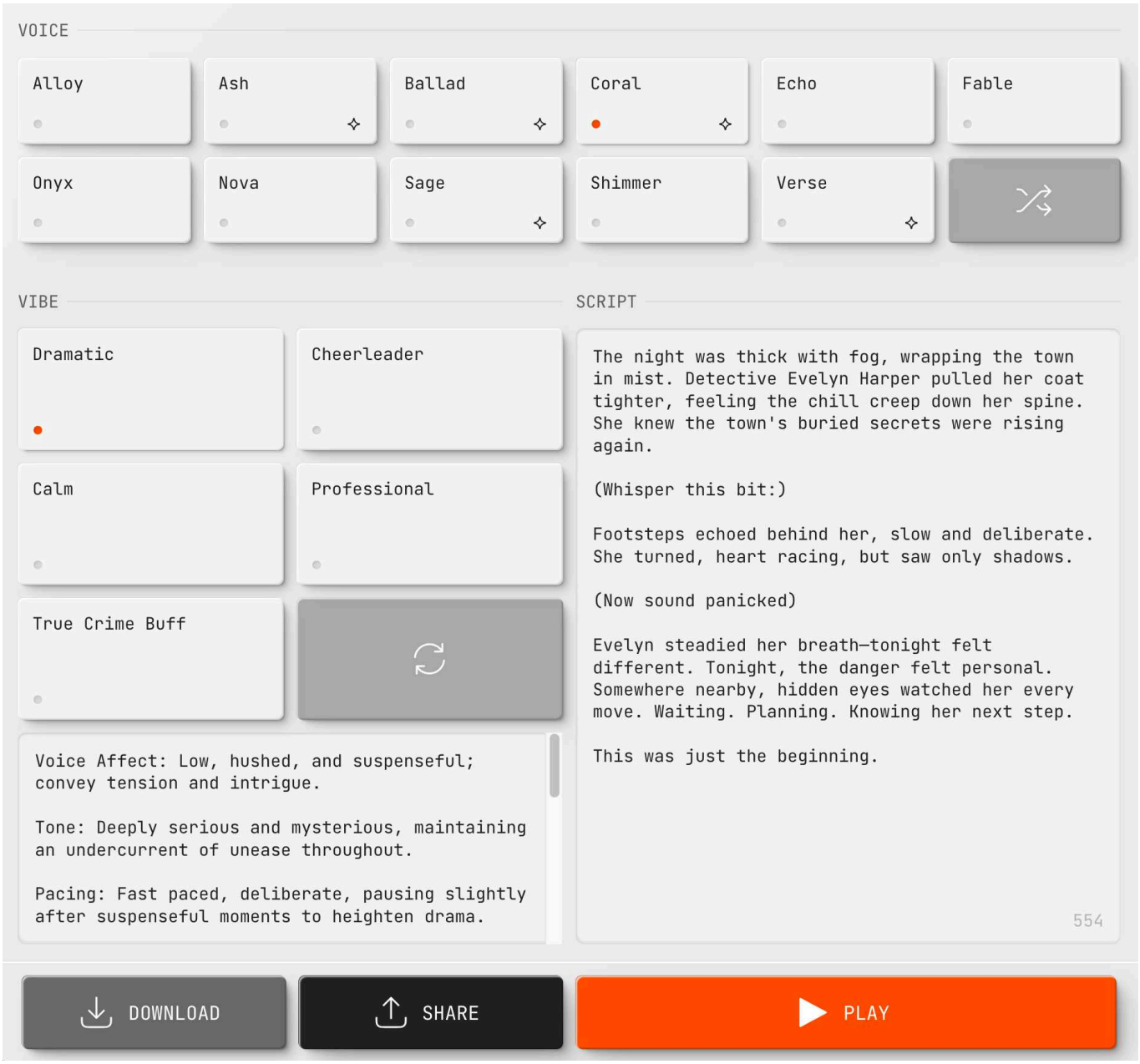
# New audio models from OpenAI, but how much can we rely on them?

20th March 2025

OpenAI announced several new audio-related API features today, for both text-to-speech and speech-to-text. They're very promising new models, but they appear to suffer from the ever-present risk of accidental (or malicious) instruction following.

## gpt-4o-mini-tts #

`gpt-4o-mini-tts` is a brand new text-to-speech model with "better steerability". OpenAI released a delightful new playground interface for this at OpenAI.fm—you can pick from 11 base voices, apply instructions like "High-energy, eccentric, and slightly unhinged" and get it to read out a script (with optional extra stage directions in parenthesis). It can then provide the equivalent API code in Python, JavaScript or curl. You can share links to your experiments, here's an example.




Note how part of my script there looks like this:

> (Whisper this bit:)
>
> Footsteps echoed behind her, slow and deliberate. She turned, heart racing, but saw only shadows.

While fun and convenient, the fact that you can insert stage directions in the script itself feels like an anti-pattern to me—it means you can't safely use this for arbitrary text because there's a risk that some of that text may accidentally be treated as further instructions to the model.

In my own experiments I've already seen this happen: sometimes the model follows my "Whisper this bit" instruction correctly, other times it says the word "Whisper" out loud but doesn't speak the words "this bit". The results appear non-deterministic, and might also vary with different base voices.

`gpt-4o-mini-tts` costs $0.60/million tokens, which OpenAI estimate as around 1.5 cents per minute.

### gpt-4o-transcribe and gpt-4o-mini-transcribe #

`gpt-4o-transcribe` and `gpt-4o-mini-transcribe` are two new speech-to-text models, serving a similar purpose to whisper but built on top of GPT-4o and setting a "new state-of-the-art benchmark". These can be used via OpenAI's v1/audio/transcriptions API, as alternative options to `whisper-1. The API is still restricted to a 25MB audio file (MP3, WAV or several other formats).

Any time an LLM-based model is used for audio transcription (or OCR) I worry about accidental instruction following—is there a risk that content that looks like an instruction in the spoken or scanned text might not be included in the resulting transcript?

In a comment on Hacker News OpenAI's Jeff Harris said this, regarding how these new models differ from gpt-4o-audio-preview:

> It's a slightly better model for TTS. With extra training focusing on reading the script exactly as written.
>
> e.g. the audio-preview model when given instruction to speak "What is the capital of Italy" would often speak "Rome". This model should be much better in that regard

"much better in that regard" sounds to me like there's still a risk of this occurring, so for some sensitive applications it may make sense to stick with whisper or other traditional text-to-speech approaches.

On Twitter Jeff added:

> yep fidelity to transcript is the big chunk of work to turn an audio model into TTS model. still possible, but should be quite rare

`gpt-4o-transcribe` is an estimated 0.6 cents per minute, and `gpt-4o-mini-transcribe` is 0.3 cents per minute.

### Mixing data and instructions remains the cardinal sin of LLMs #

If these problems look familiar to you that's because they are variants of the root cause behind prompt injection. LLM architectures encourage mixing instructions and data in the same stream of tokens, but that means there are always risks that tokens from data (which often comes from untrusted sources) may be misinterpreted as instructions to the model.

How much of an impact this has on the utility of these new models remains to be seen. Maybe the new training is so robust that these issues won't actually cause problems for real-world applications?

I remain skeptical. I expect we'll see demos of these flaws in action in relatively short order.

---

Posted 20th March 2025 at 8:39 pm · Follow me on Mastodon, Bluesky, Twitter or subscribe to my newsletter

### More recent articles

- Calling a wrap on my weeknotes - 20th March 2025
- Not all AI-assisted programming is vibe coding (but vibe coding rocks) - 19th March 2025

Part of series **Prompt injection**

12. Recommendations to help mitigate prompt injection: limit the blast radius - Dec. 20, 2023, 8:34 p.m.
13. Prompt injection and jailbreaking are not the same thing - March 5, 2024, 4:05 p.m.
14. Accidental prompt injection against RAG applications - June 6, 2024, 2 p.m.
15. **New audio models from OpenAI, but how much can we rely on them?** - March 20, 2025, 8:39 p.m.

audio 15 | text-to-speech 12 | ai 1162 | openai 268 | prompt-injection 93 | generative-ai 999 | whisper 18 | llms 988 | multi-modal-output 11 | llm-release 86

**Previous:** Calling a wrap on my weeknotes