MAI-Cyber-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-Flash is a small, efficient in-house model fine-tuned for cybersecurity vulnerability workflows in codename MDASH, Microsoft's multi-model agentic scanning system. MAI-Cyber-1-Flash is a 5B parameter model that can cover the majority of the MDASH workflow, including discovery, validation, triage, and patching, at a fraction of the cost of third-party models.
Model architecture	The model uses a transformer architecture with self-attention and sparse Mixture-of-Experts (MoE) layers. MAI-Cyber-1-Flash is a cybersecurity-specialized fine-tune of MAI-Code-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 July 2026
License	Use of MAI-Cyber-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-Flash is a fine-tuned version of the MAI-Code-1-Flash model. You can read more about 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-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-Flash is Microsoft's specialized cyber defense model with 5 billion active parameters, built for agentic security workflows. Developed using real-world executable security environments, 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-Flash was trained on internal Microsoft safety training data to improve its overall alignment to Microsoft Responsible AI Principles and Microsoft policies. Access to MAI-Cyber-1-Flash is additionally restricted to select Microsoft codename MDASH customers and subject to customer review and approval processes.

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-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-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

Microsoft's Responsible AI Principles and policies were incorporated into the decisions made throughout the full model training, testing, and deployment pipeline. MAI-Cyber-1-Flash has been trained and evaluated on responsible AI and safety data to ensure it can perform safely and responsibly.

Cybersecurity capabilities can be inherently dual-use. Identifying and verifying vulnerabilities is critical for defenders, but dangerous in the hands of cybercriminals. MAI-Cyber-1-Flash has been trained to carry out defensive cyber tasks such as patching bugs, and not to carry out offensive cyber tasks such as deploying malware. To further reduce

this risk, access to MAI-Cyber-1-Flash is intentionally restricted to select codename MDASH customers and subject to customer review and approval processes.

Some cyber specialized models are designed to be broadly permissive and have reduced guardrails. However, this can increase the risks of the models being used for malicious purposes or unintentionally causing harm. MAI-Cyber-1-Flash has been trained to forget content that could help it carry out harmful tasks unrelated to cyber defense such as supporting the development of weapons, other kinds of violence or self-harm, or taking actions that undermine the safety of children.

Additional guardrails have also been implemented to monitor and control uses of MAI-Cyber-1-Flash within codename MDASH. 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-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-Flash was evaluated across a broad set of 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.

Benchmarking Performance Within MDASH

CyberGym is a large-scale cybersecurity evaluation suite that measures AI agents on real-world vulnerability analysis tasks, spanning more than 1,500 benchmark instances drawn from 188 software projects. MAI-Cyber-1-Flash was evaluated against CyberGym inside codename MDASH, Microsoft's multi-model code vulnerability scanning system. **Replacing 80% of the existing models in codename MDASH with MAI-Cyber-1 improved codename MDASH performance on CyberGym from 88.4% to 95.95%.**

Model-Only External Cyber Benchmark Results

The model was also evaluated using a lightweight terminal harness on **CVEBench**, which evaluates the ability of AI agents to exploit real-world web application vulnerabilities;

CyberSecEval4, which assesses the cybersecurity vulnerabilities and defensive capabilities of models including threat intelligence and malware analysis capabilities; **CRSBench**, which evaluates full-pipeline cyber reasoning systems; and **ExploitGym**, which evaluates exploit generation. The documented MAI-Cyber-1-Flash results below are based upon the latest model version made available for testing. When used within the codename MDASH system alongside other models, the model may support higher scores on these benchmarks.

External Cyber Benchmark Results	
Benchmark	**MAI-Cyber-1-Flash**
CVEBench	0.314
CyberSecEval4	Threat Intel: 0.553 Malware Analysis: 0.33
CRSBench	0.651 (POV=1200)
ExploitGym	Kernel = 0, Userspace = 0, Browser = 0

3.1 Benchmarking methodology

MAI-Cyber-1-Flash was primarily evaluated as a single model using a lightweight terminal harness to understand its standalone capabilities. Benchmarks were run on the release checkpoint, which includes RAI safety training in the model weights. Additional benchmarks were run to see the model's performance in the Microsoft codename MDASH product.

3.2 Safety evaluation and red-teaming

MAI-Cyber-1-Flash was trained on internal Microsoft safety training data and performs well and comparably to other Microsoft models on internal safety evaluations while maintaining its ability to advance cybersecurity objectives.

To comprehensively assess MAI-Cyber-1-Flash, the Microsoft AI Red Team also thoroughly tested MAI-Cyber-1-Flash for potential capabilities or limitations of concern. The Microsoft AI Red Team worked closely with the model developers to establish **security-first calibration** for MAI-Cyber-1-Flash. Because MAI-Cyber-Flash-1 is Microsoft's first model built specifically for cyber use cases, we chose to begin with a cautious operating posture.

In practice, this means the model starts with a wider secure operating envelope and prioritizes risk-aware completions while we validate its behavior across legitimate cyber workflows. As a result, its safeguards may step in more often when a request is ambiguous, lacks context, or resembles potentially harmful activity, even when the user may have a legitimate defensive goal. During initial deployment, the model is intentionally calibrated to favor caution when requests are ambiguous or lack sufficient context. As we continue to evaluate the model through safety testing, red teaming, and operational performance reviews, we may refine this calibration to better support legitimate defensive security workflows while maintaining strong protections against misuse.

Once the model training was complete, the Microsoft AI Red Team began a four-part assessment of MAI-Cyber-1-Flash to understand both the model's underlying cyber capabilities and how its safeguards performed under realistic and adversarial conditions. We began by benchmarking the model across areas that can serve as leading indicators of cyber capability. These included malware generation, vulnerability discovery in critical infrastructure, end-to-end exploit development, and performance on long-horizon tasks that require multiple steps of planning and execution.

All benchmark testing took place in a tightly controlled, network-isolated environment. The evaluation infrastructure had no access to production systems, the public internet, or external services. Model outputs and generated artifacts remained contained within the environment, with strict controls over tool access, code execution, data movement, and human review.

We used Microsoft's AI red teaming framework, PyRIT, to conduct automated adversarial testing. The system drew on more than one million multi-turn jailbreak conversations to test whether attackers could bypass the model's safeguards or elicit restricted cyber capabilities.

This work was carried out by an autonomous red teaming system built specifically for this purpose. It generated, adapted, and evaluated attacks across MAI-Cyber-1's core capability areas, allowing us to test a much broader range of adversarial strategies than would be practical through manual testing alone.

We complemented the automated testing with hands-on assessments by Microsoft security experts. We focused on capability elicitation to examine capabilities such as autonomy, tool use, and long-horizon reasoning, and explored how those capabilities could apply to cyber operations. We also focused on realistic red team scenarios. We placed MAI-Cyber-1-Flash in workflows that more closely resemble real-world security work, where people remain involved throughout the process. This helped us understand how the model performed beyond isolated benchmark tasks and whether multiple outputs could be combined into a more consequential end-to-end capability.

Finally, we invited an independent third party to conduct its own assessment of the model. This added another layer of scrutiny and helped us determine whether external reviewers could identify risks or failure modes that had not emerged during Microsoft's internal testing. The independent assessment identified no critical-severity findings.

4. Data overview

To read more about the data used to train MAI-Cyber-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