Staying ahead of threat actors in the age of AI

microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai

February 14, 2024

By Microsoft Threat Intelligence

Over the last year, the speed, scale, and sophistication of attacks has increased alongside the rapid development and adoption of AI. Defenders are only beginning to recognize and apply the power of generative AI to shift the cybersecurity balance in their favor and keep ahead of adversaries. At the same time, it is also important for us to understand how AI can be potentially misused in the hands of threat actors. In collaboration with OpenAI, today we are publishing research on emerging threats in the age of AI, focusing on identified activity associated with known threat actors, including prompt-injections, attempted misuse of large language models (LLM), and fraud. Our analysis of the current use of LLM technology by threat actors revealed behaviors consistent with attackers using AI as another productivity tool on the offensive landscape. You can read OpenAI's blog on the research here. Microsoft and OpenAI have not yet observed particularly novel or unique AI-enabled attack or abuse techniques resulting from threat actors' usage of AI. However, Microsoft and our partners continue to study this landscape closely.

The objective of Microsoft's partnership with OpenAI, including the release of this research, is to ensure the safe and responsible use of AI technologies like ChatGPT, upholding the highest standards of ethical application to protect the community from potential misuse. As part of this commitment, we have taken measures to disrupt assets and accounts associated with threat actors, improve the protection of OpenAI LLM technology and users from attack or abuse, and shape the guardrails and safety mechanisms around our models. In addition, we are also deeply committed to using generative AI to disrupt threat actors and leverage the power of new tools, including Microsoft Copilot for Security, to elevate defenders everywhere.

A principled approach to detecting and blocking threat actors

The progress of technology creates a demand for strong cybersecurity and safety measures. For example, the White House's Executive Order on AI requires rigorous safety testing and government supervision for AI systems that have major impacts on national and economic security or public health and safety. Our actions enhancing the safeguards of our AI models and partnering with our ecosystem on the safe creation, implementation, and use of these models align with the Executive Order's request for comprehensive AI safety and security standards.

In line with Microsoft's leadership across AI and cybersecurity, today we are announcing principles shaping Microsoft's policy and actions mitigating the risks associated with the use of our AI tools and APIs by nation-state advanced persistent threats (APTs), advanced persistent manipulators (APMs), and cybercriminal syndicates we track.

These principles include:

Identification and action against malicious threat actors' use: Upon detection of the use of any Microsoft AI application programming interfaces (APIs), services, or systems by an identified malicious threat actor, including nation-state APT or APM, or the cybercrime syndicates we track, Microsoft will take appropriate action to disrupt their activities, such as disabling the accounts used, terminating services, or limiting access to resources.

Notification to other AI service providers: When we detect a threat actor's use of another service provider's AI, AI APIs, services, and/or systems, Microsoft will promptly notify the service provider and share relevant data. This enables the service provider to independently verify our findings and take action in accordance with their own policies.

Collaboration with other stakeholders: Microsoft will collaborate with other stakeholders to regularly exchange information about detected threat actors' use of AI. This collaboration aims to promote collective, consistent, and effective responses to ecosystem-wide risks.

Transparency: As part of our ongoing efforts to advance responsible use of AI, Microsoft will inform the public and stakeholders about actions taken under these threat actor principles, including the nature and extent of threat actors' use of AI detected within our systems and the measures taken against them, as appropriate.

Microsoft remains committed to responsible AI innovation, prioritizing the safety and integrity of our technologies with respect for human rights and ethical standards. These principles announced today build on Microsoft's Responsible AI practices, our voluntary commitments to advance responsible AI innovation and the Azure OpenAI Code of Conduct. We are following these principles as part of our broader commitments to strengthening international law and norms and to advance the goals of the Bletchley Declaration endorsed by 29 countries.

Microsoft and OpenAI's complementary defenses protect AI platforms

Because Microsoft and OpenAI's partnership extends to security, the companies can take action when known and emerging threat actors surface. Microsoft Threat Intelligence tracks more than 300 unique threat actors, including 160 nation-state actors, 50 ransomware groups, and many others. These adversaries employ various digital identities and attack infrastructures. Microsoft's experts and automated systems

continually analyze and correlate these attributes, uncovering attackers' efforts to evade detection or expand their capabilities by leveraging new technologies. Consistent with preventing threat actors' actions across our technologies and working closely with partners, Microsoft continues to study threat actors' use of AI and LLMs, partner with OpenAI to monitor attack activity, and apply what we learn to continually improve defenses. This blog provides an overview of observed activities collected from known threat actor infrastructure as identified by Microsoft Threat Intelligence, then shared with OpenAI to identify potential malicious use or abuse of their platform and protect our mutual customers from future threats or harm.

Recognizing the rapid growth of AI and emergent use of LLMs in cyber operations, we continue to work with MITRE to integrate these LLM-themed tactics, techniques, and procedures (TTPs) into the MITRE ATT&CK® framework or MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems) knowledgebase. This strategic expansion reflects a commitment to not only track and neutralize threats, but also to pioneer the development of countermeasures in the evolving landscape of AI-powered cyber operations. A full list of the LLM-themed TTPs, which include those we identified during our investigations, is summarized in the appendix.

Summary of Microsoft and OpenAI's findings and threat intelligence

The threat ecosystem over the last several years has revealed a consistent theme of threat actors following trends in technology in parallel with their defender counterparts. Threat actors, like defenders, are looking at AI, including LLMs, to enhance their productivity and take advantage of accessible platforms that could advance their objectives and attack techniques. Cybercrime groups, nation-state threat actors, and other adversaries are exploring and testing different AI technologies as they emerge, in an attempt to understand potential value to their operations and the security controls they may need to circumvent. On the defender side, hardening these same security controls from attacks and implementing equally sophisticated monitoring that anticipates and blocks malicious activity is vital.

While different threat actors' motives and complexity vary, they have common tasks to perform in the course of targeting and attacks. These include reconnaissance, such as learning about potential victims' industries, locations, and relationships; help with coding, including improving things like software scripts and malware development; and assistance with learning and using native languages. Language support is a natural feature of LLMs and is attractive for threat actors with continuous focus on social engineering and other techniques relying on false, deceptive communications tailored to their targets' jobs, professional networks, and other relationships.

Importantly, our research with OpenAI has not identified significant attacks employing the LLMs we monitor closely. At the same time, we feel this is important research to publish to expose early-stage, incremental moves that we observe well-known threat actors

attempting, and share information on how we are blocking and countering them with the defender community.

While attackers will remain interested in AI and probe technologies' current capabilities and security controls, it's important to keep these risks in context. As always, hygiene practices such as multifactor authentication (MFA) and Zero Trust defenses are essential because attackers may use AI-based tools to improve their existing cyberattacks that rely on social engineering and finding unsecured devices and accounts.

The threat actors profiled below are a sample of observed activity we believe best represents the TTPs the industry will need to better track using MITRE ATT&CK® framework or MITRE ATLAS™ knowledgebase updates.

Forest Blizzard

Forest Blizzard (STRONTIUM) is a Russian military intelligence actor linked to GRU Unit 26165, who has targeted victims of both tactical and strategic interest to the Russian government. Their activities span across a variety of sectors including defense, transportation/logistics, government, energy, non-governmental organizations (NGO), and information technology. Forest Blizzard has been extremely active in targeting organizations in and related to Russia's war in Ukraine throughout the duration of the conflict, and Microsoft assesses that Forest Blizzard operations play a significant supporting role to Russia's foreign policy and military objectives both in Ukraine and in the broader international community. Forest Blizzard overlaps with the threat actor tracked by other researchers as APT28 and Fancy Bear.

Forest Blizzard's use of LLMs has involved research into various satellite and radar technologies that may pertain to conventional military operations in Ukraine, as well as generic research aimed at supporting their cyber operations. Based on these observations, we map and classify these TTPs using the following descriptions:

- **LLM-informed reconnaissance:** Interacting with LLMs to understand satellite communication protocols, radar imaging technologies, and specific technical parameters. These queries suggest an attempt to acquire in-depth knowledge of satellite capabilities.
- **LLM-enhanced scripting techniques:** Seeking assistance in basic scripting tasks, including file manipulation, data selection, regular expressions, and multiprocessing, to potentially automate or optimize technical operations.

Similar to Salmon Typhoon's LLM interactions, Microsoft observed engagement from Forest Blizzard that were representative of an adversary exploring the use cases of a new technology. As with other adversaries, all accounts and assets associated with Forest Blizzard have been disabled.

Emerald Sleet

Emerald Sleet (THALLIUM) is a North Korean threat actor that has remained highly active throughout 2023. Their recent operations relied on spear-phishing emails to compromise and gather intelligence from prominent individuals with expertise on North Korea. Microsoft observed Emerald Sleet impersonating reputable academic institutions and NGOs to lure victims into replying with expert insights and commentary about foreign policies related to North Korea. Emerald Sleet overlaps with threat actors tracked by other researchers as Kimsuky and Velvet Chollima.

Emerald Sleet's use of LLMs has been in support of this activity and involved research into think tanks and experts on North Korea, as well as the generation of content likely to be used in spear-phishing campaigns. Emerald Sleet also interacted with LLMs to understand publicly known vulnerabilities, to troubleshoot technical issues, and for assistance with using various web technologies. Based on these observations, we map and classify these TTPs using the following descriptions:

- **LLM-assisted vulnerability research:** Interacting with LLMs to better understand publicly reported vulnerabilities, such as the CVE-2022-30190 Microsoft Support Diagnostic Tool (MSDT) vulnerability (known as "Follina").
- **LLM-enhanced scripting techniques:** Using LLMs for basic scripting tasks such as programmatically identifying certain user events on a system and seeking assistance with troubleshooting and understanding various web technologies.
- **LLM-supported social engineering:** Using LLMs for assistance with the drafting and generation of content that would likely be for use in spear-phishing campaigns against individuals with regional expertise.
- **LLM-informed reconnaissance:** Interacting with LLMs to identify think tanks, government organizations, or experts on North Korea that have a focus on defense issues or North Korea's nuclear weapon's program.

All accounts and assets associated with Emerald Sleet have been disabled.

Crimson Sandstorm

Crimson Sandstorm (CURIUM) is an Iranian threat actor assessed to be connected to the Islamic Revolutionary Guard Corps (IRGC). Active since at least 2017, Crimson Sandstorm has targeted multiple sectors, including defense, maritime shipping, transportation, healthcare, and technology. These operations have frequently relied on watering hole attacks and social engineering to deliver custom .NET malware. Prior research also identified custom Crimson Sandstorm malware using email-based command-and-control (C2) channels. Crimson Sandstorm overlaps with the threat actor tracked by other researchers as Tortoiseshell, Imperial Kitten, and Yellow Liderc.

The use of LLMs by Crimson Sandstorm has reflected the broader behaviors that the security community has observed from this threat actor. Interactions have involved requests for support around social engineering, assistance in troubleshooting errors, .NET development, and ways in which an attacker might evade detection when on a compromised machine. Based on these observations, we map and classify these TTPs using the following descriptions:

- **LLM-supported social engineering:** Interacting with LLMs to generate various phishing emails, including one pretending to come from an international development agency and another attempting to lure prominent feminists to an attacker-built website on feminism.
- **LLM-enhanced scripting techniques:** Using LLMs to generate code snippets that appear intended to support app and web development, interactions with remote servers, web scraping, executing tasks when users sign in, and sending information from a system via email.
- **LLM-enhanced anomaly detection evasion:** Attempting to use LLMs for assistance in developing code to evade detection, to learn how to disable antivirus via registry or Windows policies, and to delete files in a directory after an application has been closed.

All accounts and assets associated with Crimson Sandstorm have been disabled.

Charcoal Typhoon

Charcoal Typhoon (CHROMIUM) is a Chinese state-affiliated threat actor with a broad operational scope. They are known for targeting sectors that include government, higher education, communications infrastructure, oil & gas, and information technology. Their activities have predominantly focused on entities within Taiwan, Thailand, Mongolia, Malaysia, France, and Nepal, with observed interests extending to institutions and individuals globally who oppose China's policies. Charcoal Typhoon overlaps with the threat actor tracked by other researchers as Aquatic Panda, ControlX, RedHotel, and BRONZE UNIVERSITY.

In recent operations, Charcoal Typhoon has been observed interacting with LLMs in ways that suggest a limited exploration of how LLMs can augment their technical operations. This has consisted of using LLMs to support tooling development, scripting, understanding various commodity cybersecurity tools, and for generating content that could be used to social engineer targets. Based on these observations, we map and classify these TTPs using the following descriptions:

- **LLM-informed reconnaissance**: Engaging LLMs to research and understand specific technologies, platforms, and vulnerabilities, indicative of preliminary information-gathering stages.
- **LLM-enhanced scripting techniques**: Utilizing LLMs to generate and refine scripts, potentially to streamline and automate complex cyber tasks and operations.

- **LLM-supported social engineering**: Leveraging LLMs for assistance with translations and communication, likely to establish connections or manipulate targets.
- **LLM-refined operational command techniques**: Utilizing LLMs for advanced commands, deeper system access, and control representative of post-compromise behavior.

All associated accounts and assets of Charcoal Typhoon have been disabled, reaffirming our commitment to safeguarding against the misuse of AI technologies.

Salmon Typhoon

Salmon Typhoon (SODIUM) is a sophisticated Chinese state-affiliated threat actor with a history of targeting US defense contractors, government agencies, and entities within the cryptographic technology sector. This threat actor has demonstrated its capabilities through the deployment of malware, such as Win32/Wkysol, to maintain remote access to compromised systems. With over a decade of operations marked by intermittent periods of dormancy and resurgence, Salmon Typhoon has recently shown renewed activity. Salmon Typhoon overlaps with the threat actor tracked by other researchers as APT4 and Maverick Panda.

Notably, Salmon Typhoon's interactions with LLMs throughout 2023 appear exploratory and suggest that this threat actor is evaluating the effectiveness of LLMs in sourcing information on potentially sensitive topics, high profile individuals, regional geopolitics, US influence, and internal affairs. This tentative engagement with LLMs could reflect both a broadening of their intelligence-gathering toolkit and an experimental phase in assessing the capabilities of emerging technologies.

Based on these observations, we map and classify these TTPs using the following descriptions:

- **LLM-informed reconnaissance:** Engaging LLMs for queries on a diverse array of subjects, such as global intelligence agencies, domestic concerns, notable individuals, cybersecurity matters, topics of strategic interest, and various threat actors. These interactions mirror the use of a search engine for public domain research.
- **LLM-enhanced scripting techniques:** Using LLMs to identify and resolve coding errors. Requests for support in developing code with potential malicious intent were observed by Microsoft, and it was noted that the model adhered to established ethical guidelines, declining to provide such assistance.
- **LLM-refined operational command techniques:** Demonstrating an interest in specific file types and concealment tactics within operating systems, indicative of an effort to refine operational command execution.
- **LLM-aided technical translation and explanation:** Leveraging LLMs for the translation of computing terms and technical papers.

Salmon Typhoon's engagement with LLMs aligns with patterns observed by Microsoft, reflecting traditional behaviors in a new technological arena. In response, all accounts and assets associated with Salmon Typhoon have been disabled.

In closing, AI technologies will continue to evolve and be studied by various threat actors. Microsoft will continue to track threat actors and malicious activity misusing LLMs, and work with OpenAI and other partners to share intelligence, improve protections for customers and aid the broader security community.

Appendix: LLM-themed TTPs

Using insights from our analysis above, as well as other potential misuse of AI, we're sharing the below list of LLM-themed TTPs that we map and classify to the MITRE ATT&CK® framework or MITRE ATLAS™ knowledgebase to equip the community with a common taxonomy to collectively track malicious use of LLMs and create countermeasures against:

- **LLM-informed reconnaissance:** Employing LLMs to gather actionable intelligence on technologies and potential vulnerabilities.
- **LLM-enhanced scripting techniques:** Utilizing LLMs to generate or refine scripts that could be used in cyberattacks, or for basic scripting tasks such as programmatically identifying certain user events on a system and assistance with troubleshooting and understanding various web technologies.
- **LLM-aided development**: Utilizing LLMs in the development lifecycle of tools and programs, including those with malicious intent, such as malware.
- **LLM-supported social engineering**: Leveraging LLMs for assistance with translations and communication, likely to establish connections or manipulate targets.
- **LLM-assisted vulnerability research**: Using LLMs to understand and identify potential vulnerabilities in software and systems, which could be targeted for exploitation.
- **LLM-optimized payload crafting**: Using LLMs to assist in creating and refining payloads for deployment in cyberattacks.
- **LLM-enhanced anomaly detection evasion**: Leveraging LLMs to develop methods that help malicious activities blend in with normal behavior or traffic to evade detection systems.
- **LLM-directed security feature bypass**: Using LLMs to find ways to circumvent security features, such as two-factor authentication, CAPTCHA, or other access controls.

- **LLM-advised resource development**: Using LLMs in tool development, tool modifications, and strategic operational planning.