Attack steps	Summary
Step 1: Reconnaissance (Demonstrated in the real world)	**Qualitative:** Both Research and real-world attacks leverage LLMs in reconnaissance, including active scanning [265], victim information gathering [72, 197], and open-source database search [228].
	Quantitative: Current benchmarks rely primarily on text-based assessments [181, 269, 281], while practical evaluations like AutoPenBench [99] show AI agents' potential in reconnaissance but with limited capabilities.
Step 2: Weaponization (Demonstrated in the real world)	**Qualitative:** Research papers show LLMs can generate functional malware with high evasion rates [210] and AI agents aid in vulnerability identification and exploitation [88, 89, 171, 183], although large-scale real-world attacks remain limited to simpler attacks like DDoS [22, 276] and SQL injection [117, 197].
	Quantitative: RedCode benchmark [109] reveal LLMs' limitations in generating functional malware. Vulnerability exploitation benchmarks [247, 281, 311, 323] show Claude-Sonnet 3.5 achieving cybersecurity practitioner-level skills but falling short of expert capabilities.
Step 3-5: Delivery & Exploitation & Installation (Demonstrated in the real world)	**Qualitative:** Initial access and exploitation (installation) are demonstrated in real world [94, 140, 205] while persistence remains at research level [242].
	Quantitative: Benchmark coverage and quality for this attack step are very limited.
Step 6: Command and control (Demonstrated in the real world)	**Qualitative:** Most sub-categories demonstrate only in research papers, e.g., privilege escalation through exploit chain generation [67] and command and control through automated domain generation [12, 242], while credential access is used in real world [146, 208, 241].
	Quantitative: Benchmark coverage and quality for these attack steps are very limited.
Step 7: Action on objectives (Large-scale real world impact)	Real-world AI-enhanced attacks increase across various systems (Web, mobile, cloud) [22, 62, 135], with malicious purposes including malware deployment, business logic abuse, and credential theft, causing significant financial losses and data breaches [62, 140, 188].
Attacks against humans (Large-scale real world impact)	Frontier AI escalates attacks against humans, with studies showing increases in social engineering (135%) and voice phishing (260%) since ChatGPT's adoption [66, 129, 258]. AI is misused for identity theft [189], deepfake-based fraud [48, 186], child exploitation [262], and psychological manipulation [235].
Defense steps	**Summary**
Step 1: Proactive testing (Demonstrated in the real world)	**Qualitative:** Research explores LLMs for proactive cybersecurity testing, including automated penetration testing [72, 118, 153] and vulnerability detection through code foundation models [79, 183] and AI-enhanced fuzzing [137, 305]. Real-world demonstrations exist but lack evidence of large-scale adoption.
	Quantitative: AI penetration testing benchmarks remain nascent, with one end-to-end benchmark AutoPenBench [99]. Vulnerability detection benchmarks [79, 196, 244] show SOTA models achieve limited accuracy, but these benchmarks face challenges in label quality, limited code context, and insufficient test diversity.
Step 2: Attack detection (Demonstrated in the real world)	**Qualitative:** Frontier AI addresses key limitations of traditional AI methods by eliminating manual feature engineering, reducing dependence on labeled datasets [76, 277], and improving generalizability to out-of-distribution data [6, 115]. Real-world applications exist in malware and network intrusion detection.
	Quantitative: Many benchmarks exist for network intrusion and malware detection [18, 248, 310] but have data and label quality limitations. Foundation model-driven detection shows high accuracy [115, 175], calling for more challenging benchmarks.
Step 3: Triage & forensic (Demonstrated in research papers)	**Qualitative:** Frontier AI usages are at an early stage, including LLMs for vulnerability analysis, improving symbolic execution with PoC generation [124, 287] and developing AI agents for root cause analysis [232]. Pre-trained foundation models demonstrate superior performance in binary analysis tasks [217, 300].
	Quantitative: Recent benchmarks like CRUXEval [106] show promising results for PoC generation with GPT-4o achieving 75% pass@1 success rate, while no public benchmark exists for comprehensive AI-driven reverse engineering evaluation.
Step 4-5: Remediation development & deployment (No demonstrated effect)	**Qualitative:** Research demonstrates frontier AI can automatically generate security vulnerability patches [2, 33, 157, 174], though real-world applications like the SQLite3 Off-by-One bug fix [117] remain limited and no specific work exists on AI-assisted remediation deployment.
	Quantitative: SOTA systems using Claude-3.5-sonnet resolve about 50% of issues in SWE-bench-verified [145]. However, most SWE-bench issues are non-security bugs, revealing limitations in evaluating frontier AI's defense capabilities.
Defense for humans (Demonstrated in the real world)	Frontier AI shows promise for enhancing defenses against human-targeted attacks, including social bot detection [92], fraud detection [77, 320], deepfake detection [226], and misinformation detection [13, 78, 177]. However, defensive techniques struggle to keep pace with sophisticated attacks [5], highlighting the need for research on adversarial dynamics between defensive and attack AI systems.