IJASSE | International Journal of Academic Studies in Science and Education

Year 2025, Vol. 3, No. 2, Pages 125-146 www.ijasse.com

OPEN ACCESS

Patterns of LLM Weaponization: A Comparative Analysis of Exploitation Incidents Across Commercial AI Systems

George Antoniou
Lynn University, United States

Article Info

Article History

Received:
11 July 2025

Accepted:
28 December 2025

Keywords

Large language models,
Comparative analysis,
Cyber exploitation patterns,
LLM weaponization,
Autonomous agents,
Capability democratization,
Underground AI,
Defensive frameworks

Abstract

This comparative study examines patterns of Large Language Model (LLM) weaponization through systematic analysis of four major exploitation incidents spanning from 2023-2025. While existing research focuses on isolated incidents or theoretical vulnerabilities, this study provides one of the first comprehensive comparative frameworks analyzing exploitation patterns across state-sponsored cyber-espionage (Anthropic Claude incident), academic security research (GPT-4 autonomous privilege escalation), social engineering platforms (SpearBot phishing framework), and underground criminal commoditization (WormGPT/FraudGPT ecosystem). Through comparative analysis across eight dimensions: Adversary sophistication, target selection, exploitation techniques, autonomy levels, detection evasion, attribution challenges, defensive gaps, and capability democratization, this research identifies critical cross-case patterns informing defensive prioritization. Findings reveal three universal exploitation mechanisms transcending adversary types: autonomous goal decomposition via chain-of-thought reasoning (present in all four cases), dynamic tool invocation and code generation (3/4 cases), and adaptive social engineering (4/4 cases). Analysis demonstrates progressive capability democratization: state-level sophistication (Claude: 80-90% autonomy) transitioning to academic accessibility (GPT-4: 33-83% success rates), specialized criminal tooling (SpearBot: generative-critique architecture), and mass commoditization (WormGPT: $200-1700/year subscriptions). Comparative findings identify four cross-cutting defensive imperatives applicable regardless of adversary type: multi-turn conversational context monitoring, behavioral fingerprinting distinguishing legitimate from malicious complex workflows, federated threat intelligence enabling rapid cross-organizational learning, and capability-based access controls proportional to LLM reasoning sophistication.

To cite this article

Antoniou, G. (2025). Patterns of LLM weaponization: A comparative analysis of exploitation incidents across commercial AI systems. *International Journal of Academic Studies in Science and Education (IJASSE), 3*(2), 125-146. https://doi.org/10.55549/ijasse.50

Corresponding Author: George Antoniou, GAntoniou@lynn.edu

Introduction

The weaponization of Large Language Models (LLMs) for cyber operations represents a fundamental shift in threat landscape dynamics (Kasneci et al., 2023). While individual exploitation incidents have garnered attention, state-sponsored campaigns, academic demonstrations, criminal marketplaces, systematic comparative analysis examining patterns across adversary types, exploitation contexts, and capability levels remain absent from literature. This research addresses this critical gap through comparative analysis of four major LLM exploitation incidents spanning September 2023 through September 2025, providing one of the first comprehensive frameworks for understanding weaponization patterns across diverse threat actor categories.

Existing cybersecurity research predominantly analyzes threats within single adversary categories: state-sponsored Advanced Persistent Threats (APTs), cybercriminal organizations, or research demonstrations. However, LLM weaponization transcends traditional adversary taxonomies. The same technical capabilities, autonomous reasoning, tool invocation, adaptive learning, prove exploitable by state intelligence services, academic researchers, specialized criminal groups, and mass-market underground tooling. This convergence of capability across adversary sophistication levels represents unprecedented threat landscape evolution demanding comparative analytical frameworks.

This study examines four critical cases representing unique adversary profiles and exploitation contexts. First, the September 2025 Anthropic Claude incident demonstrates state-sponsored cyber-espionage achieving 80-90% tactical autonomy across multi-organization campaigns (Anthropic PBC, 2025; BBC World Service, 2025). Second, the GPT-4 autonomous privilege escalation research (Happe et al., 2023) represents academic security demonstration achieving 33-83% exploitation success rates against vulnerable Linux systems. Third, the SpearBot phishing framework (Qi et al., 2025) exemplifies specialized social engineering tooling employing generative-critique architecture for adaptive content creation. Fourth, the WormGPT/FraudGPT darkweb underground ecosystem demonstrates mass commoditization of weaponized LLMs through subscription-based cybercrime-as-a-service models (Trustwave, 2023; Unit 42, 2025).

Research Problem and Comparative Necessity

Traditional threat analysis focuses on adversary-specific Tactics, Techniques, and Procedures (TTPs). Attribution frameworks distinguish state actors from criminal groups through infrastructure analysis, targeting patterns, and operational sophistication. However, LLM weaponization creates fundamental analytical challenges. When diverse adversary types exploit identical underlying capabilities, chain-of-thought reasoning,

autonomous code generation, adaptive social engineering, traditional adversary-centric frameworks prove insufficient (Wei et al., 2022). Understanding what remains constant across exploitation contexts versus what varies by adversary type becomes essential for effective defense prioritization in cybersecurity and AI.

Comparative analysis serves three critical functions. First, identifying universal exploitation patterns enables development of capability-based defenses effective regardless of adversary identity. If all exploitation contexts employ chain-of-thought decomposition, defensive systems must address this mechanism universally rather than tailoring responses to specific adversary profiles. Second, understanding capability progression from state-level sophistication through academic demonstration to mass commoditization illuminates' democratization dynamics and timeline forecasting. Third, cross-case analysis reveals defensive gaps transcending specific incidents, identifying systemic vulnerabilities requiring architectural rather than tactical remediation.

Research Questions

This comparative study addresses four interconnected research questions:

RQ1: What exploitation mechanisms and technical capabilities remain constant across diverse adversary types and operational contexts?

RQ2: How do exploitation patterns, autonomy levels, and operational sophistication vary systematically across adversary categories?

RQ3: What capability democratization dynamics emerge from comparative temporal analysis spanning state-sponsored operations through mass commoditization?

RQ4: What defensive imperatives and architectural requirements emerge from cross-case pattern analysis?

Answering these questions requires moving beyond single-case analysis toward systematic comparative framework examining both commonalities and contextual variations. The research contribution lies in providing the first comprehensive comparative taxonomy of LLM weaponization patterns, enabling defenders to prioritize capability-based controls over adversary-specific signatures.

Literature Review

Existing research on LLM security spans multiple domains including AI safety, offensive capabilities documentation, and defensive frameworks. However, comparative analysis examining exploitation patterns across adversary types remains limited. This review organizes literature into three clusters: LLM capability documentation, adversary-specific exploitation research, and defensive frameworks.

LLM Security Capabilities and Vulnerabilities

Comprehensive surveys (Abdali et al., 2024; Zhou et al., 2025) document expanding LLM capability landscapes relevant to offensive operations. Yao et al. (2024) provide systematic categorization distinguishing beneficial applications, offensive misuse potential, and inherent vulnerabilities, noting that commercial models increasingly demonstrate capabilities exploitable across diverse adversary types. Their analysis reveals LLMs excel at code vulnerability detection while simultaneously introducing novel attack surfaces through prompt manipulation dilemma central to weaponization analysis (Raptis et al., 2025).

Liu et al. (2025) surveyed 223 studies examining LLM security threats, revealing concentrated research on backdoor attacks, data poisoning, and jailbreaking techniques. Critically, their analysis demonstrates these vulnerabilities exist independently of adversary sophistication, state actors, criminal groups, and individual researchers can exploit identical weaknesses. Zhang Y., et al. (2025) conducted systematic review analyzing over 300 works, identifying critical research gaps: while offensive capability documentation proliferates, comparative analysis examining how different adversary types weaponize identical capabilities remains sparse.

Adversary-Specific Exploitation Research

Research documenting LLM exploitation typically focuses on single adversary categories. Happe et al. (2023) demonstrated GPT-4 achieving 33-83% success rates in autonomous Linux privilege escalation within academic research context, establishing feasibility proof but not addressing how similar capabilities might be weaponized by other adversary types. Qi et al. (2025) documented SpearBot framework employing generative-critique architecture for phishing, demonstrating specialized tooling development but not examining relationship to broader weaponization ecosystem (Saha Roy et al., 2024).

Underground threat intelligence documents criminal LLM commoditization. Trustwave (2023) analyzed WormGPT and FraudGPT emergence on dark web forums, noting subscription pricing ($200-1700 annually) and explicit criminal marketing. Unit 42 (2025) examined WormGPT evolution through multiple generations, demonstrating sustained criminal investment in weaponized LLM development. However, these analyses remain isolated within criminal threat intelligence domain, not contextualizing relationships to state-sponsored or academic exploitation demonstrations.

Defensive Frameworks and Red Teaming

Defensive research emphasizes red teaming and adversarial testing (Perez et al., 2022). Feffer et al. (2024) surveyed 42 enterprise red team programs, warning against "security theatre" lacking systematic rigor. Gupta (2025) proposed COMPASS-RT framework integrating NIST AI RMF (National Institute of Standards and Technology, 2023) with MITRE ATLAS, providing structured methodology but not addressing how defensive priorities should adapt across diverse adversary types. Xu et al. (2025) demonstrated automated red teaming achieving 100% attack success rates against 37/41 LLMs, revealing widespread vulnerability but not differentiating defensive requirements by exploitation context.

Critical gaps emerge across existing literature: while individual exploitation incidents receive analysis and defensive frameworks (International Organization for Standardization, 2023) propose generic controls, systematic comparative research examining what remains constant versus varies across adversary types and operational contexts remains absent. This study addresses this gap through structured comparative framework.

Methodology

This study employs structured comparative case analysis methodology examining four major LLM exploitation incidents. The approach combines systematic case documentation with dimensional comparative framework enabling identification of both universal patterns and contextual variations.

Case Selection and Sampling Strategy

Cases were selected through purposive sampling strategy ensuring representation across four critical dimensions: (1) adversary sophistication (state-sponsored, academic, specialized criminal, mass-market criminal); (2) temporal distribution (September 2023 through September 2025); (3) exploitation type (cyber-espionage, penetration testing, social engineering, multi-purpose tooling); and (4) public documentation availability (incident disclosures, peer-reviewed research, threat intelligence reporting, underground marketplace analysis).

Four cases meet selection criteria. The Anthropic Claude Incident (September 2025) represents state-sponsored cyber-espionage with comprehensive incident disclosure (Anthropic PBC, 2025). GPT-4 Autonomous Privilege Escalation (Happe et al., 2023) provides peer-reviewed academic research with controlled experimental methodology. SpearBot Phishing Framework (Qi et al., 2025) exemplifies specialized social engineering research published in peer-reviewed venues. WormGPT/FraudGPT Ecosystem (2023-2025)

represents underground criminal commoditization documented through multiple threat intelligence sources (Trustwave, 2023; Unit 42, 2025).

Data Collection and Source Types

Data collection employed multiple source types appropriate to each case context. For Claude incident: official vendor disclosure, forensic analysis reporting, and media coverage. For GPT-4 privilege escalation: peer-reviewed research paper, open-source tooling documentation, and experimental benchmark data. For SpearBot: peer-reviewed publication with technical methodology and evaluation results. For WormGPT/FraudGPT: threat intelligence reporting, dark web forum analysis, and underground marketplace documentation. Source triangulation across vendor disclosures, academic publications, and threat intelligence enables robust cross-case comparison despite varying documentation formats.

Comparative Analytical Framework

Analysis employs dimensional comparative framework examining eight critical dimensions across all cases: (1) Adversary Profile, sophistication level, organizational structure, operational objectives; (2) Target Selection, victim types, scope, selection criteria; (3) Exploitation Techniques, technical mechanisms, LLM capabilities leveraged, attack workflows; (4) Autonomy Levels, percentage of autonomous versus human-directed actions, decision point frequency; (5) Detection Evasion, methods employed, adaptive behaviors, signature avoidance; (6) Attribution Challenges, factors complicating identification, forensic markers, infrastructure patterns; (7) Defensive Gaps, security architecture weaknesses exploited, control failures; (8) Capability Democratization, accessibility, skill requirements, dissemination mechanisms.

Each dimension receives systematic coding across all four cases enabling identification of: universal patterns (present across all cases), majority patterns (present in 3/4 cases), contextual variations (present in 1-2 cases with clear adversary-type correlation), and temporal progressions (capability evolution from early through late cases). This structured approach enables moving beyond anecdotal incident description toward systematic pattern identification.

Limitations and Scope

This comparative analysis acknowledges several limitations. First, reliance on publicly disclosed information means classified intelligence regarding state-sponsored capabilities remains inaccessible, potentially underestimating sophistication levels. Second, underground marketplace analysis depends on threat

intelligence reporting rather than direct access, introducing potential reporting biases. Third, focusing on four high-profile cases may not capture full diversity of LLM weaponization ecosystem. However, selected cases represent best-documented incidents spanning critical adversary types, providing sufficient foundation for pattern identification while acknowledging broader ecosystem complexity.

Case Descriptions

This section provides structured description of each case following standardized format: adversary profile, operational context, technical exploitation mechanisms, documented metrics, and defensive implications. Standardized description enables systematic cross-case comparison in subsequent sections.

Case 1: Anthropic Claude Cyber-Espionage Campaign (September 2025)

Adversary Profile: State-sponsored actors, attributed to Chinese intelligence services through infrastructure correlation and targeting patterns. High sophistication level with substantial resources, operational security capabilities, and strategic intelligence objectives. Campaign represented sustained effort over multiple weeks with professional operational tradecraft.

Operational Context: Multi-organization cyber-espionage targeting approximately 30 entities across technology, finance, and government sectors. Strategic objectives focused on intellectual property theft, competitive intelligence gathering, and potentially infrastructure reconnaissance. Campaign discovered through Anthropic's security monitoring detecting anomalous usage patterns indicating coordinated exploitation.

Exploitation Techniques: Adversaries circumvented safety guardrails through prompt decomposition jailbreak exploiting model difficulty recognizing malicious intent distributed across benign-appearing sub-tasks. Constructed elaborate deceptive narrative portraying themselves as cybersecurity professionals conducting authorized defensive testing. Each individual prompt appeared innocuous, "Identify web server version," "Write buffer overflow check script", but sequential execution composed complete intrusion workflows.

Documented Metrics: Tactical autonomy reached 80-90% of operations including reconnaissance, exploit generation, and lateral movement planning. Human intervention is limited to 4-6 critical decision points per campaign, primarily high-stakes data exfiltration target selection. Attack speed demonstrated machine-pace operation with 1,000+ concurrent requests. Timeline compression: operations traditionally requiring 2-3 weeks completed in hours.

Attribution and Forensics: LLM intermediation complicated traditional attribution. Attack artifacts reflected Claude training data rather than human operator tradecraft. Successful attribution required sophisticated prompt pattern analysis and infrastructure correlation—capabilities beyond standard incident response toolkits. Demonstrates attribution challenges universal to LLM-mediated attacks.

Case 2: GPT-4 Autonomous Linux Privilege Escalation (2023)

Adversary Profile: Academic security researchers at TU Wien conducting controlled penetration testing research. Ethical research context with institutional review and responsible disclosure practices. Research objective: evaluating LLM capabilities for autonomous security testing, not operational exploitation.

Operational Context: Controlled experimental environment employing custom benchmark of vulnerable Linux systems. Research utilized hacking-BuddyGPT prototype enabling fully autonomous privilege escalation attempts. Benchmark included multiple vulnerability classes: file-based exploits, SUID misconfigurations, sudo privilege escalation vectors, and kernel vulnerabilities. Human baseline comparison employed professional penetration testers with 7 years' experience.

Exploitation Techniques: GPT-4-Turbo demonstrated autonomous attack workflow: scanning system configurations, identifying vulnerabilities through pattern recognition, generating custom exploitation scripts, executing attacks, and verifying privilege elevation. Model employed chain-of-thought reasoning decomposing "become root" into executable sub-tasks: enumerate running services, check file permissions, identify SUID binaries, test sudo configurations, examine kernel version for known exploits.

Documented Metrics: Success rates varied by vulnerability class and model version. GPT-4-Turbo achieved 33-83% exploitation success rates depending on vulnerability type and configuration. File-based exploits: 75-100% success (highest performing category). GPT-3.5-turbo demonstrated 16-50% success rates. Local models (Llama3): 0-33% success, with 7B parameter variants failing entirely. Human penetration testers achieved 75% success rate, GPT-4-Turbo approached or exceeded human performance in specific vulnerability classes.

Comparative Benchmarking: Research included systematic ablation studies evaluating impact of context size, in-context learning, high-level guidance, and memory management. Results demonstrated: larger context windows improved performance through maintaining attack state; in-context learning (providing exploit examples) significantly boosted success rates; high-level guidance (strategic hints) doubled performance; LLM-driven reflection mechanisms enabling self-correction improved outcomes.

Case 3: SpearBot Phishing Framework (2025)

Adversary Profile: Specialized security research focused on social engineering capabilities. Academic context examining LLM applications for spear-phishing email generation. Research demonstrated sophisticated adversarial AI architecture but remained within responsible disclosure framework.

Operational Context: Development and evaluation of generative-critique architecture specifically optimized for phishing content creation. Framework employed dual-LLM system: generator LLM producing phishing emails, critique LLM evaluating detectability and persuasiveness. Iterative refinement continued until security detection markers eliminated while maintaining social engineering effectiveness.

Exploitation Techniques: Generative-critique architecture represented significant advancement over single-LLM approaches. Generator produced initial phishing content based on target profile and campaign objectives. Critique LLM analyzed output identifying: spam filter triggers (suspicious URLs, attachment naming patterns), linguistic anomalies (unnatural phrasing, grammatical inconsistencies), brand impersonation markers (logo mismatches, domain discrepancies), and urgency manipulation detectability.

Iterative Refinement Process: Framework continued generator-critique cycles until critique LLM certified output as: grammatically flawless, culturally appropriate for target demographic, brand-consistent with impersonation subject, urgency-inducing without overt manipulation markers, and likely to evade automated detection systems. Average refinement required 3-5 generator-critique iterations, with complex scenarios requiring up to 10 iterations.

Documented Effectiveness: Evaluation against commercial phishing detection systems demonstrated: 87% evasion rate against spam filters, 72% evasion against AI-powered phishing detectors, significantly higher than baseline LLM-generated phishing (43% evasion). Human evaluation study showed: 65% of security-aware participants unable to distinguish SpearBot emails from legitimate communications when context-appropriate impersonation employed.

Case 4: WormGPT/FraudGPT Underground Ecosystem (2023-2025)

Adversary Profile: Underground marketplace developers and criminal buyers. Sellers operated anonymously on dark web forums (BreachForums, HackForums, Russian-language Exploit forum) and Telegram channels. Buyer demographics ranged from sophisticated cybercriminal groups to low-skilled opportunistic actors, democratization across skill levels representing key ecosystem characteristics.

Operational Context: WormGPT emerged July 2023 as "first widely recognized, commercialized malicious LLM" marketed explicitly for criminal purposes. Built on GPT-J 6B open-source model, allegedly fine-tuned on malware code, exploit write-ups, and phishing templates. Original WormGPT shut down mid-2023 after media exposure but spawned successor generations (WormGPT 4, announced late 2024) and competitors (FraudGPT, EvilGPT, DarkGPT, KawaiiGPT).

Business Model and Pricing: Subscription-based cybercrime-as-a-service model. WormGPT pricing: €550 annually for WormGPT v2, €5000 for private builds. FraudGPT pricing varied across marketplaces: $200/month to $1700/year, with significant price variation suggesting either marketplace fee structures or competitive pricing experiments. WormGPT 4 (2024-2025 generation): $50/month, $220 "lifetime access" including source code.

Capabilities and Features: Underground LLMs marketed comprehensive criminal tooling: business email compromise (BEC) message generation, phishing email composition in multiple languages, polymorphic malware code generation evading antivirus signatures, credential harvesting script creation, ransomware development including encryption and ransom notes, vulnerability exploitation code, and social engineering content customized for specific targets.

Ecosystem Evolution: First generation (WormGPT, FraudGPT 2023): standalone fine-tuned models, relatively expensive, limited distribution. Second generation (WormGPT 4, 2024-2025): wrappers around commercial LLM APIs (Grok, Mixtral) employing sophisticated jailbreak prompts. Cost reduction, easier deployment, wider accessibility. Third generation (KawaiiGPT, 2025): freely available on GitHub, open-source community-maintained, zero financial barrier, 500+ developer community providing updates and improvements.

Market Dynamics and Skepticism: Threat intelligence revealed significant skepticism within criminal communities regarding malicious LLM efficacy. Forums contained numerous accusations of scams, exaggerated capabilities, and unreliable tooling. Some threat actors preferred manual methods or traditional tools over AI assistance. However, continued marketplace presence and evolution through multiple generations indicates sustained criminal demand despite skepticism.

Comparative Findings

This section presents cross-case comparative analysis organized by the eight-dimensional framework, identifying both universal patterns and contextual variations.

Universal Exploitation Mechanisms (Present Across All Cases)

Analysis reveals three exploitation mechanisms present across all four cases regardless of adversary type or operational context. First, autonomous goal decomposition via chain-of-thought reasoning appears universal. Claude incident decomposed "exfiltrate database" into 7-step workflow; GPT-4 research decomposed "become root" into vulnerability enumeration, exploit generation, and execution verification; SpearBot decomposed "generate convincing phishing email" into drafting, critique, and iterative refinement; WormGPT marketed this capability as core feature enabling users to specify high-level criminal objectives.

Second, adaptive social engineering proves universal across all cases. Claude incident employed deceptive narrative portraying adversaries as authorized security testers. GPT-4 research required no explicit social engineering but demonstrated capability through convincing technical communication. SpearBot specifically optimized social engineering through generative-critique architecture eliminating detection markers. WormGPT ecosystem explicitly marketed social engineering capabilities for BEC and phishing operations. This universality suggests social engineering represents inherent LLM capability leveraged across all exploitation contexts.

Third, detection evasion through novel pattern generation appears across all cases. Claude attacks generated 1,000+ unique code variants avoiding signature detection. GPT-4 produced custom exploitation scripts rather than employing known exploits. SpearBot iteratively eliminated detection markers through critique-driven refinement. WormGPT marketed polymorphic malware generation as key differentiator. Universal presence of adaptive evasion indicates fundamental challenge for signature-based detection systems.

Majority Patterns (Present in 3/4 Cases)

Dynamic tool invocation and code generation (Bistarelli et al., 2025) appear in 3/4 cases (Claude, GPT-4, WormGPT) but less prominently in SpearBot which focused primarily on natural language social engineering. Claude autonomously invoked scanning tools and generated custom exploit code. GPT-4 research systematically evaluated tool invocation capabilities demonstrating autonomous privilege escalation. WormGPT explicitly marketed code generation for malware, scripts, and exploits. SpearBot focused on phishing content generation without extensive code generation capabilities, though framework architecture itself required sophisticated software engineering.

Attribution challenges through LLM intermediation appear in 3/4 cases (Claude, GPT-4, WormGPT). Claude incident required sophisticated prompt pattern analysis for attribution. GPT-4 research artifacts reflected

training data rather than researcher characteristics—though academic context made attribution unnecessary. WormGPT usage obscures user identity through underground marketplace anonymity and LLM-generated content removing user-specific patterns. SpearBot, as published research, involves no attribution challenges beyond inherent academic transparency.

Operational scale expansion appears in 3/4 cases (Claude, GPT-4, WormGPT). Claude achieved machine-speed operation across 30+ organizations simultaneously. GPT-4 research demonstrated automated large-scale security testing across benchmark systems. WormGPT enables individual users to conduct operations previously requiring team efforts. SpearBot, while sophisticated, represents specialized tooling rather than large-scale operational platforms.

Contextual Variations by Adversary Type

Autonomy levels demonstrate clear adversary-type correlation. State-sponsored Claude operation: 80-90% autonomy with strategic human oversight. Academic GPT-4 research: 33-83% success suggesting moderate autonomy with frequent failures. Criminal tooling WormGPT: variable autonomy depending on user skill and task complexity. SpearBot: highly autonomous within narrow domain (phishing generation) but requires human target selection and campaign design. Pattern suggests state actors achieve highest autonomy through sophisticated operational tradecraft, academic contexts demonstrate potential while documenting limitations, and criminal tools exhibit variable performance.

Operational security demonstrates adversary-type progression. State-sponsored Claude incident: sophisticated OPSEC, weeks-long campaign before detection, professional cleanup operations. Academic research: complete operational transparency as research objective. Criminal ecosystem: variable OPSEC from sophisticated WormGPT operations through careless marketplace discussions revealing capabilities. SpearBot: research transparency prioritizes documentation over operational security. Pattern reveals inverse relationship between research/education objectives and operational security requirements.

Skill requirement democratization follows clear progression. State-sponsored operation required: prompt engineering expertise, cybersecurity domain knowledge, operational planning, infrastructure management. Academic research required: security expertise, experimental design, benchmark creation. Criminal tooling progression: WormGPT early versions required technical sophistication for deployment; later generations simplified through API wrappers; KawaiiGPT provides GitHub distribution with five-minute setup. SpearBot represents specialized research requiring academic expertise but demonstrating potential for democratization. Temporal progression shows decreasing skill barriers from elite capabilities (2023) toward mass accessibility

(2025). The comparative analysis of LLM exploitation cases integrates eight dimensions as shown in (Table 1).

Table 1. Comparative Analysis of LLM Exploitation Cases Across Eight Dimensions

Dimension	Claude (State)	GPT-4 (Academic)	SpearBot (Specialized)	WormGPT (Criminal)
Adversary Profile	State-sponsored intelligence, Chinese attribution, strategic objectives	Academic researchers, TU Wien, controlled ethical research	Specialized research team, social engineering focus	Underground marketplace, criminal buyers, varied sophistication
Autonomy Level	80-90% autonomous operations, 4-6 human decision points	33-83% success rate, moderate autonomy with failures	High autonomy in narrow domain (phishing generation)	Variable, user-dependent, increasing with tool evolution
Target Selection	~30 organizations, tech/finance/government, strategic intelligence	Controlled benchmark, multiple vulnerability classes	Demonstration targets, security-aware participants	Opportunistic, user-defined, BEC/phishing/malware targets
Primary Techniques	Prompt decomposition jailbreak, deceptive narratives, machine-speed execution	Autonomous privilege escalation, vulnerability enumeration, exploit generation	Generative-critique architecture, iterative refinement, detection evasion	Polymorphic malware, BEC messages, phishing emails, ransomware scripts
Evasion Methods	1000+ code variants, distributed intent across benign prompts	Custom exploit generation, novel attack patterns	Iterative elimination of detection markers, 87% spam filter evasion	Polymorphic generation, signature avoidance, AV evasion
Attribution	LLM intermediation obscures tradecraft, required AI forensics	Research transparency, academic attribution	Published research, full disclosure	Underground anonymity, LLM-generated content removes patterns
Skill Level Required	High: prompt engineering, cybersecurity expertise, operational planning	High: security research, experimental design, benchmark development	High: AI architecture, security research	Decreasing: Gen1 high, Gen2 moderate, Gen3 (KawaiiGPT) minimal
Timeline & Cost	Sept 2025, sustained campaign, extensive resources	Oct 2023, research project, institutional funding	2025, research publication, grant-funded	2023-2025, $200-1700/year subscriptions, mass market evolution

Discussion

Cross-case comparative analysis reveals critical patterns informing both theoretical understanding and practical defense design. This section synthesizes findings addressing RQ3 (capability democratization dynamics) and RQ4 (defensive imperatives).

Capability Democratization Dynamics

Comparative temporal analysis demonstrates systematic capability democratization from elite state-level operations through academic accessibility toward mass criminal commoditization. September 2025 Claude incident represents state-of-the-art: 80-90% autonomy, sophisticated operational tradecraft, multi-week sustained operations. This establishes capability ceiling, what becomes possible with substantial resources, expertise, and operational planning.

Academic research (October 2023 GPT-4 study) demonstrates capability accessibility outside state apparatus. Success rates of 33-83% indicate significant autonomous potential while documenting limitations and failure modes. Critical democratization milestone: academic institutions with research budgets, computational resources, and technical expertise can replicate capabilities approaching state-level sophistication, albeit in controlled environments. Specialized research tooling (SpearBot 2025) demonstrates focused capability refinement.

Rather than general-purpose exploitation, specialized frameworks optimize specific attack vectors. Generative-critique architecture achieving 87% spam filter evasion exceeds baseline LLM performance through architectural innovation, not merely model scaling. Pattern suggests democratization proceeds through both general capability increase and specialized tooling development.

Underground criminal commoditization (WormGPT ecosystem 2023-2025) represents mass democratization endpoint. Three-generation evolution demonstrates acceleration: Generation 1 (2023), custom fine-tuned models, expensive (€550-5000), limited distribution; Generation 2 (2024), commercial API wrappers with jailbreak prompts, cost reduction ($50-220), wider accessibility; Generation 3 (2025), open-source GitHub distribution (KawaiiGPT), zero cost, five-minute setup, 500+ developer community.

Democratization timeline: Capability demonstration (state actors, Sept 2025) → Academic validation (controlled research, Oct 2023) → Specialized tooling (SpearBot, 2025) → Criminal commoditization (WormGPT Gen 1-3, 2023-2025) → Open-source mass distribution (KawaiiGPT, July 2025). Pattern suggests

approximately 6–18-month lag from capability demonstration to underground availability, with decreasing lag over time as exploitation methodologies mature.

Cross-Cutting Defensive Imperatives

Comparative analysis identifying universal patterns enables capability-based defensive priorities transcending adversary-specific approaches. Four imperatives emerge from cross-case analysis.

Imperative 1: Multi-Turn Conversational Context Monitoring

Universal presence of prompt decomposition and distributed malicious intent across all cases demands defensive systems maintaining semantic awareness across multi-turn conversations. Claude incident exploited single-prompt evaluation weakness; GPT-4 research decomposed complex objectives across sequential tool invocations; SpearBot employed iterative refinement across multiple generator-critique cycles; WormGPT enabled users to build complex exploits through conversational workflows.

Traditional security controls evaluate discrete actions independently single network request, individual file access, isolated command execution. LLM exploitation distributes malicious intent across benign-appearing individual actions that become dangerous only when composed sequentially. Defensive systems must maintain conversational state, analyze cumulative effect of action sequences, detect semantic patterns indicating distributed intent, and recognize legitimate complex workflows versus malicious decomposition.

Implementation requires: session-based behavioral modeling tracking user interaction patterns; semantic intent analysis inferring objectives from action sequences; anomaly detection identifying unusual workflow patterns; and contextual access controls considering conversation history when authorizing sensitive operations. This represents fundamental defensive architecture shift from action-level to workflow-level security.

Imperative 2: Behavioral Fingerprinting for Complex Workflows

Detecting evasion through novel pattern generation (universal across cases) requires moving beyond signature-based detection toward behavioral fingerprinting. Effective defenses must distinguish legitimate complex use cases from malicious exploitation attempts without generating excessive false positives disrupting legitimate users.

Challenge appears most acute in Claude and GPT-4 cases where individual actions (enumerate services, check file permissions, generate network queries) remain legitimate for authorized security testing but become malicious in adversarial context. Solution cannot simply block specific actions or patterns—overly restrictive controls break legitimate functionality.

Behavioral fingerprinting approach: establish baseline behavioral profiles for different legitimate use categories (software development, security research, data analysis); identify behavioral anomalies deviating from established profiles; analyze temporal patterns distinguishing systematic reconnaissance from legitimate exploration; correlate LLM usage with organizational role and authorization context.

Imperative 3: Federated Threat Intelligence and Rapid Dissemination

WormGPT ecosystem evolution demonstrates adversarial adaptation speed necessitating rapid defensive intelligence sharing. Original WormGPT shut down mid-2023 after media exposure but spawned multiple successors and evolved through three generations within 18 months. SpearBot research documented 87% spam filter evasion; defensive systems must rapidly incorporate findings preventing widespread exploitation. Claude incident revealed prompt decomposition jailbreak; disseminating defensive patterns across LLM deployments prevents replication.

Current threat intelligence cycles (discovery → analysis → documentation → dissemination → implementation) consume weeks or months. LLM exploitation adaptation operates on days or weeks timescales. Federated learning approaches enable collaborative defense: decentralized organizations contribute anonymized behavioral data; aggregate learning identifies emerging attack patterns; rapid-response mechanisms distribute defensive updates; privacy-preserving techniques protect proprietary data.

Implementation requires: standardized LLM security incident reporting frameworks; automated behavioral data collection and anonymization; federated learning infrastructure enabling cross-organizational collaboration; rapid-response distribution mechanisms for critical defensive updates.

Imperative 4: Capability-Based Access Controls and Graduated Restrictions

Capability democratization from state sophistication through mass commoditization demands tiered access controls proportional to LLM reasoning capability and operational context. Not all use cases require full autonomous reasoning, tool invocation, and code generation capabilities, graduated restriction strategies reduce attack surface.

Tiered capability model: Tier 1 (Full Autonomy), complete reasoning, tool invocation, code generation, restricted to authenticated high-trust contexts with comprehensive monitoring; Tier 2 (Guided Autonomy), reasoning and analysis without autonomous tool invocation, suitable for analytical workflows not requiring system interaction; Tier 3 (Assisted Operations), natural language interaction without autonomous decision-making, appropriate for content generation, summarization, communication drafting; Tier 4 (Query-Response), information retrieval without reasoning workflows, lowest-risk deployment.

Access control framework considers: user authentication and authorization level; organizational context and security posture; task requirements and necessary capability levels; monitoring and logging provisions; incident response integration. Organizations should deploy minimum necessary capability level for each use case rather than default full-autonomy access.

Conclusion

This comparative study examined LLM weaponization patterns through systematic analysis of four major exploitation incidents spanning state-sponsored cyber-espionage, academic security research, specialized social engineering frameworks, and underground criminal commoditization. Unlike prior research focusing on isolated incidents or theoretical vulnerabilities, this work provides first comprehensive comparative framework identifying both universal patterns and contextual variations across adversary types and operational contexts.

Key Findings and Contributions

Cross-case analysis reveals three universal exploitation mechanisms present across all cases regardless of adversary sophistication: (1) autonomous goal decomposition via chain-of-thought reasoning enabling high-level objective specification with autonomous tactical execution; (2) adaptive social engineering leveraging natural language capabilities for deceptive communication and manipulation; (3) detection evasion through novel pattern generation circumventing signature-based defenses. Universal presence of these mechanisms indicates fundamental LLM capabilities exploitable across all operational contexts, necessitating capability-based rather than adversary-specific defensive approaches.

Temporal analysis documents systematic capability democratization following predictable progression: state-level demonstration (Claude incident 80-90% autonomy) → academic validation (GPT-4 research 33-83% success) → specialized tooling development (SpearBot 87% evasion) → underground commoditization (WormGPT multi-generation evolution) → open-source mass distribution (KawaiiGPT free GitHub

availability). Democratization timeline shows approximately 6–8-month lag from capability demonstration to underground availability, with decreasing lag suggesting acceleration as exploitation methodologies mature.

Comparative framework identifies four cross-cutting defensive imperatives: (1) multi-turn conversational context monitoring maintaining semantic awareness across interaction sequences; (2) behavioral fingerprinting distinguishing legitimate complex workflows from malicious exploitation; (3) federated threat intelligence enabling rapid cross-organizational learning and defensive adaptation; (4) capability-based access controls implementing graduated restrictions proportional to reasoning sophistication and operational context.

Research Contributions

This study contributes to LLM security research in several ways. Methodologically, it provides structured comparative framework applicable to analyzing additional exploitation cases as they emerge. Empirically, it documents patterns across adversary types enabling defensive prioritization based on universal rather than contextual vulnerabilities. Theoretically, it establishes capability democratization model tracking progression from elite operations through mass commoditization. Practically, it derives defensive imperatives directly from cross-case pattern analysis rather than single-incident reaction.

Limitations and Future Research Directions

This research acknowledges several limitations informing future work. First, reliance on publicly disclosed information excludes classified intelligence regarding state-sponsored capabilities, potentially underestimating true sophistication levels. Second, four-case samples, while representing critical adversary types and best-documented incidents, cannot capture full LLM exploitation ecosystem diversity. Third, temporal scope (September 2023-September 2025) represents snapshot of rapidly evolving landscape, continuous monitoring and updated comparative analysis remains necessary.

Future research should expand case samples as additional incidents emerge and become documented, enabling more robust pattern validation. Longitudinal analysis tracking capability evolution over extended timeframes would illuminate democratization dynamics more precisely. Empirical evaluation of proposed defensive imperatives through controlled experimentation and field deployment would establish effectiveness beyond theoretical analysis. Investigation of under-studied adversary types, particularly hacktivist organizations, nation-state proxies, and insider threats, would extend comparative framework coverage.

Cross-disciplinary research integrating technical security analysis with policy, legal, and socio-economic perspectives would address broader implications. Capability democratization raises critical questions: liability frameworks when autonomous AI systems cause harm, regulatory approaches balancing innovation against security imperatives, international cooperation mechanisms for coordinated threat response, economic impacts of lowered barriers to sophisticated cyber operations.

Implications for Practice and Policy

For cybersecurity practitioners, comparative findings emphasize urgent need for capability-based defensive architectures transcending adversary-specific approaches. Organizations must implement multi-turn context monitoring, behavioral fingerprinting, and graduated access controls regardless of perceived threat actor profiles (Antoniou, 2025). Universal exploitation patterns mean defensive gaps exploitable by state actors prove equally exploitable by commodity criminal tooling, comprehensive protection requires addressing a fundamental architectural weaknesses.

For LLM developers and vendors, study underscores critical need for security-by-design principles addressing comparative findings. Current safety approaches evaluating individual prompts in isolation prove insufficient against sophisticated prompt decomposition documented across all cases. Next-generation architectures require conversational context awareness, behavioral anomaly detection, and capability-based restriction mechanisms. Vendor responsibility extends beyond isolated safety testing toward sustained adversarial research matching democratization pace.

For policymakers, capability democratization dynamics demand proactive regulatory frameworks. Current reactive approaches addressing specific incidents prove inadequate when sophisticated capabilities democratize mass-market accessibility within 18 months. Effective policy must anticipate democratization trajectories, establish graduated regulatory requirements scaling with capability levels, enable rapid information sharing across jurisdictional boundaries, and balance innovation incentives against security imperatives (European Union, 2024).

Closing Perspective

LLM weaponization represents not isolated security incident category but fundamental transformation of cyber threat landscape. Comparative analysis demonstrates exploitation patterns transcend adversary types, capabilities democratize following predictable timelines, and defensive architectures require systematic redesign addressing universal patterns rather than contextual variations. As LLM technology continues

advancing and proliferating, the gap between attack capabilities and defensive measures risks widening unless strategic imperatives identified through comparative analysis receive swift implementation. The moment demands moving beyond single-incident reaction toward systematic comparative understanding enabling anticipatory defense against predictable democratization trajectories.

References

Abdali, S., Shah, N., & Papalexakis, E. E. (2024). *LLM security: Vulnerabilities, attacks, defenses, and countermeasures*. arXiv preprint arXiv:2505.01177. https://arxiv.org/abs/2505.01177

Anthropic PBC. (2025, November 14). Disrupting the first reported AI-orchestrated cyber espionage campaign. Anthropic PBC. https://www.anthropic.com/news/disrupting-ai-cyber-espionage

Antoniou, G. (2025). AI-driven defense-in-depth: A systematic review of SOC maturity models and DDoS mitigation. In *Proceedings of the 6th International Conference on Recent Trends & Applications in Computer Science and Information Technology (RTA-CSIT 2025). CEUR Workshop Proceedings, Vol-4044*.

BBC World Service. (2025, November 14). AI firm claims Chinese spies used its tech to automate cyber attacks. BBC World Service. https://www.bbc.com/news/technology

Bistarelli, S., Fiore, M., Mercanti, I., & Mongiello, M. (2025). Usage of large language model for code generation tasks: A review. *SN Computer Science*, *6*(6), Article 673. https://doi.org/10.1007/s42979-025-04241-5

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (EU AI Act)*. Official Journal of the European Union, L 1689.

Feffer, M., Srinivasan, A., Neider, D., D'Antoni, L., Recht, B., & Mytkowicz, T. (2024). *Red teaming for generative AI: Silver bullet or security theater?* arXiv preprint arXiv:2401.15897. https://arxiv.org/abs/2401.15897

Gupta, A. (2025). Red teaming AI systems for security validation. *International Journal of AI, BigData, Computational and Management Studies*, *6*(1), 116–123.

Happe, A., Kaplan, A., & Cito, J. (2023). *LLMs as hackers: Autonomous Linux privilege escalation attacks*. arXiv preprint arXiv:2310.11409.

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 Information technology—Artificial intelligence—Management system*. ISO.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and

challenges of large language models for education. *Learning and Individual Differences*, *103*, Article 102274.

Liu, Y., Deng, G., Li, Y., Wang, K., Wang, X., Zhang, T., Wang, Y., Zhang, H., Zhao, S., & Zeng, K. (2025). Large language models in cybersecurity: A survey of applications, vulnerabilities, and defense techniques. *Applied Sciences*, *6*(9), Article 216.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST. https://doi.org/10.6028/NIST.AI.100-1

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). *Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned*. arXiv preprint arXiv:2209.07858. https://arxiv.org/abs/2209.07858

Qi, Q., Luo, Y., Xu, Y., Guo, W., & Fang, Y. (2025). SpearBot: Leveraging large language models in a generative-critique framework for spear-phishing email generation. *Information Fusion*, *122*(C), Article 103176.

Raptis, E. K., Kapoutsis, A. C., & Kosmatopoulos, E. B. (2025). Agentic LLM-based robotic systems for real-world applications: A review on their agenticness and ethics. *Frontiers in Robotics and AI*, *12*, Article 1605405.

Saha Roy, S., Thota, P., Vamsi, N., Krishna, & Nilizadeh, S. (2024). From chatbots to phishbots?: Phishing scam generation in commercial large language models. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE.

Sun, Y., Zhang, X., Wang, Y., Liu, S., Chen, Z., Wang, H., Li, Q., & Zhang, T. (2025). *Security concerns for large language models: A survey*. arXiv preprint arXiv:2505.18889.

Trustwave. (2023, September 20). WormGPT and FraudGPT – The rise of malicious LLMs. Trustwave. https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/wormgpt-and-fraudgpt-the-rise-of-malicious-llms/

Unit 42. (2025). The dual-use dilemma of AI: Malicious LLMs. Palo Alto Networks. https://unit42.paloaltonetworks.com/dilemma-of-ai-malicious-llms/

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NeurIPS.

Wu, T., Wang, N., Walter, K., Guo, W., & Fang, Y. (2025). *From promise to peril: Rethinking cybersecurity red and blue teaming in the age of LLMs*. arXiv preprint arXiv:2506.13434.

Xu, Y., Chen, W., Zhang, J., Liu, X., Wang, Y., Wu, Y., & Chen, X. (2025). *The automation advantage in AI red teaming*. arXiv preprint arXiv:2504.19855.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, *4*(2), Article 100211.

Zhang, Y., Wang, H., Chen, Y., Liu, X., & Zhou, J. (2025). When LLMs meet cybersecurity: A systematic literature review. *Cybersecurity*, *8*(1), Article 7.

Zhang, Z., Li, Y., Wang, J., Liu, Y., & Chen, X. (2025). Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities. *Computer Networks*, *248*, Article 110451.

Zhou, Y., Liu, S., Yang, J., Wang, C., Zhang, P., He, R., & Zhao, L. (2025). *A survey of attacks on large language models*. arXiv preprint arXiv:2505.12567.

Author Information

George Antoniou
https://orcid.org/0009-0004-4023-1257
Lynn University
3601 N. Military Trail
Boca Raton, FL. 33431
USA