



NORTHWESTERN UNIVERSITY

Computer Science Department

Technical Report

Number: NU-CS-2026-08

March 2026

Automated Penetration Testing with LLM Agents and Classical Planning

**Lingzhi Wang, Xinyi Shi, Ziyu Li, Yi Jiang, Shiyu Tan, Yuhao Jiang,
Junjie Cheng, Wenyuan Chen, Xiangmin Shen, Zhenyuan LI, Yan Chen**

Abstract

While penetration testing plays a vital role in cybersecurity, achieving fully automated, hands-off-the-keyboard execution remains a significant research challenge. In this paper, we introduce the "Planner-Executor-Perceptor (PEP)" design paradigm and use it to systematically review existing work and identify the key challenges in this area. We also evaluate existing penetration testing systems, with a particular focus on the use of Large Language Model (LLM) agents for this task. The results show that the out-of-the-box Claude Code and Sonnet 4.5 exhibit the strongest penetration capabilities observed to date, substantially outperforming all prior systems. However, a detailed analysis of their testing processes reveals specific strengths and limitations; notably, LLM agents struggle with maintaining coherent long-horizon plans, performing complex reasoning, and effectively utilizing specialized tools. These limitations significantly constrain their overall capability, efficiency, and stability. To address these issues, we propose CHECKMATE, a framework that integrates enhanced classical planning with LLM agents, providing an external, structured "brain" that mitigates the inherent weaknesses of LLM agents. Our evaluation shows that CHECKMATE outperforms the state-of-the-art system (Claude Code) in penetration capability, improving benchmark success rates by over 20%. In addition, it delivers substantially greater stability, reducing both time and monetary costs by more than 50%.

Keywords

Cyberattacks, Penetration Testing, LLM, Classical Planning.

Automated Penetration Testing with LLM Agents and Classical Planning

Lingzhi Wang*, Xinyi Shi*, Ziyu Li*, Yi Jiang[†], Shiyu Tan[†], Yuhao Jiang*, Junjie Cheng[†], Wenyan Chen[†],
Xiangmin Shen[‡], Zhenyuan LI[†], Yan Chen*

*Northwestern University, [†]Zhejiang University, [‡]Hofstra University

Abstract—While penetration testing plays a vital role in cybersecurity, achieving fully automated, hands-off-the-keyboard execution remains a significant research challenge. In this paper, we introduce the “Planner-Executor-Perceptor (PEP)” design paradigm and use it to systematically review existing work and identify the key challenges in this area. We also evaluate existing penetration testing systems, with a particular focus on the use of Large Language Model (LLM) agents for this task. The results show that the out-of-the-box Claude Code and Sonnet 4.5 exhibit superior penetration capabilities observed to date, substantially outperforming all prior systems. However, a detailed analysis of their testing processes reveals specific strengths and limitations; notably, LLM agents struggle with maintaining coherent long-horizon plans, performing complex reasoning, and effectively utilizing specialized tools. These limitations significantly constrain its overall capability, efficiency, and stability. To address these limitations, we propose CHECKMATE, a framework that integrates enhanced classical planning with LLM agents, providing an external, structured “brain” that mitigates the inherent weaknesses of LLM agents. Our evaluation shows that CHECKMATE outperforms the state-of-the-art system (Claude Code) in penetration capability, improving benchmark success rates by over 20%. In addition, it delivers substantially greater stability, cutting both time and monetary costs by more than 50%.

Index Terms—Cyberattacks, Penetration Testing, LLM, Classical Planning

I. INTRODUCTION

Penetration testing, hereafter used interchangeably with “pentesting,” has become a vital component of cybersecurity. It enables organizations to proactively identify and mitigate vulnerabilities before exploited by adversaries. The U.S. Cybersecurity and Infrastructure Security Agency (CISA) highlights pentesting as a core service for protecting critical infrastructure [1]. MarketsandMarkets [2] forecasts that the global pentesting market was valued at USD 1.7 billion in 2024, and is projected to grow to USD 3.9 billion by 2029 [3]. The PenTesting-as-a-Service (PTaaS) segment alone is expected to expand from USD 118 million in 2024 to USD 301 million by 2029. This significant and rising demand has prompted the adoption of artificial intelligence (AI) to meet the need for more efficient and scalable penetration testing solutions. A recent research [4] suggests that by 2028, AI-powered security testing tools will outnumber human pentesters in mainstream security operations, signaling a transformative shift in the field.

The high reliability of skilled human pentesters has made fully automated penetration testing a long-standing challenge in cybersecurity research. However, existing approaches

still require nontrivial human intervention and have yet to achieve truly “hands-off-the-keyboard” automation [5]. As a result, users are often forced to act as supervisors, executors, prompters, and decision-makers, limiting the practical utility of these systems. Moreover, human intervention hinders rigorous and unbiased evaluation, as it is difficult to standardize the extent of human involvement during evaluation, leading to inconsistent assessments across studies. In real-world use, different human knowledge, experience, and skill levels make it harder for the system to achieve its expected effectiveness.

Multiple challenges have prevented truly “hands-off-the-keyboard” automation in pentesting. Earlier work based on formal models [6], [7] is unable to interpret heterogeneous, unstructured information or generate specific commands. Therefore, they either operate on overly simplified scenarios or produce abstract attack plans without concrete commands or instructions. Recent advances in LLMs enable heterogeneous information processing and detailed instruction generation, yet human involvement remains necessary in LLM-based pentesting. First, LLM hallucinations often lead to incomplete, incorrect, or even fabricated commands, requiring human supervision to identify and correct them before execution [5]. Second, limited context memory and logical reasoning capabilities [8]–[11] make LLMs ineffective at performing long-term, multi-step pentesting tasks [5]. They often become stuck in repetitive and unproductive iterations and require human input and guidance.

Recently, LLMs have been widely applied to complex tasks such as software development [12]–[14], code auditing [15], [16], and data engineering [17], prompting a focus on *LLM Agents*. These systems are characterized by their ability to automatically plan for complex tasks, execute code and commands, and iteratively refine them based on execution results, significantly reducing the need for human guidance and intervention. In this paper, we first try to answer the question: *Can LLM agents conduct pentesting independently?* To answer this, we evaluated and compared their penetration capabilities against prior work. Our results show that the out-of-the-box Claude Code [18] and Sonnet 4.5 can autonomously complete pentesting tasks and achieve state-of-the-art (SOTA) capability, substantially outperforming all prior work in this field. To understand the strong performance of Claude Code, we further analyze its testing process and find that it demonstrates superior code refinement and subtask management capabilities. However, we also identify several limitations in Claude Code

and other LLM agents in pentesting, including difficulties in maintaining coherent long-horizon plans, performing complex reasoning, and leveraging specialized tools. Such issues lead to lower success rates, efficiency, and unstable performance.

In this paper, we address these limitations by integrating classical planning techniques with LLM agents. Our approach is motivated by several key insights. First, classical planning represents plans as directed acyclic graphs (DAGs) [19], which provide an explicit and logically structured view of the overall task. This structure enables the agent to maintain a coherent long-horizon plan and mitigates common failure modes, such as erratic attack-path selection, incomplete or repetitive execution, and forgetting previously gathered information. Second, classical planning explicitly encodes causal relationships through preconditions and effects. This allows us to generate long reasoning chains without relying on the LLM itself, thereby ensuring their correctness while maintaining efficient and stable reasoning. Moreover, this explicit causal structure enables us to incorporate customized domain knowledge and reasoning patterns that go beyond the LLM’s internal knowledge. Last but not least, the use of modular actions defined in classical planning allows us to easily integrate uncommon or highly specialized pentesting tools, reducing reliance on the LLM’s internal knowledge for command generation. These predefined actions also let us specify critical execution details in advance, eliminating the need for repeated LLM iterations to refine simple commands on the fly. This design further improves accuracy, efficiency, and overall system stability.

However, traditional classical planning is limited to fully observable and deterministic environments, making it unsuitable for pentesting scenarios. To address this gap, we propose *Classical Planning+*, the first LLM-augmented classical planning framework capable of dynamic updates. Classical planning+ preserves the core principles of classical planning while leveraging LLMs to update action effects and state information on the fly, thereby removing the requirement to fully specify all domain knowledge beforehand. As a result, it extends classical planning to partially observable and non-deterministic domains, significantly broadening its applicability to real-world pentesting tasks.

In this paper, we aim to enhance LLM-based pentesting with classical planning+. We decompose a pentesting system into three core components: a planner, an executor, and a perceptor. Using this design paradigm, we review the design choices in existing work and introduce our system, CHECKMATE. In our system, classical planning+ serves as the planner: it infers the set of feasible actions given the current attack state and selects the most appropriate next step. LLM agents act as the executor that carries out these actions. Their execution outputs are then converted by an LLM into predicates compatible with classical planning+, enabling the planner to update the attack state and determine subsequent steps. Our experiments on the Vulhub [20] dataset demonstrate that CHECKMATE significantly outperforms existing systems, including Claude Code, in penetration capability, efficiency, and stability.

Our main contributions are summarized as follows.

- 1) **Proposed a Unified Design Paradigm for Pentesting.** We propose a design paradigm for automated pentesting systems, which consists of three fundamental components: a Planner, an Executor, and a Perceptor (PEP). We use this paradigm to review and categorize existing work. This paradigm helps us understand the distinctions among existing works and provides guidance for the future development and improvement of such systems.
- 2) **Performed the Largest Evaluation on Existing Systems.** We systematically evaluate existing pentesting systems on the Vulhub dataset. We find that out-of-the-box Claude Code, powered by Sonnet 4.5, achieves the strongest performance with the least human intervention, showing a substantial improvement over all prior work in this area. We further analyzed the penetration workflow of Claude Code and identified three major limitations.
- 3) **Proposed Classical Planning+ and Developed CHECKMATE.** To address these limitations, we propose classical planning+, the first dynamically updating classical planning scheme powered by LLMs, which extends conventional classical planning to partially observable and non-deterministic tasks. Based on it, we develop CHECKMATE, which integrates classical planning and LLM agents for pentesting. Our extensive evaluation demonstrates that CHECKMATE substantially outperforms prior systems, improving success rates on Vulhub benchmark by over 20% and cutting both time and monetary costs by more than 50%.

II. THE PEP PARADIGM AND RELATED WORK

A. The PEP Designing Paradigm for Pentesting Systems

With the growing volume of work in automated pentesting, we consolidate a design paradigm that decomposes an automated pentesting system into three cooperating components: a Planner, an Executor, and a Perceptor (PEP). This decomposition provides a clear way to dissect systems and a framework for future research, allowing each component to be independently analyzed, improved, and benchmarked. Using the PEP paradigm, we taxonomized prior work and, for each component, reviewed representative solutions, remaining challenges, and open problems. The prior work taxonomy is summarized in Table I.

1) *Planner*: The planner aims to answer: (1) what actions are feasible now; (2) among those feasible actions and their following attack paths, which one has the highest value and should be given priority for execution. Earlier work used formal methods as the planner. For example, some work [6], [21]–[24] presents pentesting as a partially observable Markov decision process (POMDP) to model the uncertainty and incomplete information, where the planner must choose between information-gathering (e.g., scans) and offensive (e.g., exploit) actions, update a Bayesian-style belief state from observations, and enumerate feasible actions under the current belief state. Follow-up research has extended POMDP to account for factors such as defender responses [21] or observation noise [23]–[25]. However, these systems suffer computational

blow-up as the problem scales [6]. Moreover, their estimated probability models (e.g., scan success rates or exploitability probabilities) are difficult to obtain in the real world. Another line of work [7], [26]–[29] formulates planning as a path-search problem. CHAINREACTOR [26], as a representative work, casts privilege-escalation attacks as a classical planning task. However, those approaches treat pentesting as a static and deterministic problem, and thus do not address how to plan under those real-world uncertainties and incomplete information.

LLMs can be naturally used as a planner by supplying pentesting information as context and requesting the next action. However, it suffers from some limitations, such as hallucinations, short-term memory, and limited context windows. Therefore, existing work has designed various mechanisms to help LLMs plan. For example, some systems maintain structured textual representations of the plan, including the Penetration Tree (PTT) and its variants in PENTESTGPT [30], AUTOPENTESTER [31], and PENHEAL [32], as well as the Situation Summaries in AUTOATTACKER [33]. VulnBot [34] leverages LLMs to convert the planning problem to a Penetration Task Graph (PTG). PENTESTAGENT [35] uses LLMs together with CVE/service keywords to drive a two-stage planning. It leverages LLMs to interpret the reconnaissance results and extract CVEs or service names, which are used to search for exploits online. Multi-agent systems like CAI [4] rely on LLMs to coordinate available agents or tools. In these frameworks, the LLM has function descriptions and pentesting information as context and determines which agent or tool to invoke at each step. Some systems also maintain a to-do list to guide the agent’s progress and prevent LLMs from being distracted.

Challenges and open questions: Existing pentesting planners can be broadly classified into two categories based on whether they rely on LLMs. The first category comprises traditional planning methods such as POMDPs and classical planning. These approaches offer a clear logical structure. However, they struggle to scale to real-world pentesting environments, where both the state space and action space are large, dynamic, and highly complex. The second category relies on LLMs to perform planning. These approaches do not require formal definitions of actions or states, which greatly simplifies system design. However, the resulting plans often suffer from logical inconsistency and poor long-term coherence. Moreover, the black-box nature of LLMs makes the planning difficult to control, interpret, or systematically improve. Therefore, there are two research directions. The first one is automatically extracting structured knowledge to enable formal planning algorithms in complex penetration scenarios. The second one is enhancing LLM-based planning so that the generated plans exhibit stronger logical consistency and coherence.

2) *Executor*: The executor is responsible for (1) translating the planning results into concrete, executable commands and (2) executing those commands on real systems. Systems without LLMs can only provide commands within a very narrow scope (e.g., a small set of Metasploit exploitations). Some

of them (e.g., CHAINREACTOR) require human operators to execute the generated attack plans [26]. LLMs enable fine-grained generation of commands and code, leading to LLM-driven executors in pentesting, such as the executor of PENTESTGPT. However, these executors cannot interact with target environments, so the generated commands still depend on human operators to execute. With the rise of LLM-based tool calling, systems like PENTESTAGENT, AUTOPENTESTER, CAI are able to synthesize, execute, and iterate commands, significantly reducing human intervention. Retrieval-augmented generation (RAG) has also been widely adopted in executor. PENTESTAGENT, AUTOPENTESTER, VULNBOT, and PENHEAL, for example, combine retrieved code snippets, articles, and previous actions with RAG pipelines to improve the quality of LLM-generated commands.

Challenges and open questions: A major challenge for executors is simulating human-like behaviors and interactions. Many attack vectors only appear through GUIs or interactive workflows where text-only commands and tools are less effective, especially in web penetration scenarios. Mimicking human behaviors (e.g., mouse and keyboard actions) triggers such attack vectors and helps evade defenses. Although the Computer-User Interaction Simulation Agent (CUA) has been proposed to mimic human behaviors in interface operations [36], [37], no existing work has yet applied it for pentesting. Another challenge is effectively leveraging specialized tools that may lie outside an LLMs’ training data.

3) *Perceptor*: The perceptor is responsible for converting heterogeneous, unstructured data, such as tool outputs and error messages, into representations that the planner can use. When planning relies solely on LLMs, this unstructured data can be provided directly to the LLM as context, so a dedicated perceptor is unnecessary. For planners that depend on structured intermediate representations (e.g., PTTs or to-do lists), the perceptor uses an LLM to translate heterogeneous data into the data structures, such as a PTT branch or an item in the to-do list. For a classical planner, unstructured information is mapped to symbolic predicates, either through manually crafted rules or by an LLM.

Challenges and open questions: Existing work focuses primarily on textual data, while visual information is also important in pentesting. For example, analysts may need to infer a web application’s functionality from its user interface or extract data embedded within images (e.g., reading CAPTCHA). Developing future perceptrs, therefore, requires addressing the challenges of robust visual understanding. Although an increasing number of LLMs and multimodal models now offer image-analysis capabilities, to the best of our knowledge, no prior work has leveraged visual artifacts effectively in the context of pentesting.

B. Classical Planning

Classical planners operate on state representations explicitly defined by predicates. Every action includes clearly defined preconditions and effects, which allow the planner to know exactly which actions are applicable at any given state. This

TABLE I: Taxonomy of automated penetration testing systems based on the PEP design paradigm.

System	Planner	Executor	Perceptor
CHAINREACTOR	Classical Planning	Predefined Actions + Human Operators	Rules + LLM (PDDL predicates)
PENTESTGPT	LLM + Penetration Tree	LLM + Human Operators	LLM
AutoPT	LLM + Finite State Machine	LLM + Agents	LLM
PENTESTAGENT	LLM + CVE-Exploit Mapping	LLM + RAG (code snippets) + Agents	LLM
AutoAttacker	LLM + Situation Summary	LLM + RAG (previous tasks) + Agents	LLM
VULNBOT	LLM + Penetration Task Graph	LLM + RAG (previous tasks) + Agents	LLM
PENHEAL	LLM + Penetration Tree	LLM + RAG (previous commands) + Agents	LLM
CAI	LLM	Multiple Tool Agents	-
AutoPentester	LLM + Modified Penetration Tree	LLM + RAG (articles) + Agents	LLM
CheckMate	Classical Planning+	LLM + Predefined Actions + Agents	LLM

symbolic grounding guarantees that valid actions are not overlooked, actions are only applied when their preconditions are satisfied, and all changes to the world state are explicitly and consistently tracked throughout the planning process. In addition, classical planning algorithms guarantee that if a valid sequence of actions exists, the planner must be able to discover it [38]. Moreover, every intermediate step in the resulting plan is logically consistent with the defined preconditions and effects, ensuring that each causal dependency is correctly maintained. Consequently, even for long action chains, the planner constructs solutions in a step-by-step manner, preserving the causal structure throughout the entire planning process.

In contrast, LLM-based planning relies on implicit, language-based reasoning [39], [40]. It lacks a persistent and structured memory of the world state, which makes it prone to forgetting past actions, repeating steps, or hallucinating outcomes, especially as the reasoning chain grows longer [41]–[43]. This also leads to skipped steps in the action sequence or invalid transitions. It also suffers from the limited context windows [44] (e.g., 8K–128K tokens), which restrict its ability to retain long-term planning structure [40], especially in complex tasks like penetration testing.

III. EVALUATION OF EXISTING PENTESTING SYSTEMS

A. Experimental Methodology & Setup

1) *Benchmark Datasets*: We adopt Vulhub [20], a community-maintained collection of containerized vulnerable environments, as the basis of our benchmark. From this repository, we randomly sampled 120 containers for evaluation. All target Docker images were anonymized, preventing the evaluation system from recognizing them as Vulhub challenges. Compared with recent work [5], [30], [45], our benchmark is the largest of its kind to date. We exclude puzzle-like challenges [46] such as those from HackTheBox [47], which emphasize more on CTF-style tricks. In addition, these challenge sets (e.g., [48]) have extensive public writeups, many of which are likely included in LLM training corpora, posing a risk of data contamination. To maintain experimental fairness and unbiasedness, we therefore do not incorporate them into our benchmark.

2) *Metrics*: To measure the actual penetration capability and progress of a pentest, we propose eleven milestones that cover the typical pentesting lifecycle. Please note that we did

not adopt the “sub-tasks” used in some prior work [30] as the metric. Because those sub-tasks emphasize completing specific activities rather than demonstrating meaningful impact. For example, a sub-task such as “web enumeration” can be checked off after enumeration was performed; that alone does not show that the tester discovered key information or used it effectively in later phases of the test. For this reason, we define milestones to assess real progress. For each pentesting engagement, we manually compare the testing process against the ground truth and judge which milestones have been achieved. The milestones are as follows:

- M1: Successfully enumerating network hosts, open ports, and running services.
- M2: Discovering multiple potential attack vectors (e.g., target services or software) without confirming the actual exploitable one.
- M3: Confirming and precisely localizing specific attack vectors susceptible to exploitation.
- M4: Obtaining or generating an exploitation command, code, or method.
- M5: Successfully executing the exploit that triggers the vulnerability or verifies the PoC.
- M6: Successfully executing arbitrary commands on the target system.
- M7: Establishing an interactive shell session with user-level privileges.
- M8: Discovering a viable privilege escalation method.
- M9: Establishing an interactive shell with elevated privileges (root on Linux/Unix, Administrator/SYSTEM on Windows).
- M10: Successful lateral movement.
- M11: Obtaining authentication credentials or private data in any format.

The milestones exhibit both sequential dependencies and parallel paths. In a typical linear progression, the sequential flow follows the pattern from M1 to M9, with each milestone requiring completion before advancing to the next. However, M10 and M11 can be pursued in parallel. For example, the tester may simultaneously explore lateral movement and privilege elevation on compromised systems, and user privacy and credentials can be obtained at any stage throughout the pentesting process. This milestone framework enables a more nuanced assessment of penetration progress and penetration capabilities.

3) *Baselines and Evaluation Criteria:* We chose four open-source, well-recognized pentesting systems, PENTESTGPT [30], PENTESTAGENT [45], CAI [4], and AutoPenTester [49], as the baselines. Related works such as AUTOATTACKER [33] and Penheal [32], as well as commercial systems like XBOW [50], do not release their code, making independent reproduction infeasible. CHAINREACTOR [26] provides an open-source implementation but narrowly focuses on privilege escalation attacks. Additionally, while several open-source pentesting toolkits [31], [35], [51] support automation primitives, they lack autonomous planning modules and thus cannot perform end-to-end automated pentesting. Lastly, traditional automated pentesting works are excluded because they neither offer reproducible implementations nor have automation capabilities.

Minimal Human Intervention. We noticed that some systems still rely on human intervention, such as interpreting system outputs, extracting key information, and guiding LLMs in choosing tools. We believe rigorous evaluation should minimize human involvement to ensure consistency, fairness, and unbiasedness. Accordingly, we followed the principle of minimal human intervention. For systems that could not operate in a fully hands-off-the-keyboard manner, we allowed only essential interactions, strictly limited to selecting default options, executing provided commands, reporting execution outcomes, etc, without offering external knowledge or guidance.

B. Comparative Evaluation of Existing Systems

In addition to the baselines representing the prior work on automated pentesting, we also evaluated the penetration capabilities of three out-of-the-box LLM agents: Claude Code [18] + Sonnet 4.5, Codex [52] + o4-mini, and Gemini Code Assistant [53] + Gemini Pro 2.5. We fed each system the same initial prompt with the task description and allowed them to use any tool that ships in a standard Kali Linux distribution. Beyond that, we provided no additional hints or human intervention. Any single step that stalled for more than two hours was terminated and counted as a failure. For PENTESTGPT [30], PENTESTAGENT [35], and CAI [4], we employed the most powerful LLM that each system supports. We measured the percentage of targets that each system advanced to each milestone.

The results shown in Figure 1 indicate that Claude Code + Sonnet 4.5 consistently outperforms all other systems across almost all milestones. The performance of PENTESTGPT drops sharply after M1, indicating its limited pentesting ability without human intervention. Although the remaining systems completed the early milestones involving basic reconnaissance and enumeration, their performance diverged significantly once the workflow demands deeper reasoning and planning, as well as exploit development. PENTESTAGENT outperforms CAI, Codex, and Gemini Code Assist due to its online exploit-search strategy, but it still fails to make progress beyond M4. In contrast, Claude Code maintains strong performance through M7 and still achieves some success in later stages, demonstrating a significantly better capability in multi-step

penetration tasks. Please note that M8 to M11 correspond to lateral movement, privilege escalation, and credential leakage. Because the Vulhub dataset simulates single-application vulnerabilities, it may not provide the attack paths necessary to reach these milestones. Overall, these results highlight two key findings. First, the out-of-the-box Claude Code + Sonnet 4.5 demonstrates substantially stronger capabilities for automated penetration testing than all prior systems evaluated in this domain. Second, this level of capability is not uniform across LLM-based code agents: Codex and Gemini Code Assist fail to progress beyond basic scanning and enumeration, whereas Claude Code consistently performs a larger number of successful follow-on actions after initial discovery. A detailed analysis of each system’s penetration process and the factors causing the performance gaps is presented in §III-C.

C. Discussion on Capabilities and Limitations

To investigate the reasons behind the strong performance of Claude Code + Sonnet 4.5, we conducted an ablation experiment with two modified configurations: (1) replacing Claude Code with an alternative agent, OpenInterpreter, while retaining Sonnet 4.5 as the backend LLM, and (2) retaining Claude Code as the agent but substituting Sonnet 4.5 with GPT-o4-mini. All other setups were the same. The results show that both alternative configurations had a substantial drop in performance and failed to achieve M3 on any task. The pentesting processes also revealed two major weaknesses. First, the agents occasionally failed to proceed independently and required human input to determine the next step. Second, they generated redundant steps, such as creating local files or performing unrelated checks before executing actions like port scanning. Overall, these results suggest that the strong pentesting capability derives from the combination of both Claude Code’s agentic control and Sonnet 4.5’s underlying model capabilities, and that removing either component leads to a significant loss in capability.

We then analyzed the detailed pentesting processes of all systems to answer two key questions: how does Claude Code outperform other systems, and where does it still fall short?

1) *Advantages of Claude Code:* PENTESTGPT relies on a human-in-the-loop workflow, while other systems incorporate command-execution capabilities, enabling a higher degree of automation. The LLM code agents are better at iteratively debugging and modifying commands based on the execution results compared to CAI. Although PENTESTAGENT excels in searching and leveraging online exploits, its performance lags in non-exploit tasks, such as enumeration and application probing. Among those LLM agents, Claude Code stands out with several strengths. First, Codex and Gemini Code Assist discover narrower attack surfaces and often select slower and less effective approaches (e.g., excessive password brute-forcing or path enumeration), while Claude Code can discover broader attack surfaces. Second, Codex and Gemini Code Assist demonstrate limited self-reflection and adjustment capabilities. They often fail to recognize when a chosen route is unproductive and may remain stalled on the same

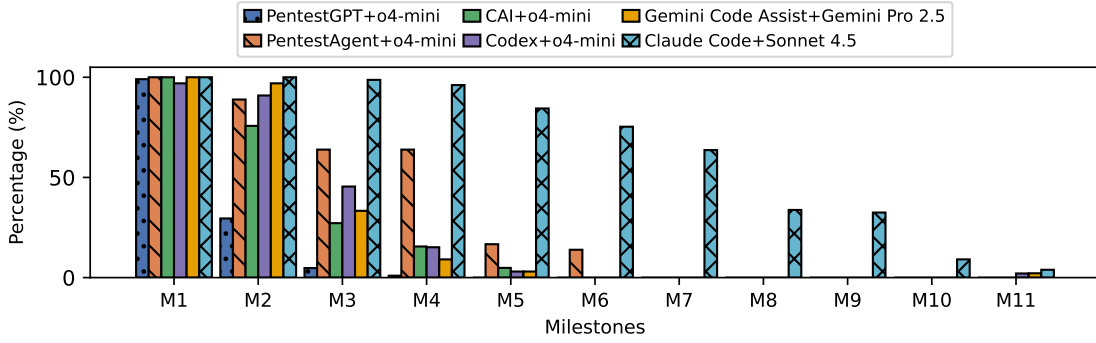


Fig. 1: Comparison of penetration capabilities of existing automated pentesting systems on Vulhub benchmark.

command for extended periods. In contrast, Claude Code continuously monitors command execution, detects potential deadlocks, and autonomously terminates stalled processes to pursue alternative actions. Claude Code is even capable of parallel multitasking. It can start a new task while another is still running, and if the new task produces more valuable findings, it will automatically kill the previous ones. Lastly, despite being explicitly instructed not to request human input, Codex and Gemini Code Assist frequently paused to request user decisions or inputs.

2) *Limitations of Claude Code:* Despite its leading performance, we identify three limitations of Claude Code, which also exist in other LLM agents.

Claude Code often fails to maintain a coherent attack plan: Claude Code does not follow a consistent, strategic sequence of actions, leading to repeated work, abandoned partial attempts, and unstable performance. Its decision-making tends to execute whatever actions come to “mind” first. For example, after identifying vulnerable applications, the agent may initially search for exploits in Metasploit, then abruptly switch to GitHub. Even after downloading a potential exploit script, it may suddenly abandon that path to write a new exploit for a different target or pursue an unrelated attack vector. This lack of strategic consistency also appears in basic tasks such as port scanning. Claude Code may perform a full port scan in one session, scan only the top 1000 ports in another, or scan its own list of “common ports” in another task. Such unpredictability makes the pentesting workflow difficult to anticipate or control. It causes the agent to diverge from optimal methodologies, overlook viable attack vectors, and waste significant time.

Claude Code struggles with long-term and experience-driven reasoning: Reasoning in pentesting means inferring causal relationships from received information to feasible actions. For example, when a web application is discovered, a typical reasoning process would involve enumerating version information, identifying all relevant exploits, and then selecting the most promising one to attempt. However, this type of long-term, multi-step reasoning is difficult for LLMs. They may skip enumeration or investigation steps and jump directly to generating an exploit based on their internal knowledge. While this can occasionally speed up

the process, it also increases the risks of hallucinated steps, inconsistent performance, and unnecessary token usage. In addition, LLMs struggle with experience-driven reasoning, the ability to leverage subtle cues in pentesting. For instance, a URL pattern like “/node/{number}” can hint at a Drupal backend. An experienced human pentester would immediately consider Drupal-specific attack paths after identifying such URLs. In contrast, LLMs often fail to recognize this kind of implicit linkage, leading to missed opportunities.

Claude Code has difficulty using specialized pentesting tools: We also observed a tendency that Claude Code favors crafting custom scripts instead of first leveraging established, specialized pentesting tools, which diverges from human pentesting methodology. For example, Claude Code frequently generates custom curl commands to probe web applications, even though thousands of highly effective Nuclei scanning templates already exist and would provide broader, faster, and more reliable coverage. We assume this is largely because such tools appear less frequently in LLMs’ training data.

IV. SYSTEM DESIGN

A. Overview

In this paper, we present CHECKMATE, a system designed to overcome limitations of existing LLM-based pentesting frameworks. Following the PEP diagram proposed in §II-A, CHECKMATE consists of three major components: classical planning+ as the planner, an LLM agent as the executor, and an LLM as the perceptor. The overall design of CHECKMATE is illustrated in Figure 2. Specifically, we introduced predefined attack actions to expand LLM’s knowledge on the specialized tools. Classical planning+ is leveraged to plan the next action, which is executed by an LLM agent. An LLM is used to interpret execution results and update the planner for further planning. Instead of relying on the LLM agent for the entire pentesting workflow, CHECKMATE restricts the LLM’s role to a pure perceptor and a simple-task executor. This design leverages the LLM agent’s strong executing and interpreting capabilities while relieving it of long-horizon planning and reasoning, which are handled by the classical planner.

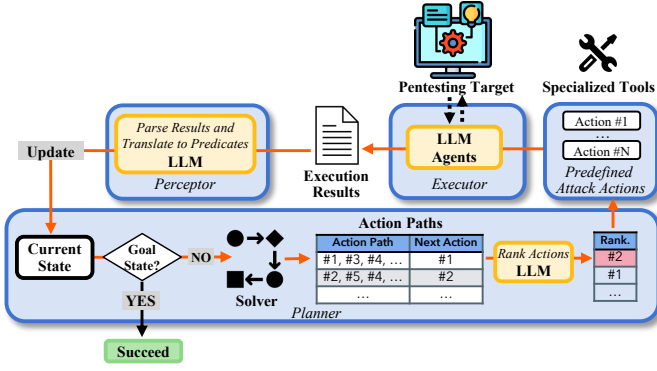


Fig. 2: Overview of CHECKMATE. The orange arrow shows the iterative loop of classical planning+. The current state is initialized before the planning starts.

B. Predefined Attack Actions

As previously mentioned, existing general-purpose LLM agents lack knowledge of specialized tools during pentesting. To address this, we introduce predefined attack actions to expand their knowledge base. We explicitly predefine niche and fine-grained tools such as Metasploit modules, NSE scripts, and Nuclei templates as “actions”, which are considered by the planner. Predefined attack actions also help avoid the inconsistency and errors in LLM-command-generation. In pentesting, most commands adhere to a consistent structure. For example, when executing a default port scan using `nmap -Pn -sC -sV -p- -oN - #{target}`, the command structure remains largely consistent, while the only part that usually changes is `#{target}`. However, the next-token prediction mechanism of LLMs is increasingly unstable, and error-prone when generating long commands. In contrast, predefined attack actions provide the core structure and options of the command, leaving only parameters like `#{target}` to be specified, significantly reducing the risk of generating incorrect commands.

Predefined attack action offers an alternative approach to expanding an LLM’s knowledge base without relying on traditional RAG or fine-tuning. Fine-tuning LLMs is often costly, time-consuming, and difficult to scale. RAG, while flexible, depends on retrieving document snippets and the model’s ability to interpret those snippets and synthesize commands. In contrast, predefined attack actions offer explicit, well-structured, executable commands. By defining preconditions of actions (we will introduce this later), these actions can be retrieved more accurately, efficiently, and interpretably than relying solely on embedding-based similarity search in RAG.

C. Planner

We propose classical planning+ as the planner of CHECKMATE, which encodes the causal relationships explicitly and maintains a persistent and coherent plan throughout the pentesting.

1) *Causal Relationships*: Causal relationships are encoded explicitly through the preconditions and effects associated with

each action. For example, once a web enumeration action identifies a specific web application, a pentester would naturally consider all relevant Metasploit modules, NSE scripts, and Nuclei templates associated with that application. In classical planning, the discovered web application is an effect of the enumeration action and simultaneously serves as a precondition of those subsequent actions. The set of factors used as preconditions and effects is flexible and can be customized or extended as needed. In our current design, these factors include elements such as the identified application, CVEs, URLs, usernames, passwords, etc. By encoding these causal dependencies directly, the system gains stronger built-in causal reasoning capabilities and reduces the need for the LLM to perform complex long-horizon reasoning on its own.

2) *Classical Planning+*: Classical planning+ is proposed to address the limitations of traditional classical planning in dynamic, non-deterministic, and partially observable tasks.

Non-Deterministic Action Effects: Pentesting inherently involves uncertainty and incomplete information. For instance, the result of a port scan is not known until finished, and the outcome of an exploit attempt is often unpredictable until it is executed. However, traditional classical planning assumes a static, deterministic, and fully observable target, where all action effects are determined, and the state of the target is completely specified before the planning starts. Some pentesting systems use complex models to encode uncertainty, which are difficult to scale in the real world. In CHECKMATE, we propose classical planning+, leveraging LLMs to dynamically determine action effects. Since it updates action effects at runtime, complete knowledge is no longer required before planning begins. Specifically, we define the non-deterministic effect to indicate that the effect of an action is unknown until it is executed. Once an action with a non-deterministic effect is executed, LLMs are invoked to analyze the execution outcome and generate concrete effect predicates. We describe this process in §IV-C2, along with a concrete example. Through this mechanism, we successfully extend classical planning to dynamic, non-deterministic, partially-observable scenarios.

Classical planning+ begins from the initial state representing all prior knowledge about the target. In each iteration, the planner checks whether the goal is reachable under the current state. If it is, the planner executes the action sequence leading to the goal. If not, the planner produces a list of applicable actions by checking the preconditions of each action. Next, the LLM executes the optimal action from this list based on its knowledge and updates the initial state. If the executed action has a non-deterministic effect, the LLM is invoked to analyze the execution output and translate it into concrete predicates. This process is repeated iteratively until either the goal is met or all possible actions have been explored. Compared to LLM agents, classical planning+ provides a more structured planning engine by presenting a plan as a directed acyclic graph. It offers several advantages in pentesting planning. First, it exhaustively explores the entire action space, avoiding missing available actions, especially in scenarios with a large number of actions or long action sequences. Second,

it avoids repeating previously executed actions or jumping across different directions, which is a common failure in LLM-based planning. Moreover, the planning process is both visible and interpretable. We illustrate classical planning+ using an example in Figure 3.

Algorithm 1 Iterative Planning for Penetration Testing under Partial Knowledge

```

1: Input: Domain  $D$  with predefined actions, initial knowledge  $I_0$ 
2: Initialize:  $S \leftarrow I_0$  ▷ Initial state from known information (e.g., IP)
3: while termination condition not met and actions remain do
4:   applicableActions  $\leftarrow \{\}$ 
5:   for all action  $a$  in domain  $D$  do
6:     if  $a$  is reachable from state  $S$  then
7:       seq  $\leftarrow \text{plan}(S, a)$ 
8:       applicableActions.add(seq.first())
9:     end if
10:  end for
11:  nextAction  $\leftarrow \text{LLM\_Select}(\text{applicableActions})$ 
12:  result  $\leftarrow \text{Execute}(\text{nextAction})$ 
13:  if nextAction has deterministic effects then
14:     $S \leftarrow S \cup \text{effects}(\text{nextAction})$ 
15:  else
16:    preds  $\leftarrow \text{Parse\_NonDeterministic\_Effects}(\text{result})$ 
17:     $S \leftarrow S \cup \text{preds}$ 
18:  end if
19: end while
20: if goal is not achieved then
21:   Report failure: challenge unsolvable.
22: end if

```

D. Executor

Once the next action is determined, the system should execute it without human intervention. To do so, it must select the appropriate tools, reliably generate precise, executable instructions, and configure all required parameters. Given the strong execution abilities of LLM-based code agents, CHECKMATE employs an LLM agent as its executor. Each predefined action is paired with a concise, action-specific prompt that guides the agent. These prompts specify the required tool and command structure, along with placeholders for parameters. For instance, for a network-scanning action, the prompt clearly outlines the expected flags and arguments, while still allowing task-specific values (the IP address) to be injected. These placeholders are automatically populated by the classical planner, ensuring that critical parameters, such as a module name for an exploitation step, are determined deterministically rather than by the LLM, thereby reducing the risk of hallucinations. After the planner selects an action, CHECKMATE provides the corresponding prompt to the LLM executor, which performs the command and returns the resulting output for downstream processing.

E. Perceptor

The perceptor bridges the executor and the planner: it analyzes the execution results in heterogeneous formats and content, translates them into the representation that the planner can use for subsequent planning. In CHECKMATE, the perceptor translates the outputs into predicates defined in classical planning+, which are then used to update the current state. CHECKMATE has two types of perceptrors: a rule-based perceptor and an LLM-based perceptor. The rule-based perceptor parses structured outputs and maps them to the corresponding predicates, avoiding the randomness introduced when using LLMs. For example, the JSON result returned from a Metasploit search can be directly mapped to a predicate (*msf-module-available ?exploit-name*) for simplicity. The LLM-based perceptor leverages LLMs to interpret unstructured outputs and produce predicates defined in classical planning+.

V. EVALUATION

A. Penetration Capability

We first evaluate the penetration capability of CHECKMATE, compared to existing work. We adopt the same benchmark dataset, metrics, and experimental setup as described in Section 3. The results are shown in Figure 4. CHECKMATE demonstrates substantially stronger penetration capability than all baselines, as evidenced by its progress across all milestones. Notably, 88% of its penetration attempts reach milestone M7, whereas prior work, except Claude Code, rarely progresses beyond M4. Furthermore, CHECKMATE shows advantages over Claude Code at the higher milestones, particularly in the success rates for M6 and M7, indicating improved effectiveness in executing exploits and successfully obtaining a shell. These gains result from the explicitly defined, fine-grained actions and a structured planning strategy. By planning all available actions before committing to any specific attack path, CHECKMATE avoids becoming trapped in unproductive branches and maintains steady progress toward deeper system compromise.

B. Efficiency

In this section, we evaluate both the efficiency and cost of CHECKMATE. We selected 20 penetration tasks that CHECKMATE and Claude Code were both able to successfully complete. Under the same LLM model setting, we compared the total monetary cost, representing the amount of LLM tokens consumed, and the time required to finish each task. The results are summarized in Figure 5. On average, CheckMate has a total cost of \$0.68, which is 53% lower than that of Claude Code under identical conditions. This reduction in token consumption can be attributed to the use of classical planning for strategy formulation. In contrast, Claude Code relies entirely on text-based reasoning, where every intermediate thought and plan must be expressed in natural language, leading to substantial token overhead. By adopting a symbolic and formalized planning mechanism, CHECKMATE avoids using the LLM to “generate” its reasoning process, thereby concentrating the model’s generation capacity on executing

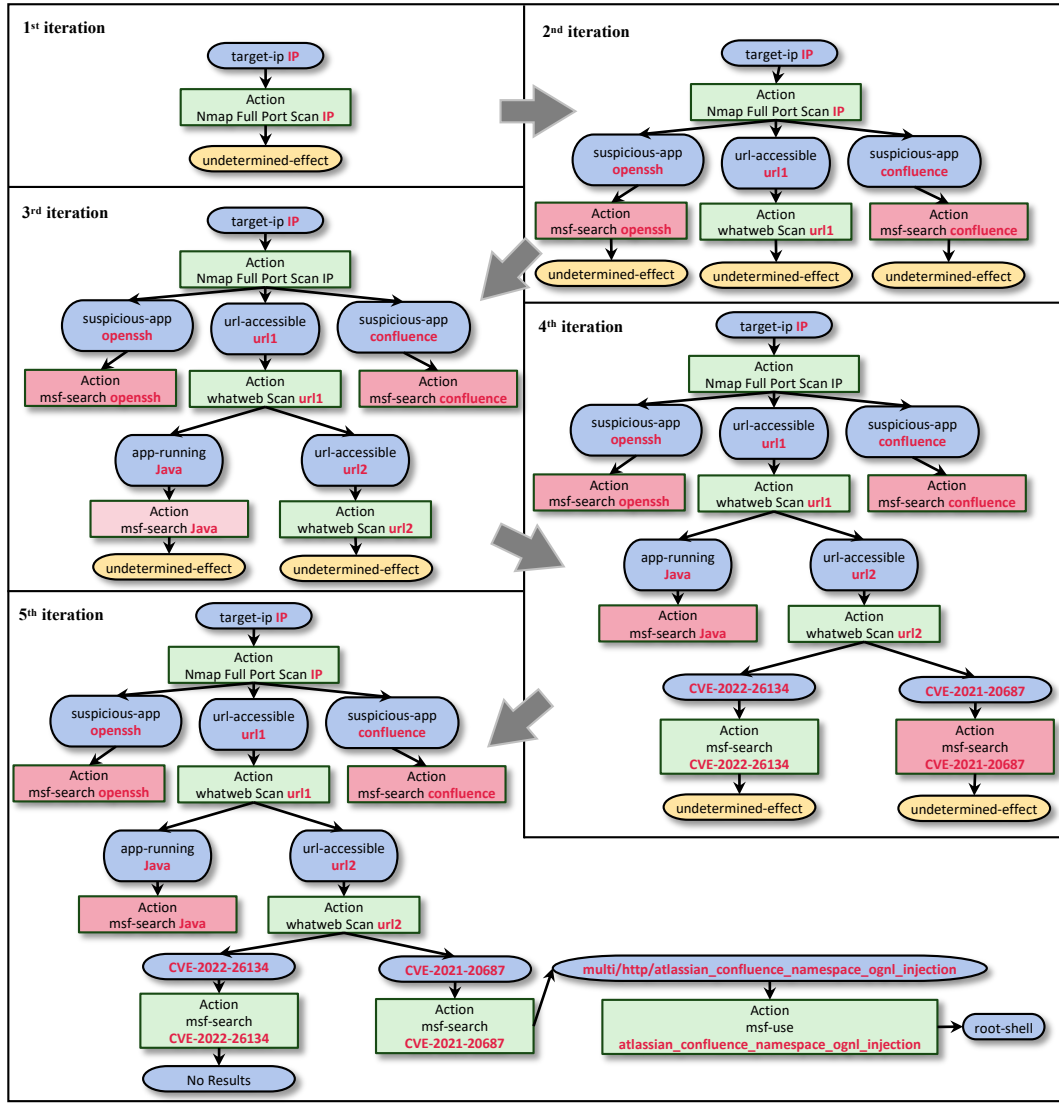


Fig. 3: A pentesting workflow driven by classical planning+. Each panel shows one planning-execution-perception iteration. Blue rounded ovals are predicates that link actions across iterations; yellow rounded ovals denote non-deterministic action effects. Rectangular boxes list feasible actions available during the engagement, and light-green rectangles indicate the actions chosen by the planner for execution in that iteration. Arrows show how actions are connected with predicates.

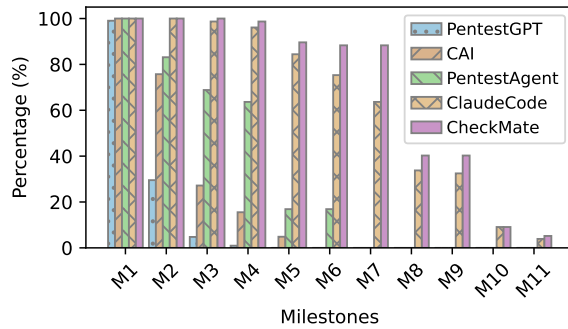
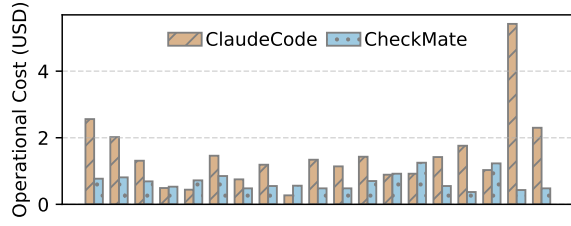


Fig. 4: Comparison of Claude Code with CHECKMATE on Vulhub benchmark

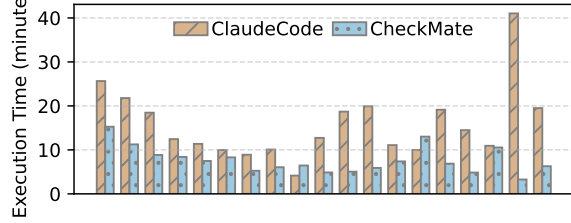
actions and interpreting outputs. The average time consumed for CHECKMATE is 7.75 minutes, which is 54% lower than Claude Code.

C. Stability

In this section, we evaluate the stability of the pentesting process, i.e., whether the system demonstrates consistent performance across repeated executions of the same task. To assess this, we execute each task three times and record the results, costs, and time consumption for each run. We then compute the success rate (i.e., the proportion of runs in which all three attempts successfully achieve penetration) and the *Coefficient of Variation* (a scale-independent measure of dispersion) of both cost and time. The aggregated results are summarized in Table II. About 25% of the tasks



(a) Monetary costs.



(b) Time consumed.

Fig. 5: Efficiency comparison between ClaudeCode and CHECKMATE.

cannot be solved consistently by Claude Code across all three attempts. In addition, CHECKMATE demonstrates higher consistency in both LLM token usage and execution time. These improvements stem from CHECKMATE’s adoption of a more structured planning engine, which reduces unnecessary fluctuations introduced by the LLM.

TABLE II: Stability comparison between CHECKMATE and Claude Code.

	CHECKMATE	Claude Code
Success Rate for all Attempts (↑)	100%	75%
Coefficient of Variation - Cost (↓)	0.129	0.451
Coefficient of Variation - Time (↓)	0.093	0.325

D. Case Study

We analyze a specific example in detail to illustrate differences during pentesting between CHECKMATE and Claude Code. In this case, CHECKMATE completed the penetration in only three steps; Claude Code, by contrast, used 26 steps, many of which were added because of redundancy, premature abandonment, distractions, and incomplete planning and reasoning. The target is an old version of Apache ActiveMQ (an open-source messaging middleware that supports Java messaging services, clustering, and the Spring framework) from Vulhub. CHECKMATE began with a full-port Nmap scan plus fingerprinting and script probes. It discovered two open ports (22 and 8191), identified that Apache ActiveMQ was running, and associated that service with likely CVEs and corresponding Metasploit modules. Rather than rushing straight to exploitation, CHECKMATE chose, from the feasible action set, to analyze the web interface to further confirm the ActiveMQ version. That analysis verified that an ActiveMQ Console was running and revealed the

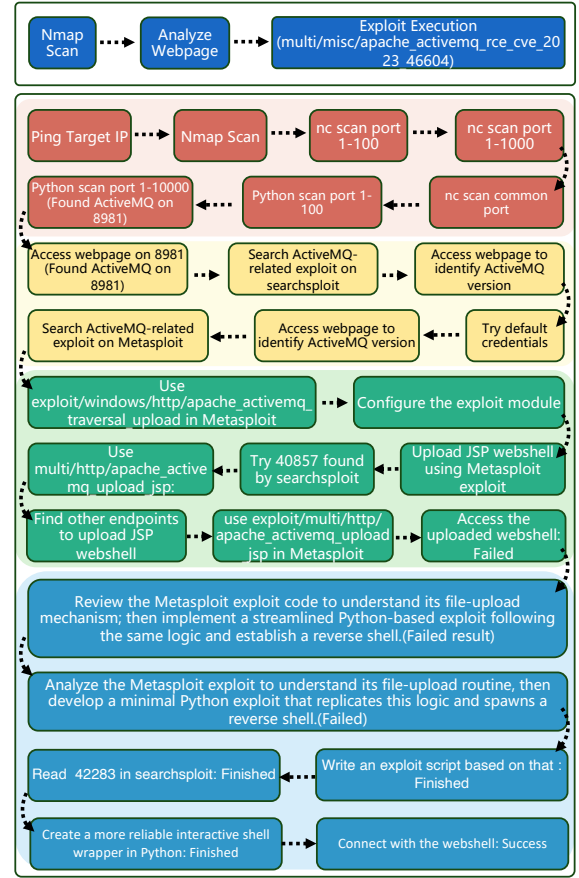


Fig. 6: Top box: CHECKMATE’s workflow. Bottom box: Claude Code’s workflow. Colors show stages. pink: reconnaissance, yellow: search/analysis, green: Metasploit/SearchSploit exploitation (failed), blue: autonomous exploitation.

precise version, 5.11.1. Armed with that version information, in the third step, CHECKMATE selected Metasploit’s “multi/misc/apache_activemq_rce_cve_2023_46604” module and ran the exploit, ultimately obtaining a root shell on the target. At each stage, the LLM evaluated all available actions and prioritized them, demonstrating a highly planned and systematic approach.

By contrast, Claude Code’s test was a largely ad-hoc process, showing exploratory and blind trial-and-error behavior. Claude Code first tried to ping the target IP and run nmap scans, but abandoned both because of insufficient permissions. In fact, both commands failed due to missing socket privileges and would have succeeded with “sudo”. Instead of modifying the commands to add the necessary privileges, Claude Code pivoted to using Netcat and writing Python scripts, which are more complex methods for port scanning. Its port-scanning also lacked a coherent plan: it scanned the first 100 ports, then 1,000 ports, then “common” ports, and only later broadened the range, thus finally finding port 8191 open. It is a risky strategy because when Claude Code hit a rabbit hole on common ports, it tended to pursue the wrong path, wasting

time or risking failure. Claude Code also struggled to remain focused on a single attack path. While attempting to determine the ActiveMQ version, it would abruptly switch to trying the default-credential brute force. After selecting and spending a long time configuring a Metasploit module, it might suddenly divert to investigating another script found on Exploit-DB, creating needless context switches and time loss. Finally, because Claude Code lacked explicit, structured reasoning, it failed to map the discovered ActiveMQ version to the most appropriate CVE. As a result, it missed the more effective Metasploit module and wasted excessive time on two suboptimal exploits.

E. Ablation Study

In this section, we conduct an ablation study by comparing CHECKMATE with two commonly used strategies for enhancing LLM-based systems. First, we compare CHECKMATE against the RAG-based approach, which is an alternative strategy for expanding an LLM’s knowledge base. We embedded metadata of specialized penetration tools, including more than 14 thousand Metasploit modules, NSE scripts, and Nuclei templates, as the document database and implemented a RAG pipeline. We aim to evaluate whether LLM agents can effectively use external knowledge to improve their penetration capabilities without relying on predefined actions and classical planning+. Second, we let Claude Code maintain a structured planning file in JSON format rather than using its default to-do list. The prompting was modified so that after each command execution, Claude Code updates this structured planning file and infers the next step based on the revised state. This approach reflects common methodologies in prior work that employ structured planning representations to improve an LLM agent’s planning consistency. For each method, we evaluated the performance on 20 tasks, running each task three times. All four methods successfully obtained a remote shell at least once. As shown in Figure 7, CHECKMATE achieves the lowest overall cost and the shortest execution time, while also delivering the most consistent and efficient performance across test cases. These results indicate that although incorporating RAG or structured planning files can enhance the efficiency of LLM-based agents, the classical planning+ approach provides the most substantial gains in both efficiency and consistency.

VI. DISCUSSION AND FUTURE WORK

A. Actions and States in Pentesting

Existing work leaves two fundamental questions unanswered: (1) What actions and skills does pentesting require? and (2) How should we represent the state of the target system? The difficulty arises from the open-ended nature of pentesting. Unlike tasks with well-defined action and state spaces, pentesting spans the full breadth of a system’s architecture, configurations, vulnerabilities, and defenses, and demands a wide and adaptable skill set. Current approaches either define fixed, finite sets of skills and states, or depend heavily on black-box LLMs to infer target states and propose actions. The fixed schemas are too restrictive, while relying on opaque LLMs makes it difficult to systematically improve penetration

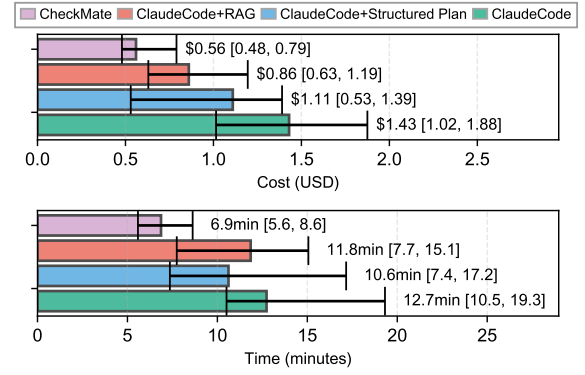


Fig. 7: Cost and time comparison. (a) Median API costs in USD. (b) Median execution time in minutes. Error bars represent the interquartile range (25th-75th percentile).

capabilities. This gap highlights the need for future work on representing, organizing, extracting, and operationalizing the fragmented knowledge on actions and states in pentesting.

B. Multimodal and UI-Aware Pentesting

Existing pentesting systems struggle in scenarios that require rich human-computer interaction, as traditional LLM agents are not good at interpreting non-textual information and operating web user interfaces (UIs) like a human. As a result, tasks that involve understanding visual elements or manipulating dynamic, interactive web components still depend heavily on humans. Recent advances in multimodal learning and Customizable UI Automation (CUA) offer promising avenues for addressing these limitations [36], [37], opening up new possibilities for pentesting in complex UI environments.

VII. CONCLUSION

In this paper, we first presented a systematic review of existing automated pentesting work through the lens of our Planner-Executor-Perceptor (PEP) paradigm. Our evaluation shows that the out-of-the-box Claude Code+Sonnet 4.5 substantially outperforms all prior systems in this area. However, further analysis revealed three limitations of Claude Code. We thus proposed CHECKMATE, a framework that couples classical planning+ with LLM agents to address these limitations. Experimental evaluations demonstrated that CHECKMATE outperforms existing systems in penetration capability, efficiency, and stability.

ETHICAL CONSIDERATIONS

This paper presents a practical study on using LLM Agents for pentesting. All techniques and systems involved are publicly accessible; we have not developed any new zero-day attacks. All experiments were conducted within authorized virtual environments. We will contact the service providers to inform them of the potential for their products in offensive scenarios. This work is intended solely for research and educational purposes, and we do not encourage or endorse any misuse of the discussed techniques.

REFERENCES

- [1] Cybersecurity and Infrastructure Security Agency, “Penetration testing services,” <https://www.cisa.gov/resources-tools/services/penetration-testing>, 2023, u.S. Department of Homeland Security.
- [2] MarketsandMarkets, “Penetration testing market size, size, growth & latest trends,” <https://www.marketsandmarkets.com/Market-Reports/penetration-testing-market-13422019.html>, 2024, accessed: 2025-10-03.
- [3] —, “Penetration testing as a service market size & share analysis – global forecast to 2029,” <https://www.marketsandmarkets.com/Market-Reports/penetration-testing-as-a-service-market-36245315.html>, 2024, accessed: 2025-10-03.
- [4] V. Mayoral-Vilches, L. J. Navarrete-Lozano, M. Sanz-Gómez, L. S. Espejo, M. Crespo-Álvarez, F. Oca-Gonzalez, F. Balassone, A. Glerapicón, U. Ayucar-Carbajo, J. A. Ruiz-Alcalde, S. Rass, M. Pinzger, and E. Gil-Uriarte, “Cai: An open, bug bounty-ready cybersecurity ai,” 2025. [Online]. Available: <https://arxiv.org/abs/2504.06017>
- [5] B. Wu, G. Chen, K. Chen, X. Shang, J. Han, Y. He, W. Zhang, and N. Yu, “Autopt: How far are we from the endend automated web penetration testing?” *arXiv preprint arXiv:2411.01236*, 2024.
- [6] C. Sarraute, O. Buffet, and J. Hoffmann, “Pomdp make better hackers: Accounting for uncertainty in penetration testing,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, 2012, pp. 1816–1824.
- [7] J. Hoffmann, “Simulated penetration testing: from” dijkstra” to” turing test++,” in *Proceedings of the international conference on automated planning and scheduling*, vol. 25, 2015, pp. 364–372.
- [8] I. Mirzadeh, K. Alizadeh, M. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, “Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models,” *arXiv preprint arXiv:2410.05229*, 2024.
- [9] B. Y. Lin, R. L. Bras, K. Richardson, A. Sabharwal, R. Poovendran, P. Clark, and Y. Choi, “Zeblogic: On the scaling limits of llms for logical reasoning,” *arXiv preprint arXiv:2502.01100*, 2025.
- [10] K. Yamin, S. Gupta, G. R. Ghosal, Z. C. Lipton, and B. Wilder, “Failure modes of llms for causal reasoning on narratives,” *arXiv preprint arXiv:2410.23884*, 2024.
- [11] H. Chi, H. Li, W. Yang, F. Liu, L. Lan, X. Ren, T. Liu, and B. Han, “Unveiling causal reasoning in large language models: Reality or mirage?” *Advances in Neural Information Processing Systems*, vol. 37, pp. 96 640–96 670, 2024.
- [12] J. Liu, K. Wang, Y. Chen, X. Peng, Z. Chen, L. Zhang, and Y. Lou, “Large language model-based agents for software engineering: A survey,” *arXiv preprint arXiv:2409.02977*, 2024.
- [13] H. Jin, L. Huang, H. Cai, J. Yan, B. Li, and H. Chen, “From llms to llm-based agents for software engineering: A survey of current, challenges and future,” *arXiv preprint arXiv:2408.02479*, 2024.
- [14] Y. Wang, W. Zhong, Y. Huang, E. Shi, M. Yang, J. Chen, H. Li, Y. Ma, Q. Wang, and Z. Zheng, “Agents in software engineering: Survey, landscape, and vision,” *Automated Software Engineering*, vol. 32, no. 2, pp. 1–36, 2025.
- [15] S. Ullah, M. Han, S. Pujar, H. Pearce, A. Coskun, and G. Stringhini, “Llms cannot reliably identify and reason about security vulnerabilities (yet?): A comprehensive evaluation, framework, and benchmarks,” in *2024 IEEE symposium on security and privacy (SP)*. IEEE, 2024, pp. 862–880.
- [16] J. Guo, C. Wang, X. Xu, Z. Su, and X. Zhang, “Repoaudit: An autonomous llm-agent for repository-level code auditing,” *arXiv preprint arXiv:2501.18160*, 2025.
- [17] M. Rahman, A. Bhuiyan, M. S. Islam, M. T. R. Laskar, R. Mahbub, A. Masry, S. Joty, and E. Hoque, “Llm-based data science agents: A survey of capabilities, challenges, and future directions,” *arXiv preprint arXiv:2510.04023*, 2025.
- [18] Anthropic, “Claude code,” <https://www.claude.com/product/claude-code>, 2025, accessed: 2025-11-07.
- [19] M. Ghallab, D. Nau, and P. Traverso, *Automated Planning: theory and practice*. Elsevier, 2004.
- [20] “Vulhub: Open-source vulnerable docker environments,” <https://vulhub.org/>, accessed: 2025-12-04.
- [21] J. Schwartz, H. Kurniawati, and E. El-Mahassni, “Pomdp+ information-decay: Incorporating defender’s behaviour in autonomous penetration testing,” in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 30, 2020, pp. 235–243.
- [22] C. Sarraute, O. Buffet, and J. Hoffmann, “Penetration testing== pomdp solving?” *arXiv preprint arXiv:1306.4714*, 2013.
- [23] J. Schwartz and H. Kurniawati, “Autonomous penetration testing using reinforcement learning,” *arXiv preprint arXiv:1905.05965*, 2019.
- [24] M. C. Ghanem, T. M. Chen, and E. G. Nepomuceno, “Hierarchical reinforcement learning for efficient and effective automated penetration testing of large networks,” *Journal of Intelligent Information Systems*, vol. 60, no. 2, pp. 281–303, 2023.
- [25] S. Zhou, J. Liu, D. Hou, X. Zhong, and Y. Zhang, “Autonomous penetration testing based on improved deep q-network,” *Applied Sciences*, vol. 11, no. 19, p. 8823, 2021.
- [26] G. D. Pasquale, I. Grishchenko, R. Iesari, G. Pizarro, L. Cavallaro, C. Kruegel, and G. Vigna, “ChainReactor: Automated privilege escalation chain discovery via AI planning,” in *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 5913–5929. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity24/presentation/de-pasquale>
- [27] J. L. Obes, C. Sarraute, and G. Richarte, “Attack planning in the real world,” *arXiv preprint arXiv:1306.4044*, 2013.
- [28] Z. Chen, F. Kang, X. Xiong, and H. Shu, “A survey on penetration path planning in automated penetration testing,” *Applied Sciences*, vol. 14, no. 18, p. 8355, 2024.
- [29] Z. Wang, Y. Zhang, Z. Liu, X. Wei, Y. Chen, and B. Wang, “An automatic planning-based attack path discovery approach from it to ot networks,” *Security and Communication Networks*, vol. 2021, no. 1, p. 1444182, 2021.
- [30] G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, and S. Rass, “{PentestGPT}: Evaluating and harnessing large language models for automated penetration testing,” in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 847–864.
- [31] Armur-Ai, “Auto-pentest-gpt-ai: Llm powered pentesting for your software,” <https://github.com/Armur-Ai/Auto-Pentest-GPT-AI>, 2025, accessed: 2025-10-16.
- [32] J. Huang and Q. Zhu, “Penheal: A two-stage llm framework for automated pentesting and optimal remediation,” in *Proceedings of the workshop on autonomous cybersecurity*, 2023, pp. 11–22.
- [33] J. Xu, J. W. Stokes, G. McDonald, X. Bai, D. Marshall, S. Wang, A. Swaminathan, and Z. Li, “Autoattacker: A large language model guided system to implement automatic cyber-attacks,” *arXiv preprint arXiv:2403.01038*, 2024.
- [34] H. Kong, D. Hu, J. Ge, L. Li, T. Li, and B. Wu, “Vulnbot: Autonomous penetration testing for a multi-agent collaborative framework,” *arXiv preprint arXiv:2501.13411*, 2025.
- [35] GH05TCREW, “Pentestagent: All-in-one offensive security toolbox with ai agent and mcp architecture,” <https://github.com/GH05TCREW/PentestAgent>, 2025, accessed: 2025-10-16.
- [36] X. Wang, B. Wang, D. Lu, J. Yang, T. Xie, J. Wang, J. Deng, X. Guo, Y. Xu, C. H. Wu *et al.*, “Opencua: Open foundations for computer-use agents,” *arXiv preprint arXiv:2508.09123*, 2025.
- [37] Y. Yang, D. Li, Y. Dai, Y. Yang, Z. Luo, Z. Zhao, Z. Hu, J. Huang, A. Saha, Z. Chen *et al.*, “Gta1: Gui test-time scaling agent,” *arXiv preprint arXiv:2507.05791*, 2025.
- [38] A. L. Blum and M. L. Furst, “Fast planning through planning graph analysis,” *Artificial intelligence*, vol. 90, no. 1-2, pp. 281–300, 1997.
- [39] S. Kambhampati, K. Valmeekam, L. Guan, M. Verma, K. Stechly, S. Bhambri, L. Saldyt, and A. Murthy, “Llms can’t plan, but can help planning in llm-modulo frameworks,” *arXiv preprint arXiv:2402.01817*, 2024.
- [40] P. Cao, T. Men, W. Liu, J. Zhang, X. Li, X. Lin, D. Sui, Y. Cao, K. Liu, and J. Zhao, “Large language models for planning: A comprehensive and systematic survey,” *arXiv preprint arXiv:2505.19683*, 2025.
- [41] H. Zhang, H. Deng, J. Ou, and C. Feng, “Mitigating spatial hallucination in large language models for path planning via prompt engineering,” *Scientific Reports*, vol. 15, no. 1, p. 8881, 2025.
- [42] Z. Ji, D. Wu, P. Ma, Z. Li, and S. Wang, “Testing and understanding erroneous planning in llm agents through synthesized user inputs,” *arXiv preprint arXiv:2404.17833*, 2024.
- [43] Z. Yao, Y. Liu, Y. Chen, J. Chen, J. Fang, L. Hou, J. Li, and T.-S. Chua, “Are reasoning models more prone to hallucination?” *arXiv preprint arXiv:2505.23646*, 2025.

- [44] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, “Lost in the middle: How language models use long contexts,” *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.
- [45] X. Shen, L. Wang, Z. Li, Y. Chen, W. Zhao, D. Sun, J. Wang, and W. Ruan, “Pentestagent: Incorporating llm agents to automated penetration testing,” in *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, 2025, pp. 375–391.
- [46] “picoCTF,” <https://picoctf.org/>, accessed: 2025-12-04.
- [47] “Hack The Box,” <https://www.hackthebox.com/>, accessed: 2025-12-04.
- [48] M. Shao, S. Jancheska, M. Udeshi, B. Dolan-Gavitt, K. Milner, B. Chen, M. Yin, S. Garg, P. Krishnamurthy, F. Khorrami *et al.*, “Nyu ctf bench: A scalable open-source benchmark dataset for evaluating llms in offensive security,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 57 472–57 498, 2024.
- [49] Y. Ginige, A. Niroshan, S. Jain, and S. Seneviratne, “Autopentester: An llm agent-based framework for automated pentesting,” *arXiv preprint arXiv:2510.05605*, 2025.
- [50] “Xbow: Ai-powered penetration testing platform,” <https://xbow.com/>, XBOW USA Inc., 2025, accessed: 2025-10-16.
- [51] 0x4m4, “Hexstrike ai mcp agents,” <https://github.com/0x4m4/hexstrike-ai>, 2025, accessed: 2025-10-16.
- [52] OpenAI, “Openai codex,” <https://openai.com/codex/>, 2025, accessed: 2025-11-07.
- [53] Google, “Code assist,” <https://codeassist.google/>, 2025, accessed: 2025-11-07.