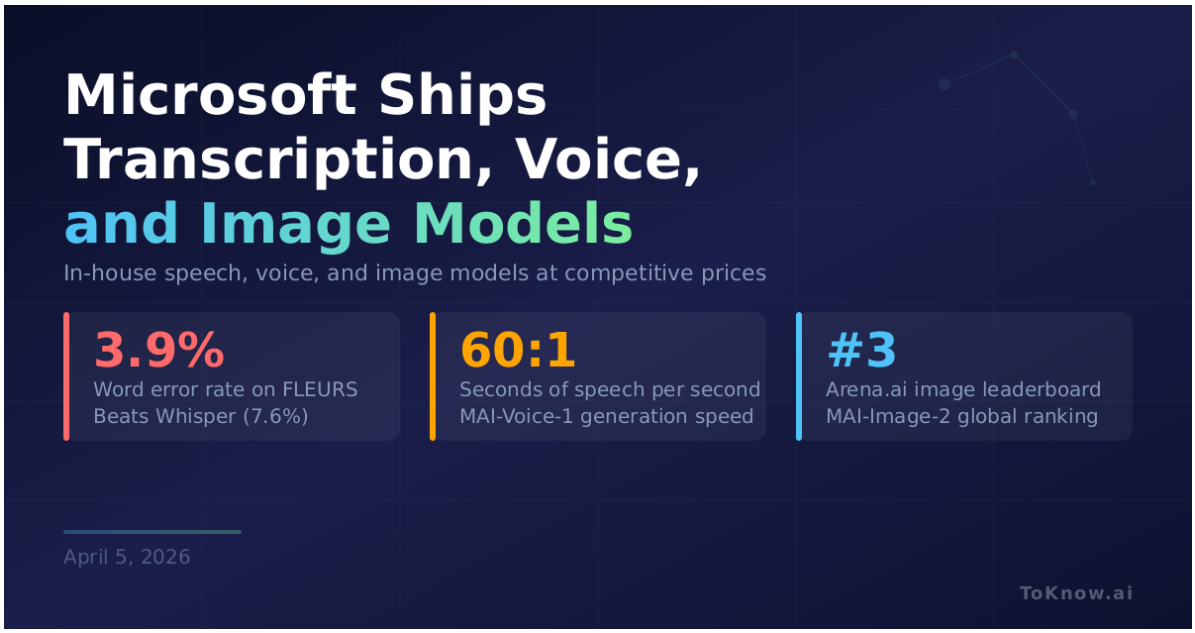
Microsoft Ships In-House Transcription, Voice Cloning, and Image Generation Models

Kabui, Charles

2026-04-05



Microsoft released three in-house AI models through its Foundry platform: MAI-Transcribe-1 for speech-to-text, MAI-Voice-1 for voice synthesis, and MAI-Image-2 for image generation. MAI-Transcribe-1 averages 3.9% word error rate on the FLEURS benchmark across 25 languages, beating Whisper-large-v3 (7.6%) and Gemini 3.1 Flash (4.9%), while running batch transcription 2.5x faster than Microsoft's previous Azure Fast offering. MAI-Voice-1 generates

60 seconds of speech in one second and can clone a voice from just a few seconds of audio. MAI-Image-2 ranks #5 on the Arena.ai text-to-image leaderboard and delivers 2x faster generation, with phased rollouts underway in Bing and PowerPoint.

The pricing is what developers will notice most. Transcription starts at $0.36 per hour, voice synthesis at $22 per million characters, and image generation at $5 per million input tokens. For anyone building voice agents, podcast tools, or creative apps, the combination of speed and cost matters more than marginal leaderboard differences. MAI-Voice-1's custom voice cloning from a few seconds of sample audio makes previously expensive voice production accessible to solo developers and small teams.

This is Microsoft building a full first-party model stack across speech, voice, and vision, separate from its OpenAI partnership. With these models already powering Copilot and rolling into Bing and PowerPoint, the strategy is clear: own the entire pipeline from model training to product deployment rather than reselling third-party models. For how Google is approaching the same image generation space, see Nano Banana 2: Google's 4K Image Generation in Gemini.

Sources:

- Microsoft AI Official Announcement
- MAI-Image-2 Introduction
- Arena.ai Text-to-Image Leaderboard
- MAI-Transcribe-1 Model Card (PDF)
