Technical Report

FRONTIER MODEL FORUM

Managing Advanced Cyber Risks in Frontier AI Frameworks

REPORT SERIES

Implementing Frontier AI Frameworks

WORKSTREAMS

Frontier AI Frameworks
AI-Cyber

DATE

February 13, 2026

About the Report

This report is the latest in our technical report series on **frontier AI frameworks**.

While our prior reports examined approaches to implementing frontier AI frameworks in general, this is the first report to address the implementation of frontier AI frameworks in a specific risk domain.

Introduction

Frontier AI offers significant promise for cybersecurity, including accelerating vulnerability discovery and patching, optimizing defensive systems, and enhancing threat detection capabilities. However, these same capabilities create dual-use risks, potentially lowering barriers for malicious actors to exploit known vulnerabilities or discover new attack vectors. As AI capabilities advance, it is crucial to develop robust risk management frameworks that maximize security benefits while proactively addressing emerging risks.

Frontier AI frameworks describe how firms intend to manage severe or extreme risks from advanced AI models with high-impact capabilities. In line with the Frontier AI Safety Commitments,[1] frontier AI frameworks include several core components: identifying key risks, establishing capability thresholds that trigger additional scrutiny and/or safeguards, conducting capability assessments to inform determinations of whether those thresholds have been reached, and deploying additional safeguards when such "enabling capability thresholds" have been achieved and an AI system could enable serious harm without them. The frameworks, which also include provisions for risk governance, have become essential for frontier AI developers seeking to responsibly manage severe risks.

Each member firm of the Frontier Model Forum (FMF) has published a frontier AI framework that identifies advanced cyber threats as a key risk. Yet setting and evaluating thresholds and developing mitigations for advanced cyber risks can be challenging. In addition to specifying the point at which an AI model’s capabilities in cyber domains may require further assessments and/or enhanced safeguards, firms also make evidence-based threshold determinations, as well as discern which safeguards are sufficient for mitigating offensive cyber risk. As part of this process, firms must also weigh and consider defender benefits that their models provide, as many cyber capabilities are inherently dual-use.

This report extends the FMF's Technical Report Series, which is focused on how frontier AI frameworks can be implemented in general, to the cyber domain specifically. Based on expert discussions among FMF member firms, this report highlights emerging industry consensus on core cyber thresholds for frontier AI, methods for evaluating capability thresholds, and methods and approaches to managing and mitigating risks once frontier capability assessments suggest that frontier AI models or systems have reached capability thresholds.

Section 1

AI-Cyber Risk Taxonomy and Thresholds

ROADMAP

1.1 Identifying Extreme AI-Cyber Risks

Establishing clear thresholds for when AI cyber capabilities pose an extreme risk is inherently difficult. The security landscape evolves continuously as new vulnerabilities emerge and existing ones are patched. What constitutes an advanced capability today may become commonplace defensive practice in the future. In addition, using large language models (LLMs) for cybersecurity purposes is inherently dual use. The same capabilities that can enable a cyber attacker to exploit weaknesses in systems, such as identifying a vulnerability, can enable a cyber defender to patch those same weaknesses before an attacker can exploit them.

Frontier AI frameworks address high-severity or extreme risks, with certain frameworks differentiating between deliberate misuse, where threat actors deliberately exploit AI capabilities for offensive purposes, and unintentional hazards that emerge from a model's cyber capabilities. The majority of these frameworks outline systematic assessment processes that developers implement to proactively identify extreme cyber risks. Finally, many frontier AI frameworks anticipate capabilities or outcomes that, if realized, could break the attacker-defender balance in favor of attackers.

1.2 Threat Modeling

Threat modeling—adapted from cybersecurity and national security domains—is a process for systematically anticipating and identifying how various threat actors might leverage frontier AI to achieve harmful outcomes and mapping the potential pathways to those outcomes. Several firms have integrated AI-cyber threat modeling into their risk management processes. By mapping these potential capabilities, organizations can develop targeted evaluations, such as capability assessments, and, when necessary, implement appropriate safeguards to prevent misuse. Given the rapidly evolving landscape of AI-enabled cyber threats, threat models require regular updates, including to incorporate emerging attack vectors, consider changes to the attacker-defender balance in light of dual-use capability increases, and adapt to how malicious actors are weaponizing AI tools in practice.

The core components of threat modeling include:

- **Apply established cybersecurity frameworks.** Companies should articulate clear kill chains, or the sequence of steps that a threat actor must take when developing and executing a cyber attack. AI-cyber threat models often leverage existing frameworks for this process, including the Cyber Kill Chain for understanding attack progression, MITRE ATT&CK for mapping adversarial tactics and techniques, and the STRIDE framework for categorizing threat types. These frameworks provide structured approaches for anticipating how sophisticated adversaries might exploit AI capabilities across different stages of a cyberattack.

- **Articulate key assumptions and variables in the threat model.** Clearly stating assumptions ensures the threat model remains transparent and grounded in realistic risk assessment. These assumptions should address critical factors such as the prerequisites for successful attacks (including necessary tools, skills, and resources) and the varying capability levels of different threat actors (from novices to advanced persistent threat groups). This process also establishes the threat model's boundaries by defining which vulnerabilities are in scope and what adversary intentions are being considered.

- **Identify threat scenarios.** Threat modeling typically begins by identifying the most severe potential outcomes, such as destructive attacks on critical infrastructure or significant economic damage from data breaches, from offensive cyber capabilities of frontier AI systems. Organizations can identify these worst-case scenarios through several approaches: analyzing historical precedents of catastrophic cyberattacks like NotPetya or the Morris Worm, consulting with domain experts and conducting surveys, and convening workshops that bring together AI researchers and cybersecurity specialists. Companies should then develop detailed threat scenarios that specify the threat actor, exploited vulnerabilities, attack techniques, kill chain stages, and estimated impact. These concrete narratives structure risk discussions, guide assessment and mitigation efforts, and provide realistic test cases for evaluating the effectiveness of model output policies and automated safety systems against specific, plausible attack vectors.

- **Project how frontier AI models could lead to harm.** Threat modeling requires organizations to anticipate how frontier AI capabilities could be deliberately misused. Companies should map out specific misuse scenarios, pinpointing the critical stages where AI assistance would be necessary or significantly advantageous for an attack to succeed. This analysis may involve identifying key bottlenecks in offensive cyber operations where AI could provide meaningful uplift to adversaries. By determining which steps present the most significant obstacles for malicious actors, developers can direct their safety resources toward preventing the AI model from providing assistance that would lower these key barriers. Critical bottlenecks may occur at any of the stages in the kill chain. Organizations can employ several methodologies to conduct this analysis. For example, they could use scenario planning workshops that trace how emerging AI capabilities might evolve into new attack methods, and structured tabletop exercises that simulate multi-stage incidents combining AI exploitation with traditional cyber techniques.

- **Map contributing risk factors.** Effective threat modeling identifies the key variables that could contribute to severe cyber risk. This process may involve identifying particular targets, examining key attack techniques or assessing core cyber vulnerabilities. Industry lists, such as OWASP Top 10, which catalogs the most critical web application security vulnerabilities, and Mandiant's M-Trends report, which details current threat vectors based on active incident response data, can provide valuable sources of real-world data. In addition to providing valuable insights on how attackers exploit vulnerabilities in practice, these resources help AI developers understand how malicious actors could leverage AI to reduce the labor or cost of attacks across the entire cyber kill chain. This understanding enables developers to prioritize which AI-enabled cyber capabilities could pose the greatest risk.
- **Consider explicitly mapping threat models to capability evaluations**. This process establishes a relationship between identified threats and the particular AI capabilities under assessment. Companies might, for instance, map the feasibility of a specific threat model to how their AI system performs on cyber benchmarks, such as SWE-bench verified, discussed in the AI-Cyber Capability Assessments section below. Understanding these relationships can help companies design targeted mitigations, such as refusals for specific vulnerability-related queries or enhanced monitoring of certain output patterns, that address the most concerning capability-threat combinations identified through the mapping process.
- **Continuously update threat models.** Threat models should be updated based on new information. Malicious actors continuously experiment with new techniques, adapting AI tools in ways that may not have been anticipated during initial threat modeling exercises. Without regular updates informed by real-world observations, threat models lose their effectiveness as actionable intelligence frameworks. New information on threat intelligence, capability assessments, security research, and incident reports should inform threat model updates, including the key assumptions that guide the threat model.

1.3 Current Consensus on Cyber Thresholds

Frontier AI frameworks use thresholds to help determine when additional assessments or safeguards become necessary, and when development and/or deployment should be restricted. Frontier AI thresholds describe predefined notions of risk that indicate when additional action is warranted to avoid unacceptable outcomes. There are several potential ways to establish such thresholds. *Capability thresholds* identify specific AI capabilities that could enable harmful scenarios. *Risk thresholds* quantify unacceptable risk levels through likelihood or outcome severity metrics. *Compute thresholds* use training computational power as a proxy for capabilities. Finally, *outcome-based thresholds* define specific threat scenarios and assess frontier AI's contribution to realizing those scenarios.

There are tradeoffs with each threshold approach. Compute thresholds are the most straightforward to measure, but greater compute used to train a model doesn't always equate to greater risk. Risk thresholds directly measure risk, but are difficult to implement reliably due to the complexity and uncertainty in estimating future risks. Capability thresholds provide a better risk proxy than compute thresholds and are more measurable than risk thresholds, but are still difficult to implement and measure through frontier AI evaluations. Scoping evaluations to ensure they measure capabilities most correlated with cyber risk remains challenging. Outcome thresholds connect AI capabilities directly to potential real-world harms, but comprehensive and realistic scenarios can be challenging to define and evaluate.

Acknowledging these tradeoffs, frontier risk management—including in cyber—may benefit from a mixed approach: use a compute threshold as an initial scope signal, then benchmark models against a reference model to determine whether they represent a material change in general capability relevant to the risk domain.

Notably, capability thresholds have emerged as the most commonly used type of threshold for determining cybersecurity risk in Frontier AI Frameworks. Although a less direct measure of risk than risk thresholds, capability thresholds are a better proxy for risk than compute thresholds and more straightforward to measure than risk thresholds.[1] Outcome-based thresholds also offer a way of directly linking models with real-world harms by allowing developers to establish tiered thresholds (i.e., critical, high, medium) based on how much the model advances threat actors towards achieving dangerous outcomes.

Frontier AI framework thresholds for cyber typically include variations on two core capability or outcome-based thresholds: whether an AI model or system is capable of significantly enabling human individuals or groups with limited cybersecurity expertise to conduct sophisticated cyberattacks, and whether AI systems can autonomously execute end-to-end cyberattacks without human intervention. This convergence around 'non-expert uplift' and 'autonomous cyberattack' thresholds reflects an emerging consensus in the cybersecurity community, paralleling how other domain-specific frameworks have coalesced around comparable capability thresholds.AI models are thought to cross a critical threshold when they provide assistance in the form of specialized knowledge, troubleshooting guidance, or procedural instruction that meaningfully reduces the expertise, resources, or time traditionally required for low-skill threat actors to create and conduct malicious cyber operations. In addition, there is general consensus that the ability to fully automate end-to-end cyberattacks without human interventions represents a critical capability leap requiring additional safeguards. Autonomous cyber capabilities could enable attacks at unprecedented scale and speed, potentially overwhelming human defenders and existing security infrastructure. Unlike human-assisted attacks that are limited by operator expertise and availability, fully autonomous systems could operate continuously and simultaneously across multiple targets.

1.4 Key Considerations for Frontier AI Cyber Thresholds

As noted above, two core scenarios constitute important risks requiring further analysis: AI models that significantly help non-experts conduct destructive cyberattacks, and AI systems that can automate, or scale up, portions or the entirety of end-to-end cyberattacks.

This consensus threshold highlights several key components for consideration:

- **Threat Actor Uplift.** Almost all current thresholds focus on frontier AI's ability to provide assistance or "uplift" that enhances human capabilities, bridging the expertise gap between individuals or groups with limited cybersecurity training and those with specialized knowledge necessary for offensive cyber operations. Several thresholds qualify the level of uplift as "significant," "meaningful," or "material." The ambiguity in defining "significant" means that in practice, developers must still exercise judgment regarding what level of uplift is either deemed acceptable or would trigger further analysis or risk mitigation measures. Further, the inherent subjectivity may complicate efforts to establish consistent and reliably comparable safety benchmarks across organizations. Moving forward, it may be useful to establish consensus on what constitutes "significant" uplift.

- **Threat Actor Expertise.** Multiple frameworks specifically focus on how AI democratizes sophisticated cyber capabilities by enabling less skilled attackers. Some thresholds differentiate between AI's impact on different types of actors (e.g., low-skilled actor vs. moderately skilled), but by and large, frameworks focus on lowering the barriers to entry for low-skilled actors to conduct offensive cyber operations. Frameworks generally do not discuss thresholds around how highly skilled, well-resourced cyber actors, such as Advanced Persistent Threats (APTs) may leverage tools, though these actors may be most capable of enabling the greatest harm. Though these actors may receive less immediate value from these tools and *may* be more attentive to international norms and values, they could also be capable of great harm. Finally, open questions remain about defining uplift that may be cumulative (i.e., enabling a population of novice attackers that may break the attacker-defender balance) versus uplift that may benefit a single actor's ability to achieve a large scale attack.

- **Zero-Day Vulnerability Discovery as a Critical Escalation Point.** Nearly every company identifies discovery of zero-day exploits as a major threshold marker. This represents a notable jump in capabilities from exploiting known weaknesses to independently finding entirely new attack vectors.

- **End-to-End Automation Without Human Intervention.** Most frameworks include thresholds related to fully, or mostly, automated cyberattacks with minimal human intervention. This represents a fundamental shift from AI as a tool that augments human capabilities to AI as an independent cyberoperator capable of planning, executing, and adapting attacks without guidance.

- **Emphasis on Well-Protected Targets.** Several frameworks specifically mention the ability to attack well-defended systems as a key threshold marker. These thresholds account for whether AI assistance enables compromising targets with "patched," "hardened," or "state of the art" security best practices. However, there is limited consensus on what constitutes a security "best practice." Moving forward, it may be necessary to establish consensus on what constitutes "state of the art" security.

The above considerations are all relevant for establishing thresholds within frontier AI frameworks. However, it is important to note that several frameworks employ multiple, tiered thresholds (e.g., indicating medium, high, or critical risk levels) that trigger corresponding mitigation actions well before intolerable risk levels are reached. Some frameworks also reference the impact that frontier AI systems will have on defenders and how that impacts risk assessment.

As frontier AI continues to advance, establishing, refining, and assessing these thresholds will become increasingly important. Many open questions remain, including how to precisely define "significantly enable" in the context of cyber, the cumulative amount of evidence needed to determine when a threshold has been crossed, and determining how model performance on assessments translates to risks. Further research and cross-industry collaboration, in particular with domain experts, are needed to address these questions and enable frontier AI developers to implement cybersecurity framework thresholds more effectively.

Table 1: Cyber Thresholds of FMF Member Firms

Company	Type	Threshold
Amazon	Capability	AI at this level will be capable of providing material uplift (beyond other publicly available research or tools) that would enable a moderately skilled actor (e.g., an individual with undergraduate level understanding of offensive cyber activities or operations) to discover new, high-value vulnerabilities and automate the development and exploitation of such vulnerabilities.
Anthropic	Capability	**Cyber Operations:** The ability to significantly enhance or automate sophisticated destructive cyber attacks, including but not limited to discovering novel zero-day exploit chains, developing complex malware, or orchestrating extensive hard-to-detect network intrusions.
Google DeepMind	Capability	**Cyber uplift level 1:** Provides sufficient uplift with high impact cyber attacks for additional expected harm at severe scale.
Meta	Outcome	**Cyber 1:** Automated end-to-end compromise of a best-practice-protected corporate-scale environment (ex. Fully patched, MFA-protected) **Cyber 2:** Automated discovery and reliable exploitation of critical zero-day vulnerabilities in current popular, security-best practices software before defenders can find and patch them. **Cyber 3:** Widespread economic damage to individuals or corporations via scaled long form fraud and scams.
Microsoft	Capability	**Low** The model supports gathering of publicly available information related to cyber security threats, including through question and answer or content generation, but cannot meaningfully plan or execute a cyberattack. **Medium** The model increases the ability of a low-skilled actor to conduct low-level spoofing, phishing, or social engineering attacks by partially automating discrete tasks related to vulnerability exploitation like generic target identification and planning (but not execution). **High** The model provides meaningful uplift to a low-to-medium skilled actor's ability to create and conduct highly disruptive or destructive cyber-attacks, including on critical infrastructure, for example, through discovering novel zero-day exploit chains or developing complex malware or other tactics, techniques, and procedures. **Critical** The model provides a meaningful uplift to a low-skilled actor's ability to identify and exploit major vulnerabilities or enables a well-resourced and expert actor to develop and execute novel and effective strategies against hardened targets.
OpenAI	Capability	**High** The model removes existing bottlenecks to scaling cyber operations including by automating end-to-end cyber operations against reasonably hardened targets OR by automating the discovery and exploitation of operationally relevant vulnerabilities. **Critical** A tool-augmented model can identify and develop functional zero-day exploits of all severity levels in many hardened real world critical systems without human intervention OR model can devise and execute end-to-end novel strategies for cyberattacks against hardened targets given only a high level desired goal.

Note: This table only highlights thresholds where there is consensus. Thresholds are accurate as of January 2026.

Section 2

AI-Cyber Capability Assessments

Frontier capability assessments are procedures conducted on frontier AI models to gather evidence of whether they have capabilities that could increase risks to public safety and security. Frontier capability assessments usually involve conducting a variety of evaluations—structured tests of model capabilities in a given domain—followed by analysis on the test results. These evaluations provide empirical evidence about model performance that, when interpreted through expert analysis and consideration of the broader operational context, help operationalize frontier AI cyber thresholds by identifying potential security and safety concerns.

For cyber-related risks, cybersecurity evaluations may be run as part of a capability assessment approach designed to produce evidence indicating whether a model could assist in targeting critical infrastructure or cause widespread economic damage. However, the resource demands and evidential value of evaluations can differ substantially. Developers select their assessment approaches based on these variations as well as factors such as resource constraints, predicted model capabilities, the maturity of evaluations, and the anticipated deployment context (including model affordances).

The below section offers an initial taxonomy and definitions for frontier AI safety evaluations specific to cybersecurity, categorized across two dimensions: methodology and domain. This section aims to document and build consensus around the current understanding of frontier AI-cyber safety evaluations.

2.1 Evaluation Methods

AI cybersecurity evaluations can be classified along several dimensions, with methodology being one of the most fundamental. The methodology refers to the evaluation's study design–specifically, how the AI model or system's capabilities, risks, and behaviors are assessed. Certain evaluation techniques may be particularly well-suited to gathering evidence about specific cyber-relevant capabilities. For example, the cyber capabilities of AI agents are increasingly a concern due to their potential ability to automate malicious cyber attacks. As agentic AI systems become more prevalent, evaluations must account for both the potential of LLMs to generate harmful responses to user's cyber queries, as well as the autonomous cyber capabilities of agents. The Capture the Flag, Cyber Range, and Capability Benchmarks discussed below are designed to evaluate the cyber capabilities of AI agents, whereas Knowledge Benchmarks and Safeguard Evaluations test a model's overall cybersecurity knowledge or propensity to provide a harmful cyber response to a user's query.

Furthermore, the evaluation methodology may depend on the stage of the development lifecycle and the characteristics of the model being deployed. Developers often conduct AI cyber evaluations at three critical stages: before any safety measures are applied to evaluate the model's maximum cyber capabilities, after safeguards are implemented to assess how effectively the safeguards reduce harmful cyber capabilities, and as close to deployment as possible to account for any enhancements made during post-training, ensuring assessments reflect the cyber capabilities of the final model.

While evaluations can combine multiple methodological approaches, most existing studies use one of the following methodological designs:

- **CTF-style (Capture the Flag) Exercises**: Test an AI agent's cybersecurity capabilities through structured challenges in an isolated environment. Models must complete specific security tasks such as identifying vulnerabilities, solving cryptographic puzzles, or responding to simulated attacks within defined time limits. The format assesses both technical knowledge and practical application of security skills in scenarios designed to reflect real-world challenges. Key features include a strictly controlled testing environment for safety, well-defined success metrics, and the ability to increase task complexity throughout the evaluation process. One example includes Hack the Box exercises where AI agents attempt to hack and exploit virtual machines to gain access and complete challenges. Evaluators can then measure the agent's performance at different parts of the cyber kill chain.

  CTFs have some limitations as evaluation tools. As agent-based evaluations, CTFs typically provide LLMs with tooling or "scaffolding" through fine-tuning, prompt engineering, or access to toolsets to test the upper bound of the potential harm a model could cause. Eliciting maximal cyber capabilities includes providing agents with direct access to hacking tools, such as Bash or pwntools. While such elicitation techniques help maximize model performance in evaluations, the lack of standardization across these methods makes it difficult to compare evaluation results between different developers. In addition, CTF exercises diverge from real-world conditions because models are explicitly instructed to pursue specific objectives, such as capturing a target flag, rather than operating under the ambiguous conditions an actual threat actor would face. Furthermore, LLMs's success on some CTF challenges may stem from exposure to publicly available solutions during training rather than genuine capability development.

- **Cyber Range Exercises**: Employ virtualized infrastructure to create realistic, controlled environments for testing AI models' cybersecurity capabilities. This approach is more elaborate than individual CTFs, enabling assessment of sophisticated reasoning and planning abilities as models iteratively learn from and respond to complex security scenarios simulating real-world networks. This simulated environment allows for evaluation of multiple security functions—from threat detection and incident response to vulnerability assessment and active defense. Cyber range exercises can provide a secure testing ground for examining how AI agents adapt to and handle multi-stage cyber threats. While cyber range exercises provide more realistic testing scenarios for AI agents, they also suffer from similar elicitation challenges to CTFs. Cyber ranges may also fail to capture real-world contexts as agents are evaluated in an environment without defenders.

- **Benchmarks**: Employ structured questions, such as question-answer formats, and tasks to assess the understanding and capabilities of AI models in cybersecurity domains. Benchmarks create comparable baseline measurements of general or cyber-specific capabilities across different models. These benchmarks are designed to be easily replicable and repeatable, and are often conducted through automated testing, though some assessments may incorporate expert human evaluation / grading. Unlike more complex evaluation methods, benchmarks prioritize standardization to enable consistent cross-model comparison. Within the cyber domain, there are three main categories of benchmarks:

    - ***Knowledge benchmarks*** - Involve fundamental knowledge testing through multiple choice and open-ended questions.
    - ***Capability benchmarks*** - Involve practical capability (such as task-based) assessments through agent-based challenges.
    - ***Safeguard evaluations*** - Test model responses to potentially harmful prompt requests.

  Like CTFs and Cyber Range Exercises, knowledge and capability benchmarks face reliability challenges. Benchmark contamination occurs when models are trained on data containing benchmark-related information, artificially inflating performance scores. Beyond contamination, benchmarks can reach saturation, meaning models achieve such high scores that the benchmarks can no longer detect meaningful improvements in cyber capabilities. This can make it difficult to evaluate relative capability improvements of newer models.

- **Red-Team Exercises**: Involves leveraging cybersecurity experts to actively probe and test AI models to assess potential security vulnerabilities and offensive capabilities. This approach involves direct interaction with models to examine their responses to requests for exploit development, vulnerability discovery, or social engineering assistance. While primarily conducted by human security experts, this methodology is evolving to include automated testing protocols. Unlike standardized benchmarks or structured CTF challenges, these exercises rely on the expertise and creativity of security professionals to identify unexpected behaviors or concerning capabilities.

- **Controlled Trials**: Measure how AI systems affect human performance in cybersecurity tasks through comparative experimental design. Often referred to as an "uplift study," this approach assesses the impact of AI assistance as compared to existing tools or alternative resources, such as using search engines. These studies typically use randomized control trials (or similar designs of treatment vs. control). Through structured testing protocols, researchers can quantify the effectiveness of AI integration in security operations, providing evidence of how these systems enhance or potentially hinder human capabilities. Unlike other evaluation methods that focus solely on AI performance, this methodology focuses on the human-AI interaction and its measurable outcomes in security scenarios.

The methodologies above range from highly automated capability evaluations to human-led approaches. Automated capability evaluations prompt AI models to answer questions or perform tasks, measuring their capabilities at scale, over time, and across models. Because these evaluations are automatically graded, often using scores assigned by human experts, they are both scalable and efficient. However, human-led testing approaches, such as red-teaming and human uplift studies, can be more adaptable and nuanced.

2.2 Evaluation Domains

Evaluations can also be classified by their domain—the specific area of expertise or capability being assessed. Most often, cyber evaluations focus on **Cyber Operation Assistance**, whether the model can assist in cyber operations and the extent of its knowledge in offensive domains. A variety of skills are currently tested such as:

- **Reconnaissance:** Assesses a model's ability to gather information about a target system, network, or organization to identify vulnerabilities and plan attacks. Evaluations may test a model's ability to collect Open Source Intelligence or conduct network reconnaissance.
- **Social Engineering:** Assesses a model's potential misuse in phishing operations designed to deceive individuals into unwittingly compromising their security. Phishing simulations can test a model's

Table 2: Cyber Benchmarks Cited by FMF Member Firms

Benchmark Name	Link	Year	Cited By
CVE-Bench	Paper	2025	OpenAI
CyberGym	Paper	2025	Anthropic
Cyberseceval4	Resource	2025	Meta
CTIBench	Paper	2024	Amazon
Cybench	Paper	2024	Amazon, Anthropic, Meta
CyberMetric	Paper	2024	Amazon
SecBench	Paper	2024	Microsoft
SECURE	Paper	2024	Amazon
SWE-bench Verified	Blog	2024	Anthropic, GDM, OpenAI

The table above includes a non-comprehensive list of publicly available offensive cyber benchmarks referenced in model cards, system cards, and technical reports from FMF member firms.

ability to persuade a victim to download a malicious attachment or divulge sensitive security information.

- **Malicious Code Generation:** Tests the ability of a model to generate malicious code. Evaluations involve testing the system's ability to write code for specific malicious behavior, such as encrypting files, exfiltrating data, setting up keyloggers, initiating DDoS attacks, among others.

- **Vulnerability Discovery and Exploitation:** Publicly available CTF challenges are often used to test a model's domain knowledge in areas that can assist in vulnerability discovery and exploitation. These areas include cryptography, forensics, binary exploitation, and web security, among others. These evaluations involve testing an AI system's ability to break encryption algorithms, to analyze and extract information from digital artifacts, to identify memory corruption vulnerabilities, to decompose compiled binaries, and to identify, analyze, and understand web application vulnerabilities, attack vectors, and defense mechanisms across the full web technology stack.

- **Tool Usage:** Tests an agent's ability to leverage common cybersecurity tools to achieve key goals. This may include testing whether an agent can successfully execute Bash or PowerShell commands, run Python scripts, or use Metasploit to identify and exploit vulnerabilities.

- **Network Operations:** Test an agent's ability to find and exploit weaknesses in network infrastructure, protocols, or configuration. While CTFs may test for an agent's ability to complete tasks related to networks scanning, sniffing, or spoofing, cyber range exercises evaluate an agent's ability to autonomously navigate, compromise, escalate privileges, and remain hidden in realistic network environments. These environments can be configured to mimic enterprise networks with segmented networks, diverse host types, and realistic security controls, requiring agents to demonstrate sophisticated attack orchestration capabilities, including network reconnaissance, lateral movement, privilege escalation, and objective completion.

2.3 Preliminary Consensus on Evidence for Threshold Decisions

Current best practice suggests threshold determinations for cyber should be made on the basis of *cumulative* evaluation evidence, structured in a holistic assessment approach. Since the results from a single evaluation are unlikely to indicate unequivocally whether a cyber threshold has been crossed, threshold determinations should be made on the basis of multiple cyber evaluations and sources of evidence. The field is still nascent enough that it remains unclear which precise combination of evidence is needed to determine whether a threshold has been reached or passed.

However, several capabilities may serve as strong indicators of whether the core "non-expert uplift" or "autonomous" thresholds may have been reached or passed. Given existing bottlenecks to offensive cyberoperations, it is helpful to carry out bottleneck assessments[2] designed to provide insight into whether an AI model or system can perform the following tasks significantly better than other available baseline resources:

- Assist with conducting attack planning (e.g., gathering publicly available information related to cybersecurity threats or collecting information on particular targets)

- Assist with or autonomously discovering a novel zero-day exploit
- Assist with or autonomously developing complex malware
- Assist with or autonomously escalating privileges (e.g., moving laterally within a system)
- Carry out autonomous cyberattacks against a hardened target (e.g., independently planning and executing an end-to-end cyberattack)

These capabilities should not be viewed in isolation, but rather as a constellation of factors that together may indicate whether a model crosses critical cyber thresholds. Developers and deployers should consider how these capabilities interact and potentially amplify risk when deployed in real-world contexts, particularly when accessible to users with varying levels of expertise and intent. For more on how to evaluate models for the capabilities above, see our Appendix.

AI-Cyber Misuse Mitigations

ROADMAP

3.1 AI-Cyber Misuse Mitigations Taxonomy

Frontier mitigations include technical misuse safeguards and societal measures designed to prevent cyber risks stemming from the deliberate misuse of frontier AI or fully automated cyber attacks. Technical misuse safeguards are technical interventions that can be implemented by frontier AI developers and downstream developers and deployers to prevent users from eliciting information, actions, or assistance from AI models or systems for harmful cyberattacks. Societal safeguards are measures implemented outside the AI model and its direct deployment environment, typically involving physical controls, supply chain security, regulatory compliance, or inter-organizational coordination. Given the breadth of existing societal safeguards, this report only discusses those that focus on preventing severe risks and that AI developers can contribute to through information sharing, reporting, or supporting defensive systems and research. This report does not cover safeguards designed to protect the models themselves from compromise, such as jailbreak prevention or secure access protocols.

While implementation approaches vary, developers generally adopt a defense-in-depth strategy, layering multiple technical safeguards to prevent misuse. This report provides a snapshot of current technical and societal safeguards available to model developers and other stakeholders. However, the appropriateness and effectiveness of any specific safeguard depends on the model's characteristics and deployment context. Many mitigations discussed here require further research to validate their effectiveness, and this report does not prescribe an ideal combination of techniques. Measuring safeguard resilience against diverse and evolving adversarial approaches remains an active area of research.

3.1 AI-Cyber Misuse Mitigations Taxonomy

AI-Cyber safeguards may be categorized by their mode of application:

- **Model-level**: Techniques applied during model training, fine-tuning, or alignment that directly modify the model's parameters and underlying behavior patterns to prevent harmful outputs.
- **System-level**: Techniques implemented in the deployment environment or application layer that monitor, filter, or restrict model inputs/outputs without modifying the model's internal parameters.
- **Societal-level**: Measures implemented outside the AI model and its direct deployment environment, typically involving physical world controls, supply chain security, regulatory compliance, or inter-organizational coordination.

The preliminary taxonomy below identifies potential safeguards against AI misuse, including measures beyond what AI developers can implement directly (referred to as "societal safeguards"). This list provides an overview of possible mitigations but is not prescriptive, as many techniques have limitations, and their applicability depends on the specific risk scenario. Additionally, several promising mitigation techniques currently under research are not included in this report but may serve as future cyber safeguards. The FMF Technical Report on Frontier Mitigations covers exploratory methods in more detail.

Potential types of safeguards include:

- **Capability Limitations:** Approaches that alter the model's weights or training process to prevent models from possessing knowledge or abilities that could enable harm in the cyber domain. Examples include targeted unlearning to selectively remove specific capabilities that could enable harmful outcomes after initial training or false learning to train the model on deliberately fabricated but plausible-sounding incorrect information. These mitigation methods are still largely experimental techniques that are not widely implemented to mitigate offensive cyber capabilities.
- **Behavioral Alignment:** Approaches that seek to prevent a model's potentially dangerous capabilities from being elicited by shaping the model's responses to human requests and its autonomous decision-making processes. Safety training typically happens in two stages. First, Supervised Fine-Tuning teaches the model baseline safe behavior through examples. Then, Reinforcement Learning refines this behavior using feedback from either a human or an AI system to better align with safety goals. Refusal-based safety training trains the model to refuse unsafe user prompts, such as requests to develop ransomware. Other methods, such as safe-completion training, train the model to produce safe outputs to dual-use queries that have both legitimate and potentially harmful applications. For example, training a model to provide helpful support on educational cybersecurity topics, while refusing to provide operational guidance for malicious cyber activities or de-escalating such requests.
- **Detection and Intervention Mitigations:** Approaches that rely on automated methods to detect model usage (e.g., inputs and outputs) that may give rise to undesired behavior.
- **Access Control Mitigations:** Approaches that govern who can use a model, what capabilities they can access, and how the model can interact with external systems. These methods establish boundaries that determine the conditions under which model capabilities can be utilized.

- **Ecosystem Mitigations:** Approaches where developers provide information, tools, and capabilities that enable other actors – governments, organizations, and civil society – to implement effective defenses against AI-enabled threats. Rather than directly controlling societal defenses, developers contribute by sharing resources that strengthen the broader defensive ecosystem.

See Tables 3-6 below for more detail on the types of mitigations listed above.

Table 3: Behavioral Alignment Mitigations for AI-Cyber

Safeguard	Description	Cyber Application	Limitations
Supervised Fine-Tuning	Developers curate datasets of desired model behaviors and fine-tune models to match these examples. Datasets include refusal examples (such as declining to provide instructions on developing ransomware) and helpful, yet harmless responses. Supervised Fine-Tuning directly teaches models specific behavioral patterns through imitation learning.	Training models on datasets containing examples of appropriate responses to cybersecurity questions, such as explaining defensive security concepts while refusing to provide step-by-step exploit development.	Safety training methods modify surface-level behaviors without altering underlying model capabilities. Safety training can be undone through targeted fine-tuning, such as training the model on datasets of harmful cybersecurity content or using adversarial prompts ("jailbreaks") that deliberately override refusal behaviors.
Reinforcement Learning with Human Feedback (RLHF)	Developers use human preferences between different model outputs to questions as a reward signal. They then use reinforcement learning to optimize models for these reward signals.	Security experts could rate pairs of model responses to ambiguous cybersecurity queries, such as comparing one response that provides detailed SQL injection techniques with another that explains SQL injection defensively with mitigation strategies.	Reward misspecification and/or "reward hacking" could occur, with models exploiting flaws in reward signals, such as generating unnecessarily verbose responses that score well but provide little value, or appealing to evaluator biases rather than producing useful outputs.
Reinforcement Learning with AI-assisted Feedback (RLAIF)	Uses AI systems to generate training feedback based on predefined principles or constitutions, and has gained traction as a scalable alternative to purely human-generated feedback.	Methods like Anthropic's Constitutional AI, OpenAI's deliberative alignment, and the broader category of AI-assisted feedback, including RLAIF can be used to guide the model to refuse harmful request by evaluating model outputs against constitutional principles like "do not provide information that could directly enable network intrusion" or "prioritize defensive security knowledge over offensive capabilities."	The efficacy of this approach relies on how comprehensive and clearly defined the underlying principles are. Even with RLAIF, AI systems remain susceptible to jailbreaking.

Table 4: Detection and Intervention Mitigations for AI-Cyber

Safeguard	Description	Cyber Application	Limitations
LLM-based Prompted Classifiers	Language models can be prompted to act as classifiers on inputs to and outputs of the model. These classifiers analyze interactions in real time, using both simple methods like keyword detection and more sophisticated semantic analysis to identify potentially dangerous queries or responses. The classifier LLM can be another instance of the same model or a separate model optimized for classification tasks. These LLM-based prompted classifiers are relatively simple to set up and deploy compared to other monitoring approaches.	A prompted LLM could screen incoming queries for patterns indicative of malicious intent (e.g., requests that combine terms like "undetectable," "backdoor," and "critical infrastructure") and outgoing responses for dangerous content (e.g., checking if the model is providing actual exploit code versus theoretical explanations). However, detecting malicious intent in cybersecurity contexts is very difficult. Effective screening often requires examining other contextual clues, such as actor, target type, and whether the activity addresses critical bottlenecks to successful cyber attacks.	Higher input volumes increase computational costs and latency, though using smaller monitoring models reduces expenses at the cost of accuracy. LLM-based classifiers remain vulnerable to circumvention through techniques like decomposing malicious requests into benign-appearing substeps or distributing queries across multiple accounts.
Custom-trained Classifiers	Language models can be trained via fine-tuning to classify inputs and outputs from the model, producing probability scores that indicate whether content contains potential harmful materials. Compared to prompting a language model to act as a classifier, custom-trained classifier models can reduce the cost and latency involved in scaling to many inputs.	Developers may guide training using explicit rules defining permissible and restricted content, such as Anthropic's Constitutional Classifiers. A classifier fine-tuned on thousands of labeled examples of malicious versus benign cybersecurity queries can assess whether a request is seeking exploit development assistance versus legitimate security education and produce a corresponding probability score. The probability scores enable automated decisions, like blocking or allowing responses.	Custom classifiers are more complex and take more time to develop compared to linear probes and introduce more latency at runtime (less than prompted LLM classifiers but more than probes) along with higher computational costs. Cyber's inherent dual-use nature also makes training classifiers for preventing malicious use difficult.
Linear Probes	Linear probes (or simply "probes") are lightweight classifiers trained on a model's internal representations (activations) to efficiently detect specific properties, such as harmful content, without requiring complex architectures or extensive computational resources.	A linear probe could be trained on a language model's internal activations to detect when the model is generating code exploitation techniques. For instance, the probe might monitor the model's internal representations while processing prompts about network security. If the activations indicate the model is about to generate content related to SQL injection attacks, buffer overflow exploits, or ransomware implementation, the probe can flag or block the output before generation completes.	Probes typically require retraining for each model version, may struggle to generalize to novel or complex real-world examples, and involve implementation tradeoffs, such as which model layers to monitor and how to aggregate scores across output tokens, that impact detection accuracy and sensitivity.

Table 4: Detection and Intervention Mitigations for AI-Cyber (Cont'd)

Safeguard	Description	Cyber Application	Limitations
Static Analysis Tools	Static analysis tools are deterministic systems deployed to validate and secure LLM outputs through pattern matching and code inspection. These tools offer low-latency, consistent checking of output formats and prohibited content patterns.	CodeShield is a static analysis tool that detects insecure code patterns in LLM-generated code at inference, identifying vulnerabilities like weak cryptographic functions and other Common Weakness Enumeration or CWE classified security issues.	While faster and more predictable than AI-powered alternatives, static analysis tools may miss nuanced or context-specific issues and are most effective when combined with other detection methods for comprehensive coverage.
Manual and Automated Abuse Monitoring	Monitoring user interaction patterns over time to detect suspicious activity indicative of attempts to misuse the model, such as repeated probing for sensitive information or trying to circumvent safety filters.	Abuse monitoring can identify users systematically trying to gather information needed to develop cyber attacks, even if individual prompts don't trigger input filters. For example, a system might track behavior patterns, such as a user making a high number of queries in an hour about progressively more specific exploit techniques.	Manual review is inherently slow and difficult to scale compared to automated methods. Defining and detecting "abusive" patterns accurately without violating user privacy or flagging benign research behaviour is challenging.

Table 5: Access Control Mitigations for AI-Cyber

Safeguard	Description	Cyber Application	Limitations
Staged Deployment	Developers may implement staged rollouts for new models, starting with highly controlled environments and gradually expanding access as controls are validated. Initial deployment could involve small groups of external users or research partners operating under strict monitoring agreements, allowing developers to observe model behavior and identify potential risks before broader release.	Staged deployment may involve initially limiting access to vetted cybersecurity professionals for research purposes, then gradually broadening availability to a wider group.	Setting appropriate vetting requirements for access, monitoring usage as access expands, and ensuring appropriate security measures are all challenges for staged deployment. There are also tradeoffs around establishing vetting requirements, as smaller organizations may not have the resources or teams to meet stringent vetting requirements, and may lose access to beneficial AI.

Table 6: Ecosystem Mitigations for AI-Cyber

Safeguard	Description	Cyber Application	Limitations
Secure Information-Sharing Networks	Establishing trusted communication channels and protocols for relevant stakeholders to securely share threat intelligence and cybersecurity information.	Information-sharing on how models are misused by cyber actors once deployed can be a valuable source of information for updating safety mitigations.	Building and maintaining trusted communication channels between diverse organizations, especially across public and private sectors or internationally, is a significant hurdle. Information sharing must also carefully navigate legal frameworks, privacy concerns, and commercial sensitivities to be effective.
Defensive Systems and Research	Developers, other actors in the AI ecosystem, and society broadly can develop or support the development of systems specifically designed to strengthen defensive capabilities, including the use of frontier models. Furthermore, open source models, artifacts, and defensive tooling can serve as a catalyst for exploration in the security ecosystem, strengthening and advancing defender uses. Some developers are also building and sharing open source tools and measurement suites for identifying malicious activities from potential malware, extracting insights from threat intelligence reports, and examining how AI models help automatically patch vulnerable systems.	Developers are already leveraging frontier AI cyber capabilities to search and find unknown security vulnerabilities in software, including in widely used open-source projects. Developers could support the development of vulnerability detection tools for critical infrastructure.	Advanced AI defensive capabilities may only be available to well-resourced organizations, creating security disparities and leaving gaps in protection.
Programs for Cyberdefense	Some developers are piloting trusted access programs or partnerships that grant qualifying users working on cyberdefense tiered access to enhanced capabilities to be used in defensive use cases.	Through vetted partnerships, developers enable defensive applications of AI, such as emulating cyber attacks on water treatment plants, to be used to improve the security of critical infrastructure.	Determining who qualifies as a trusted cybersecurity user is complex, as credentials can be fabricated, affiliations may change, insider threats exist within legitimate organizations, and manual vetting processes are difficult to scale.

Section 4

AI-Cyber Mitigation Assessments

After implementing safeguards, developers must verify that these mitigations effectively reduce risks to acceptable levels under realistic operating conditions. AI-cyber mitigation assessments test the robustness of technical interventions put in place by frontier AI developers to prevent users from using AI tools for malicious cyber operations. For example, mitigation assessments test how well the technical controls, such as refusal policies and input and output classifiers, prevent cyber misuse of AI systems.

Comprehensive mitigation assessments require extensive internal testing by safety teams and in some cases may involve external evaluation by third-party organizations and researchers. These assessments can take two approaches: testing individual safeguards to evaluate specific components, or testing multiple mitigations together to assess how they function as an integrated defensive system. Assessing individual controls involves evaluating specific technical properties, such as testing the precision and recall of classifier models or the effectiveness of behavioral guardrails against prompt variations. By contrast, system-level testing evaluates the robustness of all safeguards working together. This integrated approach provides a more realistic picture of how the system will perform under actual deployment conditions and can reveal emergent vulnerabilities that may not appear when components are tested in isolation, even when each component functions correctly on its own.

Approaches to evaluating AI-cyber mitigations include:

- **Empirical Testing.** Teams within the organization conduct structured testing using knowledge of model architecture and training methods. This approach benefits from institutional knowledge but may suffer from blind spots due to organizational biases. This process, often referred to as “safeguard evaluations,” involves testing model responses to potentially harmful prompt requests. Developers create test prompts to evaluate whether model responses align with organizational policies. Tests prompts could include requests the model should refuse entirely, queries requiring cautious or limited responses, and prompts where full compliance aligns with the organization's acceptable output policies. To ensure comprehensive coverage of model behavior, these prompts should target critical cyber domains, including code development, infrastructure and tooling creation, vulnerability exploitation techniques, among others. To ensure more robust testing, evaluators can also test these prompts multiple times, in combination with jailbreaks, and as part of multi-turn interactions.

  In some cases, firms work with third party assessors to evaluate the safety of their models. Third party assessors can provide additional coverage by developing safeguard evaluations and may apply a different level of depth on evaluating domain areas compared to internal teams. Third party assessors may leverage publicly available benchmarks or privately develop harmful questions on cyber capabilities unlikely to be needed for benign use. Assessors can then grade whether a model complies with a harmful request and if the model provides a harmful response.

- **Red Teaming.** Effective testing requires testers with both red teaming expertise and specialized domain knowledge. This combination enables testers to identify sophisticated circumvention attempts that rely on technical knowledge, such as using obfuscated code patterns or indirect references to exploitation techniques to bypass cyber safeguards. Testing should occur on systems as they will be deployed. This approach ensures that test results accurately reflect the model's real-world defensive performance rather than theoretical capabilities.

  External teams, such as specialized cybersecurity firms that conduct adversarial testing, can offer distinct advantages. They bring established testing frameworks and cross-organizational experience from evaluating multiple models across firms, providing a broader perspective on emerging attack patterns and defensive gaps. They can also be a resource for firms that lack internal expertise.

- **Ongoing Deployment Monitoring.** Relative to capability benchmarks, mitigation assessments may maintain their usefulness over a longer period of time because they evaluate the robustness of the safeguards whereas capability benchmarks can become saturated as raw capabilities improve. Mitigation assessments can maintain their usefulness over time because they evaluate the robustness of the safeguards rather than the raw capabilities of the models. However, developers should continuously monitor mitigation effectiveness after deployment, since safeguard performance can decline as attackers develop new circumvention techniques. Operational metrics might include detection system accuracy rates, trends in attack complexity, and intervention success rates when blocking malicious requests, such as malicious code generation. These metrics enable developers to identify when protections need strengthening and deploy timely updates to maintain defensive capabilities.

Section 5

Continuing Work

REQUEST FOR COMMENT

We welcome engagement with this report and our broader technical report series on frontier AI frameworks.

Researchers and organizations interested in further refining and harmonizing the implementation of frontier AI frameworks are invited to reach out to the Frontier Model Forum.

Please offer feedback at: info@frontiermodelforum.org

This report outlines current developer approaches to evaluating and mitigating offensive cyber capabilities in frontier AI models. As model capabilities advance and our understanding of cyber risks improves, these frameworks will continue to evolve. Ongoing work spans several interconnected areas:

Cyber Capability Assessment and Threat Modeling. Developers are improving cyber threat modeling techniques and establishing more robust methodologies for assessing potential threat scenarios. As model capabilities advance, assessment methods must evolve to detect increasingly sophisticated capabilities, including developing automated evaluations and training metrics that identify concerning cyber capabilities earlier in the development process, and addressing challenges like models that might conceal their capabilities during evaluation. Future advances in model interpretability and transparency methods hold promise for deeper insights into how models acquire and apply cyber knowledge. Collaboration remains essential for developing standardized cyber evaluation frameworks, shared red-teaming methodologies, and common benchmarks.

Mitigation Design and Validation. This includes developing more robust safety mechanisms that resist jailbreaking and prompt injection attacks, exploring techniques to limit models' ability to reason about or execute offensive cyber operations, and establishing best practices for different deployment contexts. Critical needs include standardized approaches for testing how well safeguards withstand adversarial pressure, empirical validation of mitigation effectiveness against realistic attack scenarios, and security measures proportional to the sophistication of potential threat actors and model capabilities.

Risk-Benefit Tradeoffs. Determining acceptable levels of cyber-related risk from frontier AI extends beyond technical considerations to encompass broader questions about dual-use capabilities, defensive versus offensive applications, and societal security needs. While legitimate cybersecurity applications exist, the potential for misuse by malicious actors requires careful threshold-setting. Developers can provide transparency about their cyber capability thresholds and decision processes, but the ultimate determination of acceptable risk-benefit tradeoffs requires input from cybersecurity experts, policymakers, diverse stakeholders, and appropriate governance structures. Future research should focus on developing evidence-based approaches for continuous calibration of cyber risk management strategies that remain responsive to both the evolving threat landscape and societal expectations around AI-enabled cyber risk.

FOOTNOTES

[1] The Frontier AI Safety Commitments are voluntary commitments that leading frontier AI developers globally signed on to at the May 2024 AI Seoul Summit. Under these commitments, major AI companies committed to identify and manage intolerable risks from advanced models and to publicly document how they intend to fulfill their commitments.

[2] Bottleneck Assessments test whether models possess specific capabilities that domain experts believe would remove “bottlenecks” to severe real-world harm.

APPENDIX

See below for a non-comprehensive list of public documents that reference AI safety evaluations of offensive cyber risks.

Table 7: Publicly Disclosed AI-Cyber Safety Evaluations

Year	Author	Name	Method	URL
2025	Amazon	Evaluating Nova 2.0 Lite model under Amazon's Frontier Model Safety Framework	CTF Exercise, Capability Benchmark, Knowledge Benchmark	Model Card
2025	Anthropic	Claude Opus 4 & Sonnet 4	CTF Exercise, Cyber Range Exercise	Model Card
2025	Google	Gemini 3 Pro Frontier Safety Framework Report	CTF Exercise	Model Card
2025	Google	Gemini 2.5 Pro Preview Model Card	CTF Exercise, Knowledge Benchmark	Model Card
2025	Google	A Framework for Evaluating Emerging Cyberattack Capabilities of AI	CTF Exercise	Paper
2025	Kouremetis et al.	OCCULT: Evaluating Large Language Models for Offensive Cyber Operation Capabilities	Knowledge Benchmark	Paper
2025	Meta	Code World Model Preparedness Report	Capability Benchmark	Paper
2025	OpenAI	Deep Research System Card	CTF Exercise, Cyber Range	Model Card
2025	Singer et al.	On the Feasibility of Using LLMs to Execute Multistage Network Attacks	Cyber Range Exercise	Paper
2025	Wang et al.	CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale	Capability Benchmark	Paper
2025	Zhu et al.	CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities	Capability Benchmark	Paper
2025	Zhun et al.	CyberGym: Evaluating AI Agents' Cybersecurity Capabilities with Real-World Vulnerabilities at Scale	Capability Benchmark	Paper
2024	Anthropic	The Claude 3 Model Family: Opus, Sonnet, Haiku	CTF Exercise	Model Card

Year	Author	Name	Method	URL
2024	Anurin et al.	Catastrophic Cyber Capabilities Benchmark (3CB): Robustly Evaluating LLM Agent Cyber Offense Capabilities	Capability Benchmark	Paper
2024	Bhatt et al.	CYBERSECEVAL 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models	Safeguard Evaluation, Capability Benchmark	Paper
2024	Carnegie Mellon University	Pico CTFs	CTF Exercise	GitHub
2024	Google	Gemini Technical Report	CTF Exercise, Red-Team Exercises	Model Card
2024	Guo et al.	RedCode: Risky Code Execution and Generation Benchmark for Code Agents	Safeguard Evaluation, Capability Benchmark	Paper
2024	Jing et al.	SecBench: A Comprehensive Multi-Dimensional Benchmarking Dataset for LLMs in Cybersecurity	Knowledge Benchmark	Paper
2024	Li and Pan et al.	The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning	Knowledge Benchmark	Paper
2024	Liu et al.	OpsEval: A Comprehensive Benchmark Suite for Evaluating Large Language Models' Capability in IT Operations Domain	Knowledge Benchmark, Cyber Range Exercise	Paper
2024	Liu et al.	CyberBench/CyberMetric: A Multi-Task Benchmark for Evaluating Large Language Models in Cybersecurity	Knowledge Benchmark	Paper
2024	Meta	The Llama 3 Herd of Models	Uplift Study, Cyber Range Exercise, CTF Exercise, Capability Benchmark, Safeguard Evaluation	Model Card
2024	Miao et al.	An Empirical Study of Netops Capability of Pre-trained Large Language Models	Knowledge Benchmark	Paper

Year	Author	Name	Method	URL
2024	OpenAI	OpenAI o1 System Card	CTF Exercise, Red-Team Exercise	Model Card
2024	Phuong et al.	Evaluating Frontier Models for Dangerous Capabilities	CTF Exercise, Knowledge Benchmark	Paper
2024	Ristea et al.	AI Cyber Risk Benchmark: Automated Exploitation Capabilities	Capability Benchmark	Paper
2024	Shao et al.	NYU CTF Dataset: A Scalable Open-Source Benchmark Dataset for Evaluating LLMs in Offensive Security	CTF Exercise	Paper
2024	Shao et al.	An Empirical Evaluation of LLMs for Solving Offensive Security Challenges	CTF Exercise	Paper
2024	Tian and Ye et al.	DebugBench: Evaluating Debugging Capability of Large Language Models	Knowledge Benchmark	Paper
2024	UK AISI	Advanced AI evaluations at AISI	CTF Exercise	Blog Post
2024	US AISI & UK AISI	US AISI and UK AISI Joint Pre-Deployment Test – OpenAI o1	Cyber Range Exercise, CTF Exercise, Capability Benchmark	Report
2024	Wan et al.	CYBERSECEVAL 3: Advancing the Evaluation of Cybersecurity Risks and Capabilities in Large Language Models	Uplift Study, Cyber Range Exercise, CTF Exercise, Capability Benchmark, Safeguard Evaluation	Paper
2024	XBOW	XBOW Validation Benchmarks	CTF Exercise	Github
2024	Yang et al.	SecCodePLT: A Unified Platform for Evaluating the Security of Code GenAI	Capability Benchmark	Paper
2024	Zhang et al.	Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risk of Language Models	Cyber Range Exercise, CTF Exercise	Paper
2023	Bhatt et al.	Purple Llama CYBERSECEVAL: A Secure Coding Benchmark for Language Models	Safeguard Evaluation	Paper
2023	Garza et al.	Assessing Large Language Model’s knowledge of threat behavior in MITRE ATT&CK	Knowledge Benchmark	Paper

Year	Author	Name	Method	URL
2023	Li et al.	SecEval: A Comprehensive Benchmark for Evaluating Cybersecurity Knowledge of Foundation Models	Knowledge Benchmark	Github
2023	Liu et al.	SecQA: A Concise Question-Answering Dataset for Evaluating Large Language Models in Computer Security	Knowledge Benchmark	Paper
2023	Moskal et al.	LLMs Killed the Script Kiddie: How Agents Supported by Large Language Models Change the Landscape of Network Threat Testing	Knowledge Benchmark	Paper
2023	Tann and Liu et al.	Using Large Language Models for Cybersecurity Capture-The-Flag Challenges and Certification Questions	CTF Exercise	Paper
2023	Tony et al.	LLMSecEval: A Dataset of Natural Language Prompts for Security Evaluations	Knowledge Benchmark	Paper
2023	Yang et al.	InterCode: Standardizing and Benchmarking Interactive Coding with Execution Feedback	Capability Benchmark	Paper