

Mistral's new voice model stays on Mistral's servers

October 26, 2026 · by Mark Bartlett

[Download this post as PDF](#)

InsiderLLM Weekly issue 27 – October 26, 2026

Mistral shipped Voxtral 3 last week, and unlike every Voxtral before it, there's nothing on Hugging Face to download. So the main story is the voice models you can download, and which of them fit a 24 GB card. First, check your LiteLLM version.

Quick Hits

- **LiteLLM 1.107.0 shipped a credential stealer.** The poisoned release was live on PyPI and Docker for about two hours on October 21. If LiteLLM fronts your local models, check your version, pin 1.106.2 as its maintainers advise, and rotate every key that machine could reach. ([how it happened](#))
 - **Microsoft's decision model is cloud-only. H2O's isn't.** Microsoft built Decision-1 on open Qwen3.5-9B weights but serves it only through Foundry and OpenRouter. H2O Lightning 4B answers the same kinds of questions in llama.cpp, from a 4.48 GB file under Apache 2.0. ([Microsoft](#))
 - **DeepSeek's open engine needs a data center.** DeepInfer, the Apache 2.0 engine behind DeepSeek's own service, lists only NVIDIA SM100 and Huawei Ascend 950 hardware. Its decode benchmarks ran on 64 GPUs, so it's for providers, not your 3090. ([DeepSeek's repo](#))
 - **llama.cpp adds Soprano, an 80M English voice.** Soprano is a 79.7M-parameter, Apache 2.0 text-to-speech model that llama.cpp can now load. It's English-only with no voice cloning, and small enough to sit beside any LLM you run. ([model card](#))
-

Which Open Voice Models Fit a 3090

A voice assistant on your own box needs two models besides the LLM: one to turn your speech into text, and one to speak the answer back. Here are the ones you can download. Granite's multilingual update is the only one from this week: the Qwen models date from January and Fish Drama 3 from October 13. Every size below comes from the model's own card or file listing, not from our testing.

Speech-to-text

- **Qwen3-ASR-1.7B**. Alibaba's speech recognizer, listed at 2B parameters on its own card despite the name. It covers 30 languages plus 22 Chinese dialects. It's Apache 2.0, so commercial use is fine, and the weights are 4.7 GB.
- **IBM Granite Speech 5.0 TurboCTC**. A 470M-parameter recognizer that went from English-only to seven languages on October 21: English, French, German, Spanish, Portuguese, Italian and Japanese. It's Apache 2.0 and the weights are under 1 GB. Skip the NC version, which is English-only and non-commercial.

Text-to-speech

- **Fish Audio Drama 3**. A 35B mixture-of-experts model with 3B active, covering 83 languages. The full weights take 37 GB of GPU memory, so they don't fit. An [unofficial int4 build](#) needs 20.8 GB for weights, but it's only been tested on 80 GB and larger cards. Commercial use is free only while your company's total revenue stays under \$500,000 a year. Running it as a hosted voice service needs a separate licence.
- **Qwen3-TTS-0.6B**. Qwen's small voice-cloning model, listed at 0.9B parameters on its card. It speaks 10 languages, it's Apache 2.0, and the weights are 2.5 GB. A [community GGUF](#), not from Qwen, gets the q8_0 model and its codec to about 1.9 GB on CPU or GPU.

Mistral's new [Voxtral Mini TTS 3](#) isn't on the list because it's API-only, at \$13 per million characters.

By weight size alone, a 3090 holds either recognizer plus Qwen3-TTS and keeps most of its 24 GB for the LLM. Fish Drama 3 is the one that doesn't fit. If I were wiring one up today, I'd start with Granite and Qwen3-TTS: both are small and Apache 2.0, and Qwen3-ASR takes over if you need a language Granite lacks. I plan to measure a speech-to-text, LLM and text-to-speech pipeline on a 3090 for an upcoming piece.

📖 Our current voice-chat setup guide is [here](#). 📖 Our look at the March Voxtral release is [here](#).



Every number we've measured, with configs and raw logs: insiderllm.com/benchmarks/.

That's the week. If you want a voice you can ship in a product, the Qwen models and Granite's Apache 2.0 version carry no revenue cap.

New here, reading this on the web? [Subscribe](#) and the next one lands in your inbox.

— Mark, InsiderLLM

What are you using for speech-to-text at home, and what language do you need it in? Reply, or hit me at hello@insiderllm.com. I read everything.

A weekly email with every new guide and measured benchmark.

[Subscribe — free, no spam](#)

Source: <https://insiderllm.com/blog/newsletter-2026-10-26/>

Free guides for running AI locally