Microsoft's New AI Models *Field Guide*

Seven in-house models in five families, shipped at Build 2026. What each does, where to find it, and the shift underneath.

V1.0 · 2026 · BUILD 2026 · MAI FAMILY

By Bas van Kaam

REASONING · FOUNDRY PREVIEW · 1

MAI-Thinking-1

WHAT IT DOES
Microsoft AI's first reasoning model, for multi-step logic, math, and software-engineering work.

HOW IT WORKS
35B active-parameter MoE, **256K** context. Trained from scratch on clean licensed data, no distillation.

ON THE BENCHMARKS
Preferred over Claude Sonnet 4.6 in Microsoft's blind human tests; matches Opus 4.6 on SWE-Bench Pro.

WHERE TO FIND IT
Microsoft Foundry (private preview), plus Baseten for developers.

CODING · ROLLING OUT · 2

MAI-Code-1-Flash

WHAT IT DOES
A lightweight agentic coding model for writing and editing code faster.

HOW IT WORKS
5B active parameters, inference-efficient. Comparable to Claude Haiku, at a lower cost.

WHAT IT SUPPORTS
Agentic, multi-file coding tasks deep inside the Microsoft developer stack.

WHERE TO FIND IT
GitHub Copilot and **VS Code**, plus distribution on Foundry.

IMAGE · PREVIEW · 3 & 4

MAI-Image-2.5 & Flash

WHAT IT DOES
Microsoft's first text-to-image and image-editing models, with control-and-preserve edits.

ON THE BENCHMARKS
No. 2 for image editing and **No. 3** for text-to-image on the LM Arena leaderboards.

VARIANTS
A Flash variant trades a little quality for more speed and a lower price.

WHERE TO FIND IT
PowerPoint and **OneDrive** (preview), Foundry, and OpenRouter.

SPEECH TO TEXT · IN FOUNDRY · 5

MAI-Transcribe-1.5

WHAT IT DOES
Production transcription with content biasing for domain-specific terms.

HOW IT WORKS
43 languages, **2.4%** word error rate, No. 1 on the FLEURS benchmark.

ON SPEED
One hour of audio in under **15 seconds**, up to five times faster than rival models.

WHERE TO FIND IT
Foundry, wired into Copilot, Teams, and Dynamics 365.

TEXT TO SPEECH · FOUNDRY · 6 & 7

MAI-Voice-2 & Flash

WHAT IT DOES
Expressive speech with voice cloning from a short sample, plus emotion and role styles.

HOW IT WORKS
15 languages, 18 locales. Clones a voice from a 5 to 60 second sample. About **$22** per 1M characters.

ON SAFETY
Consent guardrails gate voice cloning. A Flash variant for low-latency agents is coming.

WHERE TO FIND IT
Foundry (Azure Speech), into VS Code and Dynamics 365 Contact Center.

FRONTIER TUNING

the real story, not the model count

Tune a model on your own workflow data, inside your own environment, using reinforcement-learning environments, private training gyms for AI.

The tuned model stays yours, and your institutional knowledge becomes part of it.

Microsoft's Excel-tuned MAI matched **GPT-5.4** at up to **ten times** the efficiency.

One tuned model reached the highest win rate of any tested, at roughly ten times lower cost.

THE PLATFORM PLAY

why build all this, why now

Built in-house for long-term self-sufficiency and less reliance on a single outside supplier.

Trained from scratch on clean, licensed, traceable data, co-designed with Microsoft's own **Maia 200** silicon for about a **1.4x** efficiency gain.

They sit on Foundry next to OpenAI and Anthropic's Claude, including Claude Opus 4.8.

Own the model, and Microsoft pays itself across hundreds of millions of users.

WHAT IT MEANS FOR YOU

the part worth keeping

The model is not the moat.

Build model-agnostic, keep a thin layer between you and whichever model you call, so you can move between Claude, MAI, and OpenAI without rebuilding the house.

Resist standardizing too early. The 43-language transcription and the voice model are useful for course and video work today.

Your data, your workflows, and your review process are the durable advantage.

Built from Microsoft's Build 2026 announcements, June 2026. Verify at microsoft.ai before relying on specifics. www.eucnewsnuggets.com