

MAI-Cyber-1.1-Flash

Model summary

Developer	Microsoft Ireland Operations Limited (MIOL), 70 Sir John Rogerson's Quay, Dublin 2, D02 R296, Ireland
Description	MAI-Cyber-1.1-Flash is an efficient 5B parameter cybersecurity-specialized in-house model built from Cyber-1-Flash, which was built from MAI-Code-1-Flash. MAI-Cyber-1.1-Flash is a cyber permissive model that has been designed to deepen cybersecurity capability to support vulnerability discovery, red teaming, and pen testing. Customer access remains restricted and is only available to customers through Codename MDASH, Microsoft's multi-model agentic scanning system.
Model architecture	The model uses a transformer architecture with self-attention and sparse Mixture-of-Experts (MoE) layers. MAI-Cyber-1.1-Flash is a cybersecurity specialized fine-tune of MAI-Cyber-1-Flash.
Parameters	137B total, 5B active
Inputs	Text.
Outputs	Text.
Context length	256k
Training dates	
Release date Release date in the EU (if different)	27 Oct 2026
License	Use of MAI-Cyber-1.1-Flash is subject to applicable Microsoft terms, policies, and preview terms in your agreement with Microsoft for MDASH, including the Microsoft Enterprise AI Services Code of Conduct, the terms applicable to Microsoft Foundry in the Microsoft Product Terms and the Azure Supplemental Preview Terms of Use.
Model dependencies	MAI-Cyber-1.1-Flash is an update to MAI-Cyber-1-Flash, which is a fine-tuned version of the MAI-Code-1-Flash model. You can read more about the MAI-Cyber-1-Flash model here and the MAI-Code-1-Flash model here .
List and link to any additional related assets	The provider has not supplied this information.
Acceptable use policy	Use of MAI-Cyber-1.1-Flash is subject to applicable Microsoft terms, policies, and preview terms in your agreement with Microsoft for MDASH, including the Microsoft Enterprise AI Services Code of Conduct, the terms applicable to Microsoft Foundry in the Microsoft Product Terms and the Azure Supplemental Preview Terms of Use.

1. Model overview

MAI-Cyber-1.1-Flash is an updated specialized cyber model with 5 billion active parameters, built for a wider range of agentic security workflows, operating with greater efficiency and precision. Developed using real-world executable security environments and trained to reduce cyber over-refusals, the model helps organizations identify, prioritize, and remediate vulnerabilities across modern codebases at enterprise scale within codename MDASH.

1.1 Key model capabilities

- **Cybersecurity specialization:** Purpose-built for cyber defense use cases including vulnerability discovery, prioritization, and remediation workflows.
- **Built for agentic workflows:** Designed to power agentic security workflows that help analyze, prioritize, and remediate vulnerabilities across large software environments.
- **Coverage without compromise:** Optimized for broad and frequent security scanning, helping organizations achieve more complete code coverage without sacrificing operational efficiency.
- **Real-world security environments:** Co-designed with codename MDASH and developed using real-world executable security environments rather than solely static datasets.
- **Microsoft security expertise:** Built on Microsoft's experience defending large-scale enterprise environments and critical infrastructure.

1.2 Alignment approach

MAI-Cyber-1.1-Flash is designed to support a broader range of cybersecurity scenarios to provide Microsoft Security customers with a helpful and cost-effective method to secure their systems. Due to the dual-use nature of these capabilities, ensuring the model is evaluated and deployed safely has been a primary consideration.

MAI-Cyber-1.1-Flash was trained on internal Microsoft safety training data to improve its overall alignment to Microsoft Responsible AI Principles and Microsoft policies. The team also implemented hardened security environments and internal monitoring mechanisms to ensure safety and alignment throughout all cyber evaluations. Scaling these mechanisms is a continued priority.

Access to MAI-Cyber-1.1-Flash is additionally restricted to select Microsoft codename MDASH customers and subject to customer review and approval processes. More

information about the limited access approach and system-level safeguards is provided in 2.8 below.

2. Usage

2.1 Primary use cases

- Vulnerability detection: Identifying potential vulnerabilities across large software codebases.
- Vulnerability prioritization: Assisting security teams in evaluating and prioritizing remediation efforts.
- Agentic remediation workflows: Supporting automated and semi-automated cyber defense workflows for modern software development environments.
- Enterprise-scale code security: Enabling broader coverage across millions of lines of code through cost-efficient scanning workflows.

2.2 Out-of-scope use cases

Any use outside of defensive security operations within Microsoft codename MDASH is out of scope for this model.

2.3 Distribution channels

Some of MAI-Cyber-1-Flash's distribution channels include:

Azure AI Foundry: MAI-Cyber-1.1-Flash is available through Azure AI Foundry Private Preview solely for use within Microsoft codename MDASH. Due to the dual-use nature of advanced cybersecurity capabilities, access is restricted and subject to additional approval requirements.

2.4 Input formats

Text.

2.5 Output formats

Text.

2.6 Technical requirements and integration guidance

MAI-Cyber-1.1-Flash is only designed for integration with codename MDASH.

2.7 Supported languages

The model primarily supports English, most official languages of the European Union, and additional non-EU languages inherited from MAI-Code-1-Flash.

2.8 Responsible AI considerations

MAI-Cyber-1.1-Flash is a notable update to MAI-Cyber-1-Flash because it was trained to reduce cyber refusals to support a broader set of Microsoft Security product needs. The model was carefully tested and evaluated across multiple MAI and Microsoft teams and by an external evaluator to understand the novel risk it poses and the impact of increased cyber permissiveness on cyber harm evaluations. Because of the increased model-level risk, the model is only being made available within codename MDASH to select verified customers.

The codename MDASH product includes multiple guardrails against abuse. There is limited ability to engage with the model within MDASH and users cannot edit prompts. Any attempt to use MAI-Cyber-1-Flash outside of MDASH triggers an alert and leads to revoked access if not remediated. Red teaming the model within codename MDASH confirmed that the harness materially constrains the risk of malicious use. Users who identify issues or concerns with the model can report vulnerabilities or abuse to the Microsoft Security Response Center (MSRC).

2.9 Known limitations

The model has been primarily trained and tested on English language data and benchmarks, and may initially perform less well for non-English speaking customers. As with any large language model, AI-generated text and code from MAI-Cyber-1.1-Flash may be inaccurate, incomplete, or otherwise incorrect. Developers should review, test, and validate outputs before relying on them in production or other consequential contexts.

3. Quality and performance evaluation

MAI-Cyber-1.1-Flash was evaluated across both public benchmarks and on extensive internal evaluations. Because the model is designed to operate as part of the MDASH multi-model system, evaluation focuses on vulnerability discovery, validation, and remediation performance across real-world codebases.

Model-Only External Cyber Benchmark Results

MAI-Cyber-1.1-Flash was evaluated at the model layer across a range of cybersecurity capability and safety benchmarks. These evaluations measure the model's ability to complete end-to-end vulnerability workflows, compromise vulnerable applications, generate exploits, reason about software vulnerabilities, analyze malware, interpret cyber

threat intelligence, and perform broader security-analysis tasks. The suite also includes safety evaluations that assess behaviors such as toxic content generation and payment-related manipulation.

CyberGym-E2E evaluates the complete find, prove, and fix vulnerability workflow, while WebExploitBench evaluates web-application penetration testing using verified-effect grading. CyberGym target-crash, ExploitGym, and CVEBench assess vulnerability discovery and exploitation capabilities. CyberSecEval4 evaluates threat-intelligence and malware-analysis capabilities; CRSBench evaluates full-pipeline cyber reasoning; CTI-Realm and ExCyTIn evaluate cyber threat-intelligence analysis; and SecBench Pro evaluates applied cybersecurity analysis. ToxiGen and MakeMePay provide additional measurements of model-safety behavior.

The headline CyberGym-E2E and WebExploitBench evaluations used a lightweight terminal-tool harness and a 256K-token context window. Results with multiple configurations are reported separately rather than averaged. These results represent model-layer performance and do not include additional orchestration or capabilities provided by the codename MDASH system.

External Cyber Benchmark Results		
Category	Benchmark	MAI-Cyber-1.1-Flash
Headline capability	CyberGym-E2E	S3 = 41.68%
Headline capability	WebExploitBench	23.0%
Exploitation	CyberGym target-crash	71.70%
Exploitation	ExploitGym	0/862
Vulnerability exploitation	CVEBench	Medium = 0.371 High = 0.400
Threat intelligence and malware analysis	CyberSecEval4	Threat Intelligence (Low) = 0.575 Malware Analysis (Low) = 0.314
Cyber reasoning	CRSBench	Minimum = 0.500 Low = 0.570 Medium = 0.610 High = 0.590
Threat intelligence	CTI-Realm	Minimum = 0.530 Low = 0.556 Medium = 0.580 High = 0.610
Cybersecurity analysis	SecBench Pro – IntuneExplorer	High = 0.854
Cyber threat intelligence	ExCyTIn	Minimum = 0.800 Low = 0.870 Medium = 0.880

		High = 0.890
--	--	--------------

3.1 Benchmarking methodology

MAI-Cyber-1.1-Flash was evaluated both as a single model using a lightweight terminal harness to understand its standalone capabilities, and within the Microsoft codename MDASH product to better approximate how customers will experience the model.

3.2 Safety evaluation and red-teaming

The MAI-Cyber-1.1-Flash model was evaluated across a large suite of internal safety evaluations to assess broad safety behavior, jailbreak robustness, malicious-cyber request compliance, policy adherence, indirect prompt injection, and over-refusal. The results identified areas requiring deployment controls, particularly malicious-content generation and interactions with untrusted code. These risks are addressed through restricted access, identity verification, policy enforcement, and runtime safeguards.

MAI-Cyber-1.1-Flash was also tested via multiple distinct red-teaming operations. Microsoft engaged Casaba Security, an independent third-party firm, to evaluate the cyber capability of MAI-Cyber-1.1-Flash across harmful request containment, adversarial jailbreak techniques, XPIA, secure code generation, and public cyber benchmarks. This testing identified that the model contained 90.2% of reviewed generally harmful requests; the model had the most failures with benign requests and in low resource languages.

The MAI-Internal Red Team ran a red-team assessment of MAI-Cyber-1.1-Flash across 38 adversarial scenarios, spanning secure development practices, authorization and scope boundaries, and offensive attack capabilities including ransomware, malware, lateral movement, privilege escalation, persistence mechanisms, botnet infrastructure, and APT orchestration. This assessment uncovered several critical risks of the model alone and helped inform the decision to not allow direct model access, but only within codename MDASH.

Microsoft's AI Red Team evaluated MAI-Cyber-1.1-Flash in codename MDASH, testing whether the system could be manipulated through prompt injection, whether attacks could enable vulnerability suppression, and whether an attacker could use the model through MDASH to generate malicious content. The red teaming operation identified no critical findings and previously identified prompt injection issues did not reproduce. MDASH's system-defined prompts, narrowly scoped stages, and structured report outputs materially constrain the risk of malicious use in the product.

4. Data overview

To read more about the data used to train MAI-Cyber-1.1-Flash please see the public [data](#) summary.

5. Contact

Requests for additional information can be directed to MSFTAIActRequest@microsoft.com.

Authorized representative: Microsoft Ireland Operations Limited 70 Sir John Rogerson's Quay,
Dublin 2, D02 R296, Ireland