agents keep saying they need clarification or another plan without converging. The CORTEX study highlights risks like “inter-agent misalignment” and “specification errors,” or scenarios where agents view roles or instructions differently from the intended ones [33], and this may result in redundancy for actions, inconsistent results, or fluctuating behavior. In a similar vein, multi-agent systems exacerbate feedback loops: a mistake from one agent that receives its stamp of approval from another can proliferate fast across the system. These problems make rigorous coordination and termination rules extremely necessary. Without guardrails, multi-agent systems (and with them a lot of complex reasoning tasks) could lead to a far longer and less understandable result, compared to the output produced by one model alone. However, despite these problems, Gen-3 is still an avenue toward complete SOC workflows because it distributes the cognitive load and enables multi-tiered quality checks [33], [34].

VII. GENERATION 4 -MCP-BASED & STANDARDIZED TOOL ECOSYSTEMS

Agentic cybersecurity systems in the fourth generation are a significant development in terms of stability, predictability, and regulation, thanks to tool schemas that are standardized [35]. Whereas in earlier generations, agents would spontaneously make API calls or create ad hoc command

structures, the Gen-4 systems are highly circumscribed. Agents interact with other enterprise security solutions like SIEM platforms, EDR agents, threat-intelligence systems, and SOAR pipelines, using a set of pre-defined schemas that dictate the formats in which each interaction is generated [35]. This schema is often translated into a JSON contract, function call structure, or a domain-specific interface that makes sure there is a predictable syntax and expected output that underpins every agent-to-tool exchange [23]. Frameworks have formalized these rules, such as the MCP, creating explicit parameters for input and output per tool that an LLM is allowed to call.

This method provides several essential advantages, like standardization increases reproducibility to begin with, and the tool call provided today, when the inputs are the same, will produce the same result tomorrow, because the agent should execute following the same schema, and it cannot fall into unstructured or ambiguous instructions. This is in stark contrast to the Gen-1 and Gen-2 systems, for which small prompt changes or LLM variability could alter system behavior [35]. Second, using a schema-dependent approach in interaction enhances safety, as agents cannot, for example, run arbitrary shell commands, change high-impact settings, or provide malformed requests, because each of these tools has a whitelisted set of callable functions. Each request must be typed, verified, and recorded, allowing for update, edit, and audit for version control and security. Agents generated via playbooks can produce standard CACAO-formatted workflows with standardized workflows instead of all other free text as well, allowing remediation instructions to be syntactically correct and machine-actionable [36]. Another unique advantage of Gen-4 buildings is in governance and compliance. All tool invocations are recorded with full transparency, including which schema/schemas, what parameters were retrieved, which data sources were used, and what results were returned, and this fits with enterprise audit requirements, incident-response documentation requirements, and regulatory controls [37], [38].

In a SOC setting, such detailed provenance helps enable an analyst to examine AI-assisted actions over time, compare outputs, and identify unauthorized deviations. The bounded execution environment also reduces incidental misuse; an LLM, for example, cannot disable a firewall rule, in the absence of these actions existing in itself as schema-controlled operations. That said, the Gen-4 systems are not without their limitations. Whilst schemas strongly dictate what action is done, they don't necessarily determine the semantic accuracy of the action, an LLM will misinterpret the evidence that is given to it at hand, so it will produce a syntactically correct but context-poor log query that ends up being incorrect, or maybe it will pick out the right tool to use to run the investigation, technically wrong [39], [40]. These semantic failures are still the most frequent because LLM-based reasoning still relies on a natural language probability space. Even when all tool calls correspond to the exact schema definitions, hallucinations might still arise in the agent's internal cognition that could guide the subsequent steps [23]. This gap provides a reason why MCP-based approaches still need human supervision. As indicated

throughout the literature, LLM-produced security actions “require human verification of outputs,” as hallucinated interpretation, inappropriate comparisons, or unfair conclusions still are not acceptable in active SOC environments [23]. Schemas do not implement correct analysis, but correct structure, and the model could correctly make a call to a SIEM API, but that means it should not only give incorrect responses, but also misjudge why the output was returned, it may misclassify severity, and/or suggest methods of remediation contrary to recommended best practice. Overall, Gen-4 architectures implement strict execution using a schema that will lead to increased reproducibility, auditability, and operational safety [36]. They give frameworks for enterprise adoption, but cannot automatically ensure analytical accuracy. Human validation and upper-level verification are still necessary to confirm that planned, well-formed actions are also contextually correct and secure.

VIII. GENERATION 5 - AUTONOMOUS CYBERSECURITY PIPELINES

The fifth generation represents the most conceptually ambitious stage, aiming to support end-to-end SOC workflows with minimal human intervention [41], [42]. Rather than needing incremental nudges and operator-centered supervision, Gen-5 agents act based on high-level goals. In these, an analyst could issue a single order like “Investigate this breach,” “Analyze this alert,” or “Generate an incident report,” and the agent establishes an entire investigative pipeline. These include extracting the appropriate alerts from SIEM platforms, correlating telemetry from EDR or network logs, enriching the indicators with threat-intelligence lookups, and aggregating data into an orderly product usable by humans. The construction of these systems is highly modular. A Gen-5 IR assistant could, for instance, autonomously pull raw alerts, map them to MITRE ATT&CK techniques, execute tailored log queries, conduct anomaly detection, and piece together correlated evidence from several data platforms and sources [43], [44]. This includes creating a situational summary of the research process, drafting a mitigation playbook, or recommending actions for containment. Systems like Microsoft Security Copilot or academic prototypes like IRCopilot exhibit elements of this idea, with discrete components for detection, investigation, containment, and remediation, each with clear roles [34]. In experimental results from multi-phase research systems, the most impressive enhancements in efficiency are observed; IRCopilot reports up to ~150% more tasks completed compared to a traditional single monolithic LLM, which emphasizes what is possible with semi-automated orchestration with constrained autonomy [34].

Gen-5 systems have several potential uses, encompassing both proactive and reactive security roles. Prototype Gen-5 autonomous IR assistants aim to automate portions of workflow from ingestion to final reporting with little human interaction, and other uses include automation/inspection detection logic, where agents autonomously adjust IDS/IPS signatures or SIEM correlation rules based on observed threat behavior [45]. Gen-5 systems may autonomously generate

new detection content, validate hypotheses across multiple data environments, and even plan partial containment workflows when malicious behavior is confirmed, but the power of Gen-5 systems comes with equally grave risks. By having autonomy, agents can run directly on live infrastructure, even issuing high-priority commands. A simulated or incorrectly mapped threat scenario could cause an agent to block production servers, close functioning processes, change firewall rules, and shut down mission-critical services without human intervention. Unverified autonomy may unintentionally trigger disruptive actions such as extended outages, security holes, and a chain reaction of operational failures, underscoring the need for strict safeguards.

The dangers go well beyond operational disruption: as an autonomous agent that handles sensitive log data, users’ IDs, and internal telemetry, the literature repeatedly warns against “data leakage” and privacy breaches when AI agents are operating unmonitored [23]. In addition, without human-in-the-loop, a faulty investigative chain that is anchored only in a single wrong assumption could spread unchecked throughout the entire system, resulting in erroneous remediation strategies. Thus, Gen-5 implementation needs to have strict security mechanisms applied. SOCs will need to implement read-only modes for default behaviors, require manual approval by others to change system state, and have layered verification of high-impact decisions [23]. Accountability requires continuous monitoring, reliable telemetry, and transparent decision-making across agents. Gen-5 systems, in practice, ought to augment analysts, like free assistants whose recommendations have to be verified by humans themselves before they can be rolled out.

IX. CROSS-GENERATION CAPABILITY COMPARISON

At five generation levels, agentic AI capabilities deepen on a number of important fronts. Gen-1 LLMs offer only single-step reasoning and limited TTP mapping based largely on static recall rather than grounded analysis [25]. Gen-2 systems are capable of simple multi-step planning when applying tool calls and can automatically search log records and simple MITRE ATT&CK retrieval, but still lack long-term memory and reliable reproducibility [23]. Gen-3’s multi-agent workflows greatly support task specialization, facilitating complex multi-phase IR pipelines in which agents perform triage, analysis, evidence correlation, or reporting [33], [34]. Common context amongst agents can likewise promote long-horizon memory. Gen-4 expands reproducibility even more through schema-bound API interactions to keep tool calls logged, typed, and replayable [36]. And those very rigid interfaces also make the program much safer by guarding against malformed or unauthorized operations. Gen-3 uses cross-agent verification, and Gen-4 uses schema validation, both of which offer stronger protections than those offered by the generations preceding them. Trade-offs remain, though: the earlier generations restrict but make things predictable, while later generations provide a greater degree of autonomy to the detriment of new failure scenarios (e.g., conflicting agent outputs, semantic misunderstanding, cascading errors, unintended automation) [23], [33]. Cumulatively, however,

capability gains are accompanied by increasing requests for oversight, verification, and control.

Table 1 Comparison

Capability Dimension	Gen-1	Gen-2	Gen-3	Gen-4	Gen-5
Reasoning Depth	Single-step	Limited multi-step via tools	Multi-agent multi-step	Structured reasoning with schemas	Full autonomous pipelines
TTP Mapping	Static recall	Simple retrieval-based	Distributed mapping via agents	Schema-validated mapping	Autonomous contextual mapping
Memory	None	Short, prompt-bound	Shared long-horizon memory	Reproducible via logs	Persistent pipeline-level memory
Reproducibility	Low	Moderate	Higher via cross-agent checks	High via schemas	High but risk-sensitive
Safety & Verification	Weak	Error-prone	Critic/validator agents	Strong schema boundaries	Requires strict safeguards

X. EVALUATION LANDSCAPE AND RESEARCH CHALLENGES

There are various benchmarks out there for measuring AI agents, but the evaluation space is far from comprehensive and not always even. As a general-purpose assessment tool, such as AgentBench, we evaluate autonomous reasoning under a wide variety of simulated environments. DefenderBench is mainly used for cyber offense and defense types of tasks, like intrusion simulation, exploit generation, and malware analysis. In security-facing settings, CyBench provides CTF-style tests, and SecEval offers structured security knowledge assessments [46]. MITRE's ATLAS framework (v2) captures adversarial AI attack patterns to guide the targeting and exploitation of agentic systems [47], [48]. Even with these advances, there are profound capability inadequacies. Perhaps most importantly, there is no available benchmark that is capable of measuring end-to-end SOC workflows from alert detection to investigation, enrichment, correlation, and final reporting in actual service use cases. This makes it an essential omission as multi-step agentic behavior provides the basis for Gen-3, Gen-4, and Gen-5 systems. No less missing are metrics for multi-agent collaboration, from division of labor to conflict resolution or consensus formation. And there is no common standard for assessing “tool-use correctness,” which is to say, if an agent’s API calls were not only syntactically correct but also semantically suitable and operationally effective. Verifier and critic agents, key traits of Gen-3 and Gen-4, lack common effectiveness metrics. Most benchmark datasets, however, (despite their potential) do not come with the noise, ambiguity, and contradictory evidence found in real SOC environments, or other mixed evidence of multiple alternatives. As discussed in existing evaluations, a realistic

multi-step SOC benchmark with strong coverage of planning, memory, tool chaining, and interplay of agents on a large scale is urgently needed in the field.

Table 2 Benchmarking

Benchmark	Reasoning Tasks	Tool Use	Cyber-Specific Tasks	Multi-Agent Evaluation	Long-Horizon Tasks	Security Sensitivity	Notes
AgentBench (2023–2024)	✓	● (limited)	✗	✗	●	Medium	General-purpose; limited cyber depth [49]
DefenderBench (2024–2025)	✓	✓	✓	✗	●	High	Strong cyber tasks; single-agent focus [46]
CyberSOCeval (2025)	✓	●	✓	✗	✗	High	Malware + TI reasoning; no orchestration [50]
CyBench / CyberBattleSim Tasks	✓	✓	✓	✗	●	High	CTF-style; offensive leaning [51]
SecEval	✓	✗	✓	✗	✗	Medium	Security Q&A; no tool use [52]
AttackSeqBench	✓	✗	✓	✗	✗	Medium	Adversarial AI + ATT&CK mapping [53]
AutoPen Bench / Red-Team Pentest Benchmarks	✓	✓	✓	✗	●	High	Offensive; exploit-focused [54]

These shortcomings highlight deeper research challenges. An urgent need exists for verification and critical models to independently analyze the correctness of agent outputs, recognizing hallucinations, invalid correlations, or unjustified conclusions [23], [29]. Limits of safety are equally important here, because if Gen-5 systems will work autonomously, they should have reasonable restrictions or fail-safes to safeguard against destructive actions, misconfigurations, or data exposure [23]. To ensure cross-agent consistency of multi-

agent systems, it has become challenging to perform research on consensus protocols and shared world-model alignment. Standard schemas, like CACAO for playbooks [36], should be modified to include SIEM, EDR, TIP, and SOAR tools with full semantic definitions. Other problems are long-horizon planning and memory retention, since existing models tend not to be able to apply well across multiple workflows. There are ethical and privacy concerns as well, as agentic systems commonly handle sensitive log records, internal telemetry, and individual user information, needing strict checks and balances to prevent bias, leakage, and maintain accountability [23].

XI. CONCLUSION

By and large, this taxonomy of generations shows progress, from 1-step Gen-1 models to multi-agent collaboration and to autonomous Gen-5 pipelines, but also underscores the parallel rise in risk and sophistication. As emphasized across the literature, the future of AI's dependable security depends on robust standards, evaluation frameworks, and principled oversight to ensure deployment that is safe, reliable, and verifiable.

REFERENCES

- [1] Y. Baddi, M. A. Almaiah, O. Almomani, and Y. Maleh, *The Art of Cyber Defense: From Risk Assessment to Threat Intelligence*. CRC Press, 2024.
- [2] P. Udayakumar and Dr. R. Anandan, "Threat Detection, Investigation, and Response," in *Design and Deploy Microsoft Azure Sentinel for IoMT: Enhance IoMT Cybersecurity Operations with Intelligent Analytics*, P. Udayakumar and Dr. R. Anandan, Eds., Berkeley, CA: Apress, 2025, pp. 243–341. doi: 10.1007/979-8-8688-2040-3_4.
- [3] F. Wu, N. Zhang, S. Jha, P. McDaniel, and C. Xiao, "A New Era in LLM Security: Exploring Security Concerns in Real-World LLM-based Systems," Feb. 28, 2024, *arXiv*: arXiv:2402.18649. doi: 10.48550/arXiv.2402.18649.
- [4] S. Banerjee, A. Agarwal, and S. Singla, "LLMs Will Always Hallucinate, and We Need to Live with This," in *Intelligent Systems and Applications*, K. Arai, Ed., Cham: Springer Nature Switzerland, 2025, pp. 624–648. doi: 10.1007/978-3-031-99965-9_39.
- [5] C. Song, L. Ma, J. Zheng, J. Liao, H. Kuang, and L. Yang, "Audit-LLM: Multi-Agent Collaboration for Log-based Insider Threat Detection," Aug. 12, 2024, *arXiv*: arXiv:2408.08902. doi: 10.48550/arXiv.2408.08902.
- [6] T. Bollikonda, "Secure Pipelines, Smarter AI: LLM-Powered Data Engineering for Threat Detection and Compliance," Apr. 16, 2025, *Preprints*: 2025041365. doi: 10.20944/preprints202504.1365.v1.
- [7] J. Zhang *et al.*, "When LLMs meet cybersecurity: a systematic literature review," *Cybersecurity*, vol. 8, no. 1, p. 55, Feb. 2025, doi: 10.1186/s42400-025-00361-w.
- [8] Z. Wang, Y. Wang, X. Xiong, Q. Ren, and J. Huang, "A Novel Framework for Enhancing Decision-Making in Autonomous Cyber Defense Through Graph Embedding," *Entropy*, vol. 27, no. 6, p. 622, June 2025, doi: 10.3390/e27060622.
- [9] F. Y. Loumachi, M. Lacerda, K. Ouazzane, A. Adnane, and O. Adamyk, "AI in Control: Rethinking Cybersecurity Compliance and Auditing," Oct. 06, 2025, *Social Science Research Network*, Rochester, NY: 5731707. doi: 10.2139/ssrn.5731707.
- [10] A. Ali and M. C. Ghanem, "Beyond Detection: Large Language Models and Next-Generation Cybersecurity," *SHIFRA*, vol. 2025, pp. 81–97, Feb. 2025, doi: 10.70470/SHIFRA/2025/005.
- [11] N. O. Jaffal, M. Alkhanafseh, and D. Mohaisen, "Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defense Techniques," *AI*, vol. 6, no. 9, p. 216, Sept. 2025, doi: 10.3390/ai6090216.
- [12] H. Xu *et al.*, "Large Language Models for Cyber Security: A Systematic Literature Review," *ACM Trans. Softw. Eng. Methodol.*, Sept. 2025, doi: 10.1145/3769676.
- [13] H. Jelodar, S. Bai, P. Hamed, H. Mohammadian, R. Razavi-Far, and A. Ghorbani, "Large Language Model (LLM) for Software Security: Code Analysis, Malware Analysis, Reverse Engineering," Apr. 07, 2025, *arXiv*: arXiv:2504.07137. doi: 10.48550/arXiv.2504.07137.
- [14] H. Alturkistani and S. Chuprat, "Artificial Intelligence and Large Language Models in Advancing Cyber Threat Intelligence: A Systematic Literature Review," Nov. 27, 2024, *Research Square*. doi: 10.21203/rs.3.rs-5423193/v1.
- [15] E. Froudakis, A. Avgetidis, S. T. Frankum, R. Perdisci, M. Antonakakis, and A. D. Keromytis, "Revealing the True Indicators: Understanding and Improving IoC Extraction From Threat Reports," Oct. 23, 2025, *arXiv*: arXiv:2506.11325. doi: 10.48550/arXiv.2506.11325.
- [16] S. Ranjithkumar and M. Mohankumar, "Security Information and Event Management (SIEM) Performance in on-Premises and Cloud Based SIEM: A Survey," in *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics*, Virtual, India: SCITEPRESS - Science and Technology Publications, 2023, pp. 627–633. doi: 10.5220/0012613800003739.
- [17] A. K. Sood, S. Zeadally, and E. Hong, "The paradigm of hallucinations in AI-driven cybersecurity systems: Understanding taxonomy, classification outcomes, and mitigations," *Computers and Electrical Engineering*, vol. 124, p. 110307, May 2025, doi: 10.1016/j.compeleceng.2025.110307.
- [18] H. F. Atlam, "LLMs in Cyber Security: Bridging Practice and Education," *Big Data and Cognitive Computing*, vol. 9, no. 7, p. 184, July 2025, doi: 10.3390/bdcc9070184.
- [19] M. Mudassar Yamin, E. Hashmi, M. Ullah, and B. Katt, "Applications of LLMs for Generating Cyber Security Exercise Scenarios," *IEEE Access*, vol. 12, pp. 143806–143822, 2024, doi: 10.1109/ACCESS.2024.3468914.
- [20] F. N. Motlagh, M. Hajizadeh, M. Majd, P. Najafi, F. Cheng, and C. Meinel, "Large Language Models in Cybersecurity: State-of-the-Art," Jan. 30, 2024, *arXiv*: arXiv:2402.00891. doi: 10.48550/arXiv.2402.00891.
- [21] Y. Yigit *et al.*, "Generative AI and LLMs for Critical Infrastructure Protection: Evaluation Benchmarks, Agentic AI, Challenges, and Opportunities," *Sensors*, vol. 25, no. 6, p. 1666, Jan. 2025, doi: 10.3390/s25061666.
- [22] Q. Zhu, "Game Theory Meets LLM and Agentic AI: Reimagining Cybersecurity for the Age of Intelligent Threats," July 14, 2025, *arXiv*: arXiv:2507.10621. doi: 10.48550/arXiv.2507.10621.
- [23] S. Srinivas *et al.*, "AI-Augmented SOC: A Survey of LLMs and Agents for Security Automation," *Journal of Cybersecurity and Privacy*, vol. 5, no. 4, p. 95, Dec. 2025, doi: 10.3390/jcp5040095.
- [24] R. Singh *et al.*, "LLMs in the SOC: An Empirical Study of Human-AI Collaboration in Security Operations Centres," Sept. 19, 2025, *arXiv*: arXiv:2508.18947. doi: 10.48550/arXiv.2508.18947.
- [25] Y. Meng *et al.*, "Uncovering Vulnerabilities of LLM-Assisted Cyber Threat Intelligence," Oct. 01, 2025, *arXiv*: arXiv:2509.23573. doi: 10.48550/arXiv.2509.23573.

- [26] D. P. F. Möller, “Threats and Threat Intelligence,” in *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices*, D. P. F. Möller, Ed., Cham: Springer Nature Switzerland, 2023, pp. 71–129. doi: 10.1007/978-3-031-26845-8_2.
- [27] S. Paul, F. Alemi, and R. Macwan, “LLM-Assisted Proactive Threat Intelligence for Automated Reasoning,” Apr. 01, 2025, *arXiv*: arXiv:2504.00428. doi: 10.48550/arXiv.2504.00428.
- [28] S. Tian *et al.*, “Exploring the Role of Large Language Models in Cybersecurity: A Systematic Survey,” Apr. 28, 2025, *arXiv*: arXiv:2504.15622. doi: 10.48550/arXiv.2504.15622.
- [29] X. Lin *et al.*, “LLM-based Agents Suffer from Hallucinations: A Survey of Taxonomy, Methods, and Directions,” Nov. 18, 2025, *arXiv*: arXiv:2509.18970. doi: 10.48550/arXiv.2509.18970.
- [30] R. Sapkota, K. I. Roumeliotis, and M. Karkee, “AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges,” *Information Fusion*, vol. 126, p. 103599, Feb. 2026, doi: 10.1016/j.inffus.2025.103599.
- [31] Ismail *et al.*, “Toward Robust Security Orchestration and Automated Response in Security Operations Centers with a Hyper-Automation Approach Using Agentic Artificial Intelligence,” *Information*, vol. 16, no. 5, p. 365, May 2025, doi: 10.3390/info16050365.
- [32] A. Bandi, B. Kongari, R. Naguru, S. Pasnoor, and S. V. Vilipala, “The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges,” *Future Internet*, vol. 17, no. 9, p. 404, Sept. 2025, doi: 10.3390/fi17090404.
- [33] B. Wei *et al.*, “CORTEX: Collaborative LLM Agents for High-Stakes Alert Triage,” Sept. 30, 2025, *arXiv*: arXiv:2510.00311. doi: 10.48550/arXiv.2510.00311.
- [34] X. Lin *et al.*, “IRCopilot: Automated Incident Response with Large Language Models,” Oct. 30, 2025, *arXiv*: arXiv:2505.20945. doi: 10.48550/arXiv.2505.20945.
- [35] I. Adabara, B. Olaniyi Sadiq, A. Nuhu Shuaibu, Y. Ibarahim Danjuma, and M. Venkateswarlu, “A Review of Agentic AI in Cybersecurity: Cognitive Autonomy, Ethical Governance, and Quantum-Resilient Defense,” *F1000Res*, vol. 14, p. 843, Sept. 2025, doi: 10.12688/f1000research.169337.1.
- [36] M. A. Gurabi, L. Nitz, R.-M. Castravet, R. Matzutt, A. Mandal, and S. Decker, “From Legacy to Standard: LLM-Assisted Transformation of Cybersecurity Playbooks into CACAO Format,” Aug. 05, 2025, *arXiv*: arXiv:2508.03342. doi: 10.48550/arXiv.2508.03342.
- [37] S. Brohi, Q. Mastoi, N. Z. Jhanjhi, and T. R. Pillai, “A Research Landscape of Agentic AI and Large Language Models: Applications, Challenges and Future Directions,” *Algorithms*, vol. 18, no. 8, p. 499, Aug. 2025, doi: 10.3390/a18080499.
- [38] S. Raza, R. Sapkota, M. Karkee, and C. Emmanouilidis, “TRISM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems,” Sept. 15, 2025, *arXiv*: arXiv:2506.04133. doi: 10.48550/arXiv.2506.04133.
- [39] Z. Li, W. Yi, and J. Chen, “Beyond Accuracy: Rethinking Hallucination and Regulatory Response in Generative AI,” Oct. 23, 2025, *arXiv*: arXiv:2509.13345. doi: 10.48550/arXiv.2509.13345.
- [40] C. Bandara, “Hallucination as Disinformation: The Role of LLMs in Amplifying Conspiracy Theories and Fake News,” *Journal of Applied Cybersecurity Analytics, Intelligence, and Decision-Making Systems*, vol. 14, no. 12, pp. 65–76, Dec. 2024, Accessed: Nov. 30, 2025. [Online]. Available: <https://sciencespress.com/index.php/JACAIIDMS/article/view/14>
- [41] A. Sheth *et al.*, “AI Driven Self-Healing Cybersecurity Systems with Agentic AI for Adaptive Threat Response and Resilience,” in *2025 IEEE Cloud Summit*, June 2025, pp. 147–153. doi: 10.1109/Cloud-Summit64795.2025.00030.
- [42] R. V. Barenji and S. Khoshgoftar, “Agentic AI for autonomous anomaly management in complex systems,” July 21, 2025, *arXiv*: arXiv:2507.15676. doi: 10.48550/arXiv.2507.15676.
- [43] P. R. Rajgopal, “SOC Talent Multiplication: AI Copilots as Force Multipliers in Short-Staffed Teams,” *International Journal of Computer Applications*, vol. 187, no. 48, pp. 46–62, None 2025, Accessed: Nov. 30, 2025. [Online]. Available: <https://ijcaonline.org/archives/volume187/number48/soc-talent-multiplication-ai-copilots-as-force-multipliers-in-short-staffed-teams/>
- [44] H. Wang, M. Xu, Y. Guo, W. Han, H. W. Lim, and J. S. Dong, “RulePilot: An LLM-Powered Agent for Security Rule Generation,” Nov. 15, 2025, *arXiv*: arXiv:2511.12224. doi: 10.48550/arXiv.2511.12224.
- [45] A. Habibzadeh, F. Feyzi, and R. E. Atani, “Large Language Models for Security Operations Centers: A Comprehensive Survey,” Sept. 19, 2025, *arXiv*: arXiv:2509.10858. doi: 10.48550/arXiv.2509.10858.
- [46] C. Zhang *et al.*, “DefenderBench: A Toolkit for Evaluating Language Agents in Cybersecurity Environments,” Oct. 14, 2025, *arXiv*: arXiv:2506.00739. doi: 10.48550/arXiv.2506.00739.
- [47] C. Wymberry and H. Jahankhani, “An Approach to Measure the Effectiveness of the MITRE ATLAS Framework in Safeguarding Machine Learning Systems Against Data Poisoning Attack,” in *Cybersecurity and Artificial Intelligence: Transformational Strategies and Disruptive Innovation*, H. Jahankhani, G. Bowen, M. S. Sharif, and O. Hussien, Eds., Cham: Springer Nature Switzerland, 2024, pp. 81–116. doi: 10.1007/978-3-031-52272-7_4.
- [48] Y.-T. Huang *et al.*, “MITREtrieval: Retrieving MITRE Techniques From Unstructured Threat Reports by Fusion of Deep Learning and Ontology,” *IEEE Transactions on Network and Service Management*, vol. 21, no. 4, pp. 4871–4887, Aug. 2024, doi: 10.1109/TNSM.2024.3401200.
- [49] X. Liu *et al.*, “AgentBench: Evaluating LLMs as Agents,” Oct. 04, 2025, *arXiv*: arXiv:2308.03688. doi: 10.48550/arXiv.2308.03688.
- [50] L. Deason *et al.*, “CyberSOCEval: Benchmarking LLMs Capabilities for Malware Analysis and Threat Intelligence Reasoning,” Nov. 10, 2025, *arXiv*: arXiv:2509.20166. doi: 10.48550/arXiv.2509.20166.
- [51] A. K. Zhang *et al.*, “Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models,” presented at the The Thirteenth International Conference on Learning Representations, Oct. 2024. Accessed: Dec. 01, 2025. [Online]. Available: <https://openreview.net/forum?id=tc90LV0yRL>
- [52] Z. Liu, J. Shi, and J. F. Buford, “CyberBench: A Multi-Task Benchmark for Evaluating Large Language Models in Cybersecurity,” presented at the AAAI-24 Workshop on Artificial Intelligence for Cyber Security (AICS), Vancouver, 2024.
- [53] H. Ma *et al.*, “AttackSeqBench: Benchmarking Large Language Models in Analyzing Attack Sequences within Cyber Threat Intelligence,” Mar. 01, 2025, *arXiv*. doi: 10.48550/arXiv.2503.03170.
- [54] L. Gioacchini, M. Mellia, I. Drago, A. Delsanto, G. Siracusano, and R. Bifulco, “AutoPenBench: Benchmarking Generative Agents for Penetration Testing,” Oct. 28, 2024, *arXiv*: arXiv:2410.03225. doi: 10.48550/arXiv.2410.03225.