MAI-Code-1.1-Flash model card

Model Summary

Developer	Microsoft Corporation Authorized Representative: Microsoft Ireland Operations Limited (MIOL) 70 Sir John Rogerson's Quay, Dublin 2, D02 R296, Ireland
Description	MAI-Code-1.1-Flash is a text-to-text and image-to-text coding model built for fast, efficient assistance in everyday developer workflows. The model is optimized to deliver high-quality coding help with low latency and low serving cost, making it well-suited for tasks such as agentic coding in real repositories, repository question answering, refactoring, and tool-using developer scenarios.
Model architecture	The model uses a transformer architecture with self-attention using sparse Mixture-of-Experts layers.
Parameters	138B total, 5B active
Inputs	Text, Image
Context length	256K tokens
Outputs	Text
Public data summary	https://microsoft.ai/pdf/MAI-Code-1.1-Flash-Data-Card.PDF
Training dates	March 2026 - August 2026
Release date	August 11, 2026
Release date in the EU	August 11, 2026
License	Various product and service terms where the model is deployed, including, for those who purchased through GitHub, the GitHub Customer Agreement or the GitHub Terms of Services and the GitHub

	Generative AI Services terms, and for those who purchased through Microsoft, the applicable Microsoft license agreement and the Microsoft Product Terms for GitHub Offerings..
Model dependencies	MAI-Thinking-1 MAI-Code-1-Flash
Additional related assets	N/A
Acceptable use policy	As described in the License section above and the GitHub Acceptable Use Policies and the Microsoft Enterprise AI Services Code of Conduct. Subsequent integrations of MAI-Code-1.1-Flash may be subject to different terms of service and policies.

Key capabilities

About this model

MAI-Code-1.1-Flash is a text-to-text and image-to-text coding model built by Microsoft for fast, efficient assistance in everyday developer workflows. The model is optimized to deliver high-quality coding help with low latency and low serving cost, making it well-suited for tasks such as agentic coding in real repositories, repository question answering, refactoring, and tool-using developer scenarios inside GitHub Copilot.

Key model capabilities

- Agentic coding in real developer environments, trained directly with the GitHub Copilot harness.
- Adaptive solution-length control, stays concise for simple requests and spends more reasoning budget on complex tasks.
- Strong instruction-following across single-turn and multi-turn scenarios.
- Competitive reasoning across math, science, and visual coding tasks.

Key use cases

MAI-Code-1.1-Flash is a coding-focused generative model intended for interactive developer assistance and agentic coding tasks inside GitHub Copilot in Visual Studio Code. Primary use cases include code generation and completion, repository question answering, refactoring, telemetry-grounded coding tasks, and agentic tool use. GitHub Copilot CLI support is planned for a later rollout. The relevant terms of service and

acceptable use policy for GitHub Copilot Business and Enterprise users can be found here , and the terms of service and acceptable use policy for Copilot Free, Pro, and Pro+ users can be found here.

Pricing

To be finalized

Technical specs

Training cut-off date

Pretraining cut-off date: December 2025

Supported languages

English

Optimizing model performance

MAI-Code-1.1-Flash was trained directly with the GitHub Copilot harness used in production, so offline improvements translate into real-world developer quality. It also uses adaptive solution length control, staying concise about simpler requests and spending more reasoning budget on harder problems.

Training disclosure

Training, testing and validation

MAI-Code-1.1-Flash was trained end to end by Microsoft on carefully curated data using Microsoft infrastructure and data pipelines. The development pipeline spans pre-training, mid-training, supervised fine-tuning, and reinforcement learning, starting from MAI-Thinking-1's compressed 5B-active-parameter mid-training checkpoint.

A lightweight supervised fine-tuning stage on curated instruction-following and agentic task data was applied on top of that mid-training checkpoint to establish reliable instruction- and format-following behavior. An additional "mid2" training phase used approximately 2 million diverse synthetic agentic tasks, organized into two progressive stages from simpler to more complex scenarios. A final large-scale reinforcement learning stage was conducted on diverse agentic tasks spanning more than 150,000 environments to strengthen per-token intelligence and end-to-end task quality. Synthetic data techniques — including prompt rewriting, rubrics synthesization, process supervision, and repo-level synthesization — were used throughout to maintain learnability as the model improved.

For testing and validation, MAI-Code-1.1-Flash was evaluated using both public benchmarks and internal benchmarks built from real developer scenarios. All evaluations were run with the GitHub Copilot production harness used to serve users, covering software engineering tasks, repository question answering, refactoring, and telemetry-grounded tasks adapted from real GitHub Copilot usage, so offline measurements reflect real-world model behavior.

Distribution

Distribution channels

Available in GitHub Copilot for all clients.

Responsible AI considerations

Safety techniques

Safety is addressed across the full training pipeline. During pre-training, harmful content is filtered out or demoted in the data mixture to reduce exposure during base model learning. In later training phases, including supervised fine-tuning and reinforcement learning, alignment techniques are applied to shape model behavior, reinforce safe and helpful responses, and discourage harmful, unsafe, or otherwise undesirable outputs.

Safety evaluations

This model was evaluated for safety using a combination of cybersecurity benchmarks and secure coding assessments, including CyberBench, CyberSecEval, and SecRepo, to assess its ability to withstand real-world security threats, avoid introducing vulnerabilities, and align with secure coding standards. Additional safety evaluations were conducted through a structured release process using production model APIs with safety classifiers and filters applied. These layered methodologies help ensure the model meets rigorous safety and security requirements before deployment.

Known limitations

As with any large language model, AI-generated text and code from MAI-Code-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.

Acceptable use

Acceptable use policy

MAI-Code-1.1-Flash is intended for commercial use within GitHub Copilot and its supported clients. Use of the model is governed by the applicable GitHub Copilot terms of service and Microsoft's Enterprise AI Code of Conduct, including restrictions on generating harmful, unlawful, or otherwise prohibited content.

Quality and performance evaluations

MAI-Code-1.1-Flash was evaluated across a broad set of benchmarks covering agentic coding, web app generation, and image-to-code scenarios. Key improvements on top of MAI-Code-1-Flash are:

	MAI-Code-1.1-Flash		MAI-Code-1-Flash		Haiku 4.5		GPT 5.4 mini	
	Pass rate	Tokens usage	Pass rate	Tokens usage	Pass rate	Tokens usage	Pass rate	Tokens usage
SWE-Bench Verified	72.6	8.6K	71.6	10.8K	69.8	20.9K	69.2	9.4K
Terminal Bench 2.1	62.9	17.0K	51.7	14.2K	49.4	25.5K	60.7	21.9K

Enabling vision capabilities for coding tasks.

		MAI-Code-1.1-Flash		Haiku 4.5		GPT 5.4 mini	
		Pass rate	Tokens usage	Pass rate	Tokens usage	Pass rate	Tokens usage
WebApp development (internal)	Text2WebApp	74.1	17.1K	11.5	60.0K	58.3	36.4K
	ScreenShot2WebApp	42.1	10.5K	10.0	33.9K	39.3	26.6K
Vision2Web	Vision2Web Level3	11.5	15.1K	13.7	36.5K	10.1	3.7K

For more general benchmarks, refer to MAI-Code-1-Flash model card here.

Benchmarking methodology

SWE-Bench Verified, Terminal Bench 2.1 were evaluated in the same VS Code-based harness used to evaluate production GitHub Copilot workflows. Pass rates are measured end to end, with repository context, tool calls, and verification included, rather than in a stripped-down benchmark environment. Solution length is reported as the average number of tokens used per completed task; shorter responses are better when pass rates are comparable. All compared models were evaluated in the same harness with the same settings to keep the comparison consistent.

Public data summary

Pre-training used a data mixture combining diverse, high-quality sources — including web text, code, books, and technical documents — carefully deduplicated and filtered to maximize knowledge coverage while reducing memorization of low-quality content. Post-

training used synthetic and open-source coding datasets spanning SWE task completion (SWE-Bench variants, repository sampling, and bug-bounty data), codebase question answering (repository- and file-level QA), and instruction following for developer workflows (debug, explain, build, and improve). To read more about the data used to train MAI-Code-1.1-Flash, please see the public data summary here.