ETSI Security Conference 2025

Securing the Open AI Frontier

Alfonso De Gregorio

09/10/25

Open-weight Models and Offensive Cyber Operations

OCCULT: Evaluating Large Language Models for Offensive Cyber Operation Capabilities

Michael Kouremetis*, Marissa Dotter*, Alex Byrne, Dan Martin, Ethan Michalak, Gianpaolo Russo, Michael Threet, and Guido Zarrella

The MITRE Corporation, McLean, VA, USA

*Principal investigators and corresponding authors: mkouremetis@mitre.org, mdotter@mitre.org

February 25, 2025

Abstract

The prospect of artificial intelligence (AI) competing in the adversarial landscape of cyber security has long been considered one of the most impactful, challenging, and potentially dangerous applications of AI. Here, we demonstrate a new approach to assessing AI's progress towards enabling and scaling real-world offensive cyber operations (OCO) tactics in use by modern threat actors. We detail OCCULT, a lightweight operational evaluation framework that allows cyber security experts to contribute to rigorous and repeatable measurement of the plausible cyber security risks associated with any given large language model (LLM) or AI employed for OCO. We also prototype and evaluate three very different OCO benchmarks for LLMs that demonstrate our approach and serve as examples for building benchmarks under the OCCULT framework. Finally, we provide preliminary evaluation results to demonstrate how this framework allows us to move beyond traditional all-or-nothing tests, such as those crafted from educational exercises like capture-the-flag environments, to contextualize our indicators and warnings in true cyber threat scenarios that present risks to modern infrastructure. We find that there has been significant recent advancement in the risks of AI being used to scale realistic cyber threats. For the first time, we find a model (DeepSeek-R1) is capable of correctly answering over 90% of challenging offensive cyber knowledge tests in our Threat Actor Competency Test for LLMs (TACTL) multiple-choice benchmarks. We also show how Meta's Llama and Mistral's Mixtral model families show marked performance improvements over earlier models against our benchmarks where LLMs act as offensive agents in MITRE's high-fidelity offensive and defensive cyber operations simulation environment, *CyberLayer*.

Keywords: Offensive Cyber, Large Language Models, Autonomous Cyber Operations, MITRE

1 Introduction

The growing capabilities of Large Language Models (LLMs) in synthesizing knowledge, analyzing code, and utilizing software have raised concerns about their potential for misuse in various critical domains when in service to their users. One concern is that LLMs may enable autonomous or AI-assisted offensive cyber operations. Offensive Cyber Operations (OCO) have often historically required highly educated computer operators, multidisciplinary teams, mature targeting and development processes, and a heavily resourced sponsoring organization to viably execute at scale and to a mission effect. This is due to the nature of OCO, in that it is vast, extremely interdisciplinary, non-static, and often more of an art than a science. Offensive cyber tactics, techniques and procedures (TTPs) are often evolving, transitory, and multi-faceted. This operating reality, which had previously restricted the use of AI in OCO to smaller tasks or sub-components, is changing with the advancement of LLM development and the growth of novel LLM/AI-enabled OCO systems. Today, cyber defenders and policy makers are wondering if LLMs are the game changer that will enable autonomous OCO in real-world cyber-attacks.

To mitigate the potential risk that autonomous and even semi-autonomous AI-enabled OCO systems could pose, we must be able to evaluate the true capabilities of any emerging OCO AI

Agenda

- *The Mitigation Gap*
- *Evolving Policy & Accountability*
- *The Path Forward: A Multi-Faceted Approach*
- *Conclusion & Call to Action*

The Mitigation Gap

Cyber Policy When Control on AI is Lost

- API-based models: Monitoring, filtering, rate-limiting, implementing guardrails.
- Open-weight models: Control is gone.
- Our traditional security playbook becomes obsolete.

An Ineffective Security Playbook

- API-based safeguards? Irrelevant when the model runs locally.
- Watermarking to trace malicious outputs? It can often be modified or removed by a determined actor.
- Secure enclaves? Bypassed entirely.
- Safety alignments—the "guardrails" carefully trained into these models to prevent harmful behavior—can often be trivially stripped away through fine-tuning, a process sometimes called "model unfiltering."

Evolving Policy & Accountability

- EU AI Act initial drafts based on the old centralised control paradigm.
- Ambiguity about the open-source definition.
- Imposed obligations (e.g. post-market monitoring) impossible to fulfil.
- Risk of stifling innovation with impossible compliance burdens or creating loopholes.

European Commission's Public Consultation

Contribution ID: 168512a9-e480-48b5-b561-a5a6718b70ae
Date: 08/06/2025 16:57:40

Targeted stakeholder consultation on classification of AI systems as high-risk

Fields marked with * are mandatory.

Targeted stakeholder consultation on the implementation of the AI Act's rules for high-risk AI systems

Disclaimer: This document is a working document of the AI Office for the purpose of consultation and does not prejudge the final decision that the Commission may take on the final guidelines. The responses to this consultation paper will provide important input to the Commission when preparing the guidelines.

This consultation is targeted to stakeholders of different categories. These categories include, but are not limited to, providers and deployers of (high-risk) AI systems, other industry organisations, as well as academia, other independent experts, civil society organisations, and public authorities.

The Artificial Intelligence Act (the 'AI Act')[1], which entered into force on 1 August 2024, creates a single market and harmonised rules for trustworthy and human-centric Artificial Intelligence (AI) in the EU.[2] It aims to promote innovation and uptake of AI, while ensuring a high level of protection of health, safety and fundamental rights, including democracy and the rule of law. The AI Act follows a risk-based approach classifying AI systems into different risk categories, one of which is the high-risk AI systems (Chapter III of the AI Act). The relevant obligations for those systems will be applicable two years after the entry into force of the AI Act, as from 2 August 2026.

The AI Act distinguishes between two categories of AI systems that are considered as 'high-risk' set out in Article 6(1) and 6(2) AI Act. Article 6(1) AI Act covers AI systems that are embedded as safety components in products or that themselves are products covered by Union legislation in Annex I, which could have an adverse impact on health and safety of persons. Article 6(2) AI Act covers AI systems that in view of their intended purpose are considered to pose a significant risk to health, safety or fundamental rights. The AI Act lists eight areas in which AI systems could pose such significant risk to health, safety or fundamental rights in Annex III and, within each area, lists specific use-cases that are to be classified as high-risk. Article 6(3) AI Act provides for exemptions for AI systems that are intended to be used for one of the cases listed in Annex III, but which do not pose significant risk since they fall under one of the exceptions listed in Article 6(3).

Substantial Modification & Proportional Accountability

- Downstream actors becomes provider, in the event they remove guardrails and safety measures.
- Accountability and liability shifts from the original innovator to the downstream deployer or modifier.
- Loopholes are closed, without punishing responsible open development.

The Path Forward: A Multi-Faceted Approach

- Shift our focus from controlling models to evaluating capabilities;
- Use AI to fight AI;
- Urgent need for int'l standardisation.

From Controlling Models to Evaluating Capabilities

What dangerous things can this model do?

AI to Fight AI

From AI-Powered Threat Detection to Automated Incident Response at Machine Speed

Int'l Standardisation

- Standardised Evaluation: Creating a common methodology, perhaps an "Offensive AI Capability Index," so that a risk assessment in one country is understood and trusted in another.
- Secure Release Practices: Defining best practices for open-weight model releases, including mandatory transparency through detailed Model Cards and Risk Assessments.
- Enhanced CTI Sharing: Developing new protocols for Cyber Threat Intelligence that are specifically tailored to AI-driven threats, allowing us to share indicators of AI-powered attacks in near real-time.

Downstream Regulatory Focus

Conclusions & *Call to Action*

- Pragmatic Equilibrium
- An ecosystem where innovation continues to be open, but with guardrails and accountability
- Let’s empower defenders in the same way we are doing with attackers
- A collective effort

Conclusions & *Call to Action*

- Pragmatic Equilibrium
- An ecosystem where innovation continues to be open, but with guardrails and accountability
- Let’s empower defenders in the same way we are doing with attackers
- A collective effort