Copilot Goes Multi-Model as Open Voice and Local AI Accelerate

AI News Digest

2026-03-31

Copilot Goes Multi-Model as Open Voice and Local AI Accelerate

By AI News Digest • March 31, 2026

Microsoft rolled out multi-model research features in M365 Copilot, while Mistral and other open-model builders pushed audio, speech, and multilingual releases forward. Local AI also crossed a symbolic milestone with llama.cpp at 100k stars, as enterprise competition around OpenAI and Anthropic sharpened.

A few shifts stood out today

Microsoft is turning model plurality into a product, open releases are getting stronger across audio and speech, and local AI keeps looking more deployable. The market feels a bit less centered on one flagship model and more on orchestration, efficiency, and where systems actually run.

Microsoft brings multi-model workflows into Copilot

Microsoft introduced **Critique** in M365 Copilot, a multi-model deep research system that uses multiple models together to generate responses and reports; Satya Nadella said Microsoft's benchmarks show "best-in-class deep research." It also launched **Council**, which lets users run multiple models on the same prompt at once to compare alignment, divergence, and unique contributions. Both are available now in Frontier [1, 2, 3, 4, 5].

Why it matters: This is a notable product signal from a major platform vendor: instead of hiding model plurality behind one answer, Microsoft is exposing model collaboration and disagreement as a feature.

Open models broaden beyond text

Mistral's Voxtral TTS is a notable open-audio release

Mistral launched **Voxtral TTS**, an open-weight multilingual text-to-speech model that supports **nine languages** and targets **real-time streaming** for voice agents [6, 7]. Latent Space said the model posted a **68.4% win rate** against ElevenLabs Flash v2.5, while Mistral speakers described it as state-of-the-art quality at a fraction of proprietary costs [6, 7].

Its architecture mixes autoregressive semantic speech tokens with **flow matching** for acoustic tokens, backed by an in-house neural audio codec at **12.5 Hz**; the team also said the setup can extend to long generations via larger context windows [6, 7].

Mistral: Voxtral TTS, Forge, Leanstral, & Mistral 4 — w/ Pavan Kumar Reddy & Guillaume Lample (0:58)

Why it matters: Open voice models are getting closer to the quality, latency, and cost targets that matter for real-time products.

The broader open-model pipeline was unusually diverse

Interconnects highlighted an unusually broad set of open releases: NVIDIA's **Nemotron-3-Super-120B-A12B-NVFP4** with a **1M context window**, multilingual support, NVFP4 pre-training, and open pre-/post-training datasets; Cohere's **cohere-transcribe-03-2026** speech model with **14 languages** under **Apache 2.0**; Sarvam's **105B** and **30B** models with strong Indic-language positioning; and **Mistral-Small-4-119B-2603** as a hybrid reasoning model with coding abilities [8]. Interconnects argued this kind of domain-specific, cheaper model development is becoming an important complement to the strongest closed agents [8].

Why it matters: The open ecosystem is spreading across speech, multilingual, regional, and reasoning workloads instead of clustering around one general chatbot race.

Local AI looks more like infrastructure

llama.cpp reached 100k stars, and the stack around it keeps firming up

llama.cpp crossed **100k GitHub stars**, has **1,500+ contributors**, and Hugging Face said it is bringing Georgi Gerganov and ggml into the team behind what it called the most widely used open-source runtime for local AI [9, 10]. Gerganov said useful local agentic workflows became feasible as models improved tool calling on everyday devices, and Clement Delangue argued that many disappointments with smaller local models are really failures of scaffolding, chat templates, prompt construction, or fine-tuning rather than raw model capability [9, 11, 12, 13].

Gerganov also described **Qwen3.5** as a "step change" across device sizes, while Delangue urged open-source agent tools to rely primarily on open models rather than closed APIs that send data to the cloud [11, 14].

> "The technology is too important to be vendor-locked. It has to be developed in the open, by the community, together with the independent hardware vendors." [9]

Why it matters: This is starting to look less like enthusiast momentum and more like a real deployment path for private, on-device, and cross-platform AI.

Strategy watch

Ben Thompson sees OpenAI's enterprise focus as a competitive necessity

Ben Thompson argued that reports of OpenAI cutting side projects should not be overread as an exit from consumer; instead, he sees a rational shift of resources toward enterprise, where customers pay for productivity gains and **Codex** has been especially strong [15]. He framed the urgency around Anthropic's enterprise growth—described as moving from a **$14B** to **$19B** run rate—and the risk that OpenAI gets shut out if large customers standardize elsewhere [15].

He also noted OpenAI has pushed back on startup-skewed **Ramp** chart interpretations and may be stronger in the Fortune 500 than those charts suggest, while arguing that ChatGPT's massive consumer scale creates a harder monetization path because ads are difficult and compute is already heavily committed [15].

Why it matters: The center of gravity in AI competition may be shifting from consumer reach to enterprise distribution, pricing, and workflow lock-in.

Research watch

Self-improving agent scaffolds advanced, but frontier math remained hard

Import AI highlighted **Hyperagents**, a self-referential scaffold that lets LLM systems iteratively modify their own prompts and tools. In reported results, the setup improved **Polyglot** coding performance from **14% to 34%**, **paper review** from **0% to 71%**, and **robotics reward design** from **6% to 37%** [16].

The same roundup pointed to **HorizonMath**, a benchmark of **100 predominantly unsolved math problems** with automated verification, where the top model scored only **7% overall** and **50%** on the easiest subset [16].

Why it matters: The capability story remains mixed: better scaffolds are producing real gains on structured tasks, while benchmarks aimed at genuine mathematical discovery are still extremely hard.

Sources

1. X post by @satyanadella
2. X post by @satyanadella
3. X post by @satyanadella
4. X post by @satyanadella
5. X post by @satyanadella
6. Mistral: Voxtral TTS, Forge, Leanstral, & what's next for Mistral 4 — w/ Pavan Kumar Reddy & Guillaume Lample
7. Mistral: Voxtral TTS, Forge, Leanstral, & Mistral 4 — w/ Pavan Kumar Reddy & Guillaume Lample
8. Latest open artifacts (#20): New orgs! New types of models! With Nemotron Super, Sarvam, Cohere Transcribe, & others
9. X post by @ggerganov
10. X post by @ClementDelangue
11. X post by @ggerganov
12. X post by @ClementDelangue
13. X post by @ClementDelangue

14. X post by @ClementDelangue
15. Why OpenAI's Enterprise Pivot Makes Sense | Sharp Tech with Ben Thompson
16. Import AI 451: Political superintelligence; Google's society of minds, and a robot drummer