Microsoft MAI Models Explained: Reasoning, Coding, Speech

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · microsoft mai models · mai-thinking-1 · mai-code-1-flash · microsoft ai foundry · copilot ai models · microsoft reasoning model · azure ai foundry · enterprise ai models · mai-voice-1

Executive Summary

Microsoft AI, the internal organization Mustafa Suleyman has led as Executive Vice President and CEO since March 2024 (blogs.microsoft.com), now ships a family of in-house **MAI** models that run alongside, rather than in place of, the OpenAI and Anthropic models already embedded in Microsoft's Copilot products. As of **September 3, 2026**, the family spans five categories: a general foundation model (**MAI-1-preview**, announced August 28, 2025 (microsoft.ai)), a dedicated reasoning model (**MAI-Thinking-1**, reaching public preview on Microsoft Foundry on August 12, 2026 (microsoft.ai)), a dedicated coding model (**MAI-Code-1-Flash**, updated to **MAI-Code-1.1-Flash** and live inside GitHub Copilot (github.blog) (github.blog)), text-to-speech and speech-to-text models (**MAI-Voice-1/2** and **MAI-Transcribe-1/1.5/2** (techcommunity.microsoft.com) (techcommunity.microsoft.com) (techcommunity.microsoft.com), and text-to-image models (**MAI-Image-1** through **MAI-Image-2.6**, the latter ranking No. 2 on the Arena.ai leaderboard (microsoft.ai)).

MAI-Thinking-1 is a sparse mixture-of-experts model with 35 billion active and roughly 1 trillion total parameters (microsoft.ai), trained on 30 trillion tokens with a July 2025 data cut-off (microsoft.ai) (microsoft.ai), and priced on Foundry starting at $2 per 1 million input tokens and $8 per 1 million output tokens (techcommunity.microsoft.com), which Microsoft positions as substantially cheaper than comparable frontier models on the SWE-Bench Pro coding benchmark (techcommunity.microsoft.com). On Microsoft's own benchmark disclosures, it scores 97.0% on AIME 2025 and 52.8% on SWE-Bench Pro, ahead of Claude Sonnet 4.6 on the former and behind Claude Opus 4.6 and GPT-5.4 on the latter (microsoft.ai) (microsoft.ai); these are vendor-reported figures compiled from each lab's own model card, not an independently audited comparison. MAI-Code-1-Flash, a separate model tuned specifically against the GitHub Copilot production harness (microsoft.ai), scored 51.2% against Claude Haiku 4.5's 35.2% on the same SWE-Bench Pro benchmark using up to 60% fewer tokens, by Microsoft's own account (microsoft.ai).

This report's central clarification is that a **model** is distinct from a **Copilot product**: Microsoft 365 Copilot "will continue to be powered by OpenAI's latest models" (microsoft.com), while Copilot Studio lets builders choose per-agent among Microsoft's, OpenAI's, and Anthropic's models from a dropdown (learn.microsoft.com), and specific named features—including Copilot Daily, Copilot Podcasts, and the GitHub Copilot model picker—are documented as running on named MAI models (microsoft.ai) (github.blog). Microsoft AI leadership frames the in-house build as a hedge rather than a replacement: Suleyman has said Microsoft was "set free" from prior contractual limits with OpenAI roughly in early 2026 to formally pursue frontier model-building (venturebeat.com), while noting Microsoft still has "OpenAI, Anthropic, and thousands of models inside Foundry" available (venturebeat.com). Independent commentary is more skeptical of near-term competitiveness, observing that OpenAI's and Anthropic's models, not MAI, still do "the heavy lifting inside Copilot" on independent coding benchmarks (theairankings.com), and that Microsoft reported on July 30, 2026, that Microsoft 365 Copilot had surpassed 30 million paid seats (microsoft.com).

Enterprises evaluating this landscape should track three distinctions this report keeps explicit throughout: announcement dates versus general-availability dates, vendor-reported benchmarks versus independently observable leaderboard ranks such as Arena.ai's (arena.ai) (arena.ai), and model-level access (Microsoft Foundry, billed through a standard Azure subscription (learn.microsoft.com)) versus product-level access (a specific Copilot license). IntuitionLabs, a life-sciences AI consultancy founded in 2023 (intuitionlabs.ai), approaches this landscape as an implementation and governance advisor rather than as a competing model vendor.

Introduction and Background

Microsoft's artificial intelligence stack now runs on two tracks at once. One track licenses frontier models from partners, principally **OpenAI**, and surfaces them inside Microsoft 365 Copilot, Bing, and GitHub Copilot. The other track, built by the internal **Microsoft AI** organization, trains and ships Microsoft's own models under the **MAI** brand. Microsoft AI was formed in March 2024 when **Mustafa Suleyman**, co-founder of DeepMind and Inflection, joined as Executive Vice President and CEO of Microsoft AI, reporting directly to CEO Satya Nadella (blogs.microsoft.com). The reorganization brought the Copilot, Bing, and Edge product teams, along with Microsoft's generative-AI research group, under Suleyman's leadership (blogs.microsoft.com), and added **Karén Simonyan**, formerly of DeepMind, as Chief Scientist (blogs.microsoft.com).

The MAI label now spans several distinct model families rather than a single product: a general-purpose foundation model (**MAI-1-preview**), a dedicated reasoning model (**MAI-Thinking-1**), a dedicated coding model (**MAI-Code-1-Flash**, later **MAI-Code-1.1-Flash**), text-to-speech and speech-to-text models (**MAI-Voice-1/2** and **MAI-Transcribe-**

1/1.5/2), and text-to-image models (**MAI-Image-1** through **MAI-Image-2.6**). This is a source of genuine confusion for buyers because "MAI-1" and "MAI-Thinking-1" are not the same model on two release schedules; they are two different lineages announced roughly a year apart, and this report untangles which is which.

This is also a story about a distinction that matters for procurement: a **model** (an artifact accessible by API on Microsoft Foundry, formerly Azure AI Foundry) is not the same thing as a **Copilot product** (a packaged assistant such as Microsoft 365 Copilot or GitHub Copilot that may route requests to a mix of models from Microsoft, OpenAI, and other vendors). Microsoft has stated plainly that Microsoft 365 Copilot "will continue to be powered by OpenAI's latest models" even as MAI models are added elsewhere (microsoft.com), while separately confirming that products such as Bing, PowerPoint, and Azure Speech already draw on Microsoft's own MAI speech and image models (techcommunity.microsoft.com). As of **September 2026**, this report maps each MAI model to its documented architecture, interfaces, supported tasks, access conditions, and the evaluation evidence Microsoft and independent leaderboards have published for it, distinguishing vendor claims from independently observable results throughout.

What Are Microsoft's MAI Models? Definition and Taxonomy

"MAI" stands for **Microsoft AI**, the organization Suleyman leads, and it is used both as the name of that organization and as the brand prefix for models it trains in-house from scratch, as opposed to models Microsoft licenses (OpenAI's GPT family) or redistributes through its marketplace (Anthropic's Claude, Meta's Llama, xAI's Grok, and thousands of others) (azure.microsoft.com). Microsoft's own MAI model catalog groups the family into five categories: reasoning, coding, voice (text-to-speech), transcription (speech-to-text), and image generation (microsoft.ai).

Table 1 below summarizes the MAI model family as documented by Microsoft as of September 2026, including each model's category, headline specification, and how Microsoft makes it available.

Model	Category	Key Specification	Status / Availability (as of Sep. 2026)
MAI-1-preview	General foundation model	Mixture-of-experts model refined with roughly 15,000 Nvidia H100 GPUs (cnbc.com)	Announced Aug. 28, 2025 as Microsoft's first end-to-end in-house model; tested publicly on the LMArena leaderboard (cnbc.com)
MAI-Thinking-1	Reasoning	Sparse mixture-of-experts, 35B active / ~1T total parameters	Introduced at Build 2026 (June 2, 2026); reached public preview on Microsoft Foundry Aug. 12, 2026
MAI-Code-1-Flash / 1.1-Flash	Coding	Lightweight, agentic model tuned against the GitHub Copilot production harness	Announced June 2, 2026 (microsoft.ai); updated to version 1.1 and live in GitHub Copilot by Aug. 11, 2026 (github.blog)
MAI-Voice-1 / MAI-Voice-2	Text-to-speech	Generates 60 seconds of expressive audio in under one second on a single GPU (techcommunity.microsoft.com)	MAI-Voice-1 announced Aug. 28, 2025 (microsoft.ai); powers Copilot Daily and Copilot Podcasts

Model	Category	Key Specification	Status / Availability (as of Sep. 2026)
MAI-Transcribe-1 / 1.5 / 2	Speech-to-text	MAI-Transcribe-2 expands coverage to 60 languages and adds speaker diarization and word-level timestamps (techcommunity.microsoft.com)	MAI-Transcribe-2 was announced in Microsoft Foundry public preview on September 3, 2026
MAI-Image-1 through MAI-Image-2.6	Text-to-image	MAI-Image-2.5 debuted at No. 2 on the Arena.ai text-to-image leaderboard (techcommunity.microsoft.com)	MAI-Image-1 launched Oct. 13, 2025, debuting in the top 10 on LMArena (microsoft.ai)

The table shows a pattern worth stating plainly: **coding is not a subset of reasoning**. Microsoft maintains MAI-Code-1-Flash as a separate model line from MAI-Thinking-1, purpose-built for developer workflows rather than adapted from the general reasoning model (microsoft.ai). A reader searching for "Microsoft's MAI coding model" should look for MAI-Code-1-Flash specifically, not assume MAI-Thinking-1 is the enterprise coding answer, even though MAI-Thinking-1 is also evaluated on coding benchmarks as part of its reasoning workload.

MAI-Thinking-1: Microsoft's Reasoning Model

MAI-Thinking-1 is Microsoft AI's first dedicated reasoning model: a sparse mixture-of-experts transformer with **35 billion active parameters** and **approximately 1 trillion total parameters**, developed entirely by Microsoft AI rather than licensed or fine-tuned from a third party (ai.azure.com). Like other reasoning models, it produces an internal chain of thought before returning a final answer, and Microsoft designed it to spend more or less of that reasoning effort depending on how complex the prompt appears (ai.azure.com). It was previewed at Microsoft's Build 2026 developer conference on June 2, 2026, and Microsoft Foundry documents MAI-Thinking-1 as a public preview, provided without an SLA and not recommended for production workloads.

Microsoft trained MAI-Thinking-1 on **30 trillion tokens** of pre-training data plus a further 3.55 trillion mid-training tokens, drawn from a mixture of public and licensed human-generated sources, with a training data cut-off of **July 2025**. The model supports a **256,000-token context window** and can generate up to 64,000 output tokens per request (learn.microsoft.com).

On benchmarks Microsoft itself ran and published, MAI-Thinking-1 scores **97.0%** on the AIME 2025 mathematics competition set, **94.5%** on AIME 2026, **87.7%** on LiveCodeBench v6 (a coding benchmark), and **52.8%** on SWE-Bench Pro, a harder software-engineering benchmark (microsoft.ai). These are vendor-reported figures, measured by Microsoft against other labs' own published model-card numbers rather than an independent third-party test. In blind human side-by-side evaluations covering 1,276 tasks, Microsoft reports MAI-Thinking-1 was preferred over Anthropic's Claude Sonnet 4.6 in 49% of comparisons versus 45% losses, and split roughly evenly with ties, and was preferred over Claude Opus 4.6 in 43% of comparisons versus 52% losses (microsoft.ai). Microsoft's own report is candid that the model "does not lead the field" outright but performs consistently across the categories it was tested on.

For enterprise deployment, MAI-Thinking-1 is accessible via a chat-completions API compatible with OpenAI SDK-style calling patterns, authenticated through Microsoft Entra ID or an API key (learn.microsoft.com). As of September 2026 it is offered only as a pay-as-you-go "Global Standard" deployment; reserved-capacity Provisioned Throughput Unit (PTU) deployment is **not** supported (learn.microsoft.com). Foundry pricing for MAI-Thinking-1

starts at **$2 per 1 million input tokens and $8 per 1 million output tokens** (techcommunity.microsoft.com), which Microsoft explicitly frames as a lower-cost alternative to comparable frontier models on the same benchmark. By comparison, OpenAI's GPT-5.4 is listed on Azure at roughly $2.50 per 1 million input tokens and $15.00 per 1 million output tokens for standard deployments (futureagi.com), nearly double MAI-Thinking-1's output price, though the two models are not benchmark-identical and the comparison should be read as a pricing data point rather than a capability equivalence claim.

Two governance details matter for enterprise buyers. First, the API can return an encrypted, opaque representation of the model's internal reasoning trace that developers are not meant to inspect, which preserves Microsoft's proprietary reasoning content but limits debuggability of the chain of thought (learn.microsoft.com). Second, Microsoft's own documentation states MAI-Thinking-1 "is not designed or evaluated for use as an autonomous decision-maker" in consequential domains such as law, finance, or safety-critical operations (ai.azure.com), a limitation that should inform any deployment involving regulated decisions.

MAI-Code-1-Flash: Microsoft's Coding Model

Microsoft's dedicated coding model, **MAI-Code-1-Flash**, was announced alongside MAI-Thinking-1 at Build 2026 on June 2, 2026, which Microsoft describes as built for fast, efficient assistance in everyday developer workflows. Unlike MAI-Thinking-1, which is a general reasoning model that happens to be evaluated on coding benchmarks, MAI-Code-1-Flash was trained and tuned directly against the production **GitHub Copilot** harness, and Microsoft describes it as agentic, with reasoning effort that adapts to task difficulty.

On Microsoft's own benchmark disclosures, MAI-Code-1-Flash scored **51.2%** on SWE-Bench Pro against **35.2%** for Anthropic's Claude Haiku 4.5, while using up to 60% fewer tokens to solve harder problems in the same benchmark suite. As with MAI-Thinking-1's benchmark claims, these are Microsoft-reported comparisons rather than independently reproduced scores, and should be read alongside the caveat that different labs' self-reported model-card numbers are not always generated under identical evaluation conditions.

GitHub's own product changelog, a Microsoft subsidiary's first-party release notes rather than a marketing page, confirms MAI-Code-1-Flash became selectable inside the GitHub Copilot model picker in Visual Studio Code starting June 2, 2026, calling it "the first in a new wave of purpose-built coding models from Microsoft" (github.blog). By **August 11, 2026**, GitHub's changelog shows the model updated to **MAI-Code-1.1-Flash**, still available inside GitHub Copilot (github.blog). This progression, from a Build-conference announcement to a shipped, versioned update inside a widely used developer tool within ten weeks, is one of the more concrete "announced-to-available" timelines in the MAI family, and it is worth enterprise engineering leads noting that the model's availability is currently scoped to GitHub Copilot rather than being a general-purpose coding assistant sold standalone through Foundry in the way MAI-Thinking-1 is.

Microsoft's broader roadmap post describing its 2026 MAI lineup places MAI-Code-1-Flash at roughly 5 billion active parameters, a size Microsoft positions as comparable to Anthropic's Claude Haiku but cheaper, and confirms the model is built specifically for and deeply integrated into GitHub Copilot and Visual Studio Code rather than positioned as a general chat assistant. Enterprise teams evaluating a coding-specific model should therefore treat MAI-Code-1-Flash and MAI-Thinking-1 as answers to different questions: MAI-Code-1-Flash for fast, in-editor completion and agentic coding tasks bound to the GitHub Copilot experience, and MAI-Thinking-1 for open-ended reasoning tasks (including but not limited to code) accessed directly through the Foundry API.

MAI-Voice, MAI-Transcribe, and MAI-Image: Speech and Multimodal Models

Microsoft AI's speech and image models are the most mature part of the MAI family in terms of production deployment. **MAI-Voice-1**, announced August 28, 2025 as Microsoft AI's "first highly expressive and natural speech generation model", can generate a full minute of expressive audio in under one second on a single GPU (microsoft.ai), a claim Microsoft repeated in its April 2026 Foundry public-preview announcement, describing the model as "capable of producing 60 seconds of expressive audio in under one second on a single GPU". On Microsoft Foundry's model catalog, MAI-Voice-1 supports cloning a voice from an audio clip of up to 120 seconds without any fine-tuning step, plus per-turn control over emotion and tone (ai.azure.com).

Its companion speech-to-text line progressed from **MAI-Transcribe-1** to MAI-Transcribe-1.5 and then **MAI-Transcribe-2**. Announced on September 3, 2026 in Microsoft Foundry public preview, MAI-Transcribe-2 expands coverage to 60 languages and adds speaker diarization and word-level timestamps (techcommunity.microsoft.com). Microsoft's April 2026 announcement states these speech models were already powering production features inside Copilot, Bing, PowerPoint, and Azure Speech before their public Foundry preview, and specifically that MAI-Voice-1 powers the **Copilot Daily** and **Copilot Podcasts** features, an example of a Copilot product feature built on a named, in-house MAI model rather than a licensed one.

On the image side, **MAI-Image-1** launched October 13, 2025, which Microsoft says debuted in the top 10 text-to-image models on LMArena, its community-run public benchmarking site. Microsoft engineered the model to more accurately simulate physical lighting behavior, such as bounce light and reflections, and to improve text-rendering accuracy inside generated images, aiming it at branding, poster, and packaging use cases (microsoft.ai). By August 2026, Microsoft's next-generation **MAI-Image-2.5** model had improved on that standing, debuting at **No. 2** on the Arena.ai (formerly LMArena) text-to-image leaderboard alongside a companion Flash variant.

Two distinctions are worth holding onto here. First, "LMArena" and "Arena.ai" refer to the same community leaderboard organization under an evolved name; Microsoft's own posts use both names depending on publication date, and this report treats them as the same ranking source. Second, a leaderboard rank is a community-voted, independently observable signal, distinct in kind from Microsoft's own internally run benchmark comparisons cited elsewhere in this report; where this report cites an Arena/LMArena rank, that number was not produced or curated by Microsoft.

From Model to Product: MAI Inside Copilot and Microsoft Foundry

The single most common source of confusion in coverage of Microsoft's AI strategy is conflating a **model** with a **Copilot product**. Microsoft has been explicit that this is not a 1:1 relationship. Microsoft 365 Copilot, the company's flagship productivity assistant, "will continue to be powered by OpenAI's latest models" even as Microsoft expands model choice elsewhere in its stack (microsoft.com), while **Microsoft Copilot Studio**, the low-code platform for building custom agents, lets a maker pick a "primary AI model" per agent from a catalog that includes Microsoft's own models alongside OpenAI's GPT family, Anthropic's Claude, and xAI's Grok, selected "using a simple dropdown menu" (learn.microsoft.com). Copilot Studio administrators can also independently allow or block

external (non-Microsoft) models separately from preview or experimental models, which confirms Microsoft treats model selection and product governance as two separate control planes rather than a single bundled choice (learn.microsoft.com).

Where a specific MAI model IS confirmed to power a specific Copilot feature, Microsoft has said so directly: MAI-Voice-1 for Copilot Daily and Copilot Podcasts, and MAI-Code-1-Flash for the GitHub Copilot model picker in Visual Studio Code (github.blog). Outside of these named cases, Microsoft has not published a complete, feature-by-feature map of which Copilot surface uses which underlying model, and a reader should not assume a given Copilot feature runs on a MAI model unless Microsoft has said so for that specific feature.

Underneath both the "model" and "product" layers sits **Microsoft Foundry**, the unified platform Microsoft describes as an interoperable AI platform, formerly branded Azure AI Studio (azure.microsoft.com). Foundry's model catalog spans more than 11,000 models, according to Microsoft, including foundation, open-source, reasoning, and multimodal models "spanning OpenAI, Anthropic, Meta, Google, xAI, Hugging Face, and frontier models including the new MAI multimodal family" (azure.microsoft.com) (azure.microsoft.com). Models Microsoft sells directly through Foundry, MAI models among them, are billed through the customer's existing Azure subscription, covered by standard Azure service-level agreements, and supported directly by Microsoft rather than by a third-party model vendor (learn.microsoft.com). Exploring the Foundry catalog does not itself require an Azure account, though an Azure subscription is required once a developer moves from browsing to actually building an agent (azure.microsoft.com).

Accessing and Deploying MAI Models: Enterprise Interfaces

The preceding sections cover Foundry deployment, governance, and model-specific access conditions. For coding, MAI-Code-1-Flash is available through GitHub Copilot and can be selected in the Visual Studio Code model picker (github.blog).

Data Analysis and Evidence

This section isolates the quantitative claims in this report, states their originators, and flags which numbers are vendor-reported versus independently observable.

Table 2 compares MAI-Thinking-1's Microsoft-published benchmark scores against other labs' self-reported model-card numbers, as compiled by Microsoft in its own technical report; none of these figures were independently re-run by a third party for this comparison, and Microsoft's report states plainly that competitor figures come from each lab's own official model cards rather than being reproduced in-house.

Benchmark	MAI-Thinking-1	Claude Sonnet 4.6	Claude Opus 4.6	GPT-5.4
AIME 2025 (math)	97.0%	95.6%	99.8%	not disclosed in source table
SWE-Bench Pro (coding)	52.8%	not disclosed in source table	53.4%	57.7%

Benchmark	MAI-Thinking-1	Claude Sonnet 4.6	Claude Opus 4.6	GPT-5.4
SWE-bench Verified (coding)	73.5%	79.6%	80.8%	not disclosed in source table
LiveCodeBench v6 (coding)	87.7%	not disclosed in source table	not disclosed in source table	not disclosed in source table

Table 2 shows MAI-Thinking-1 leading Claude Sonnet 4.6 on the AIME 2025 math benchmark while trailing both Claude Opus 4.6 and GPT-5.4 on SWE-Bench Pro, a pattern consistent with Microsoft's own characterization of the model as consistently strong rather than field-leading. All four models' numbers come from vendor self-reporting; readers should treat cross-lab comparisons in this table as directional rather than as results of a controlled, independently audited bake-off.

Independent, community-run evidence is thinner but does exist. Arena.ai's (formerly LMArena's) own leaderboard changelog confirms MAI-1-preview was added to its public Text Leaderboard on August 28, 2025 (arena.ai), and at launch the model ranked 13th for text workloads, behind entries from Anthropic, DeepSeek, Google, Mistral, OpenAI, and xAI, according to contemporaneous CNBC reporting (cnbc.com). As of this report's research date in September 2026, MAI-1-preview no longer appears among Arena.ai's visible top-ranked text models (arena.ai), while Microsoft's newer image models occupy stronger positions on Arena.ai's separate text-to-image leaderboard, with MAI-Image-2.6 and MAI-Image-2.5 both ranked inside the top ten as of the same date (arena.ai).

On cost, Microsoft's own capital allocation gives a sense of scale behind the MAI effort: after the shift from finance to operating leases, Microsoft adjusted its calendar-year 2026 capital-expenditure expectation to approximately **$175 billion**. This figure describes Microsoft's overall infrastructure investment, not MAI-specific spending, since Microsoft does not break out MAI as a separate reporting line.

Case Studies and Real-World Examples

GitHub Copilot's coding-model rollout. The clearest documented example of an MAI model moving from announcement to daily developer use is MAI-Code-1-Flash. GitHub's own changelog shows the model entering the Copilot model picker in Visual Studio Code on June 2, 2026, the same day Microsoft announced it at Build, and being updated to MAI-Code-1.1-Flash within GitHub Copilot by August 11, 2026 (github.blog), a same-quarter cadence of iteration that is unusual for a frontier-adjacent model and suggests Microsoft is treating the coding line as a fast-moving product surface rather than an annual release.

Copilot Daily and Copilot Podcasts. Microsoft's April 2026 Foundry announcement documents MAI-Voice-1 and its companion transcription model as already running inside shipped Microsoft products, specifically naming Copilot, Bing, PowerPoint, and Azure Speech, with Copilot Daily and Copilot Podcasts named as the two Copilot features MAI-Voice-1 specifically powers. This is a useful real-world marker because Microsoft used its own products as an early production environment for MAI voice technology.

Copilot Studio's model-agnostic agent builder (Hypothetical Example). Consider a hypothetical enterprise IT team building an internal HR-policy assistant in Microsoft Copilot Studio. Because Copilot Studio exposes a per-agent model selector spanning Microsoft's own models and third-party options such as OpenAI's and Anthropic's

(learn.microsoft.com), the team could start the agent on a general-purpose model, then later swap in MAI-Thinking-1 for a reasoning-heavy escalation path without re-platforming the agent, since model choice and product configuration are governed separately in Copilot Studio (learn.microsoft.com). This scenario illustrates the architectural flexibility Microsoft has documented rather than describing any specific customer's deployment.

Implications and Future Directions

Microsoft's own AI leadership has been unusually direct about why the company is building MAI models at all, given its multi-billion-dollar commercial relationship with OpenAI. Suleyman has said Microsoft's contractual position only recently allowed it to pursue frontier model-building formally: "we were only sort of set free from our contract with OpenAI about six months ago to formally pursue superintelligence," framing MAI as an exercise of newly available strategic freedom rather than dissatisfaction with existing partners. He has also framed the goal explicitly as capability diversification: "we have the capacity not just to buy models from third parties, but to build the absolute frontier, the best models in the world" (venturebeat.com), while stressing that Microsoft is not under near-term pressure to do so because it already has broad model access: "we have OpenAI, we have Anthropic, we have thousands of models inside Foundry" (venturebeat.com). On data provenance, Suleyman has stated MAI's reasoning models are trained "from scratch," without distillation from other labs' models and without unlicensed or opaque data (venturebeat.com), a claim that lines up with Microsoft's technical report language about MAI-Thinking-1's training corpus.

Independent commentary reads this multi-model posture as a hedge rather than a replacement strategy. One trade analysis observed that "no other company sells frontier AI from three rival sources at once," referring to Microsoft's simultaneous commercial relationships with OpenAI and Anthropic alongside its own MAI models (theairankings.com), while a separate industry commentary compared the logic to a different Big Tech vertical-integration move: "just as Google built Tensor chips to reduce NVIDIA dependency, Microsoft is building MAI models to reduce OpenAI dependency" (institutepm.com). The same trade analysis cautioned that, as of its assessment, "on independent coding benchmarks the models doing the heavy lifting inside Copilot are OpenAI's GPT-5.x and Anthropic's Claude Opus 4.8, not MAI" (theairankings.com), Microsoft reported on July 30, 2026, that Microsoft 365 Copilot had surpassed 30 million paid seats (microsoft.com).

For life-sciences and other regulated enterprises evaluating this landscape, the practical decision is rarely "pick one vendor's model family" but rather how to govern model choice, data handling, and validation evidence across a stack that increasingly mixes OpenAI, Anthropic, and Microsoft's own models inside the same Copilot Studio or Foundry environment. **IntuitionLabs**, a life-sciences AI consultancy founded in 2023 (intuitionlabs.ai) and based in San Jose, California (intuitionlabs.ai), works with pharmaceutical and life-sciences organizations on exactly this kind of multi-vendor AI governance question as an implementation and advisory partner rather than as a model provider itself; where a life-sciences organization is weighing whether a reasoning-heavy validated workflow belongs on MAI-Thinking-1, a licensed frontier model, or a mix of both routed through Copilot Studio, that governance and validation work sits outside any single model vendor's own documentation. Readers comparing Copilot against other enterprise AI assistants more broadly, including Claude, ChatGPT Enterprise, and Gemini, may also find IntuitionLabs' enterprise AI assistant comparison a useful companion to the model-level detail in this report.

Looking ahead, the clearest forward signal in the record is cadence: Microsoft announced MAI-Transcribe-2 in Microsoft Foundry public preview on September 3, 2026, after the point releases MAI-Code-1.1-Flash, MAI-Image-2.5 and its Flash variant, and MAI-Transcribe-1.5, suggesting the MAI line will keep iterating on a product-style release cadence rather than an annual frontier-model cycle, and that any snapshot of "the current MAI models," including this one, should be read with its stated observation date rather than assumed to be permanent.

Frequently Asked Questions (FAQs)

What is Microsoft MAI? MAI stands for Microsoft AI, both the internal organization led by Mustafa Suleyman (blogs.microsoft.com) and the brand under which that organization ships in-house models, as distinct from OpenAI or Anthropic models Microsoft licenses or redistributes (azure.microsoft.com).

Is MAI-1 the same as MAI-Thinking-1? No. **MAI-1-preview** was Microsoft's first end-to-end in-house foundation model, announced August 28, 2025. **MAI-Thinking-1** is a separate, later reasoning-specific model that reached public preview on Microsoft Foundry on August 12, 2026. They are different models from different points in the MAI roadmap, not two versions of the same release.

Does Microsoft have a dedicated coding model? Yes. **MAI-Code-1-Flash**, updated to **MAI-Code-1.1-Flash**, is Microsoft's dedicated coding model, tuned against the production GitHub Copilot harness and available through the GitHub Copilot model picker (github.blog), distinct from the general-purpose MAI-Thinking-1 reasoning model.

Can I use MAI models outside of Copilot? Yes, for most of the family. MAI-Thinking-1, MAI-Voice models, MAI-Transcribe-2, and MAI-Image models are available as direct API products on Microsoft Foundry, billed through a standard Azure subscription (learn.microsoft.com), independent of any Copilot license. MAI-Code-1-Flash is currently the exception, reachable through the GitHub Copilot model picker rather than as a standalone Foundry API product.

Are MAI models replacing OpenAI models in Copilot? Not based on Microsoft's own statements. Microsoft 365 Copilot "will continue to be powered by OpenAI's latest models" (microsoft.com), and Copilot Studio lets builders choose among Microsoft, OpenAI, Anthropic, and other models per agent rather than standardizing on one vendor.

How much does MAI-Thinking-1 cost to use? Microsoft lists Foundry pricing for MAI-Thinking-1 starting at $2 per 1 million input tokens and $8 per 1 million output tokens (techcommunity.microsoft.com).

Conclusion

Microsoft's MAI models are not a single product but a growing family of in-house alternatives running alongside, not in place of, the OpenAI and Anthropic models already embedded in Microsoft's Copilot lineup. MAI-Thinking-1 handles general reasoning, MAI-Code-1-Flash handles coding inside GitHub Copilot, and the MAI-Voice, MAI-Transcribe, and MAI-Image lines handle speech and image generation. As of this report's September 3, 2026 observation date, MAI-Transcribe-2 is the newest announced speech-to-text model, available in Microsoft Foundry public preview. Buyers evaluating this landscape should keep three distinctions straight: a model is not the Copilot product built on top of it, an announcement date is not a general-availability date, and a vendor-reported benchmark

score is not an independently reproduced one. Microsoft's own disclosures, read carefully and dated consistently, are detailed enough to make each of those distinctions without guesswork, which is precisely what this report has tried to document rather than merely assert.
