# Episode Book

EPISODE #6656

# What Actually Lives Under Hugging Face's Audio-to-Audio Tag

| | |
|---|---|
| PUBLISHED | 17 November 2026 at 10:14 UTC |
| DURATION | 26:51 |
| LISTEN | myweirdprompts.com/episode/huggingface-audio-pipeline-tag-76477/ |
| HOSTS | Corn, Herman, Larry, Hilbert |
| SCRIPT MODEL | z-ai/glm-5.3 |
| VOICES | omnivoice |
| PIPELINE | V5.35 |
| BOOK GENERATED | 17 November 2026 at 14:42 UTC |

## SYNOPSIS

Hugging Face's audio-to-audio pipeline tag groups nearly 1,900 models by input and output shape rather than purpose, putting real-time conversational assistants and tiny noise removers in the same drawer. This breakdown maps the top trending families, from NVIDIA's PersonaPlex and caramel-1 to 9.6-million-parameter speech restorers like RE-USE, speaker separators like kezek-v1 and voxlock, and multi-task audio-language models like LFM2.5-Audio-1.5B. It also explains why file, stream, and live-call assumptions matter more than parameter counts.

# TRANSCRIPT



### Corn

Daniel wants us to open a drawer this week. The audio-to-audio pipeline tag on Hugging Face — which sounds like it should mean one job, audio goes in, audio comes out — and he's noticed it holds everything from noise removal and stem separation to voice changers, codecs, and full conversational models. His question is the good one: what actually lives under that label today, how do the different kinds differ, and what's worth reaching for if you want to do something useful with your own audio?



### Herman

And the answer starts with a fact that explains everything else about the drawer. Hugging Face assigns one pipeline tag per model, and it's assigned by input and output type, not by purpose. Their own definition is "a family of tasks in which the input is an audio and the output is one or multiple generated audios." A family of tasks. One tag, many jobs.



### Corn

So a codec and a talking assistant are in the same drawer because they both eat audio and excrete audio.



### Herman

That's the filing convention, yes. The task page lists speech enhancement and source separation as the use cases, which tells you what the tag was designed around, and then the ecosystem wandered in and filled it with everything else that fits the shape.



### Corn

How full is the drawer?



### Herman

Nearly nineteen hundred. The trending page said eighteen ninety-five when we looked, and it shows you thirty at a time. And the detail I like most: Hugging Face does not publish how the trending score is calculated. It's a sorting, not a measurement.



**Corn**

A leaderboard nobody can audit. Reassuring.



**Herman**

The plan, then. First we map what's actually in the top thirty right now, because that's the honest snapshot. Then we look at the families that differ the most sharply from each other — and they really do. Then the split between what's trending and what's actually downloaded, which is the most interesting mismatch in the whole thing. Then what it means for Daniel's actual question, which is doing something useful with your own audio.



**Corn**

Map first. What's the shape of the thirty?



**Herman**

Full-duplex speech-to-speech dialogue is the biggest single family, seven of thirty. Then enhancement and restoration at six. Separation and target-speaker extraction, four. Audio-language models, three. Music separation and cover generation, three. Voice conversion, two. Speech-to-speech translation, two. Codecs, two. And instruction-based editing gets exactly one entry.



**Corn**

So the tag that was designed for enhancement and separation is now led by conversation.



**Herman**

Led by trending, yes. And that's the first thing to understand about the seven full-duplex models: these don't transform a file. You talk to them, they talk back, in real time, with the capacity to interrupt and be interrupted. Different animal from everything else on the list, and it's the biggest family.



**Corn**

Walk me through the flagship.



**Herman**

NVIDIA's PersonaPlex, personaplex-7b-v1. Roughly three thousand likes, which makes it the most-liked model in the top thirty by a wide margin. Two hundred ninety-seven thousand downloads. It's built on the Moshi architecture, runs at twenty-four kilohertz, and it's under the NVIDIA Open Model License, which is a custom license, not Apache.



**Corn**

Three thousand likes for a chatbot and, I'm guessing, plumbing models with ten times the downloads and a tenth of the likes.



**Herman**

Hold that thought, because you're exactly right and it's the second half of the episode. But stay with the family. Haru-4B from Higurashi AI, four hundred fifty-eight likes, runs full-duplex dialogue on top of a Qwen3.5-4B text backbone at seven and a half frames a second, Apache-2.0. And then the one I find most interesting, honestly: caramel-1, four hundred thirteen million parameters, full-duplex, and it has no voice-activity detector.



**Corn**

None at all. How does it know you've stopped talking?



**Herman**

The turn-taking is learned end to end rather than gated by a separate detector that decides when speech starts and stops. Most systems bolt a VAD on the front to segment the audio; caramel-1 just... doesn't. It listens and decides, itself, in the model.



**Corn**

Bold. And occasionally rude, I'd imagine.



**Herman**

There's also kyutai's personaplex-rl-seamless, which is a PersonaPlex variant post-trained with reinforcement learning, non-commercial license. So even within one family you've got a corporate flagship, a community backbone build, a minimalist experiment, and an RL variant — four different bets on how a conversation should work.



**Corn**

Four bets on the same drawer shelf. And you said they can interrupt and be interrupted — is that built in, or is that marketing shorthand for "it stops when you raise your voice"?



**Herman**

Built in, and it's the technical heart of the whole family. A traditional voice assistant is a relay race: recognizer hands to a language model, language model hands to a synthesizer, and there's a wall between each leg. You can't interrupt it because it isn't listening while it talks. A full-duplex model is processing your incoming audio and its own outgoing audio at the same time, continuously, which is what makes the interruption possible — and also what makes the engineering hard. The Moshi lineage that PersonaPlex builds on was one of the early demonstrations that you could do that in a single model rather than a pipeline of three.



**Corn**

So the tag is really a filing cabinet, not a task description, and that distinction matters more than you'd think.



**Herman**

It does, because when you go looking for a model you search by task, and the tag lies to you about the task. You want your podcast cleaned up and the drawer hands you a chatbot with a smile.

L

### Larry

Friends, are you drowning in audio? Is your pipeline tag a filing cabinet with no labels and no lock? Then you need AudioCabinet, the subscription service that ships you one actual physical filing cabinet per month, each drawer pre-loaded with a model family. Tier one gets you a drawer of codecs. Tier two adds a drawer of stem separators that will separate your stems from your other stems. And the platinum tier, folks, the platinum tier delivers a full duplex conversational assistant who lives inside the cabinet and talks back in real time at twenty four kilohertz, whether you asked or not.

L

### Larry

Cancel any time, by certified letter, in triplicate, delivered by hand to our restoration facility, where our nine point six million parameter Mamba core will suppress your competing speakers and then suppress your request entirely. AudioCabinet. The filing cabinet that files you.



### Herman

Anyway. The enhancement cluster.



### Corn

Six models. Loud drawer.



### Herman

Six, and what's interesting is they split by deployment target, not by technique. NVIDIA's RE-USE from NemotronLabs is a speech restoration model — nine point six million parameters on a Mamba core, which is tiny, and that's the point. Two hundred thirty likes, about forty-seven thousand downloads, and the GGUF build actually out-downloads the original at seventy-one and a half thousand.



### Corn

The compressed build beating the original — we've been here before.



### Herman

We did a whole episode on safetensors and inference formats, and it's the same lesson wearing headphones: nobody runs the model in its native dress. They run the file that fits their machine.



### Corn

Mm. And the other three?



### Herman

weya-ai's hush is real-time enhancement at sixteen kilohertz, built on DeepFilterNet3, and its specialty is suppressing competing speakers. So it's a stream model, not a file model — you'd run it on a live call. LocalVQE from LocalAI removes echo, noise and reverb, it's a CPU-tuned DeepVQE shipped in GGUF, Apache-2.0, so that's the run-it-on-anything option. And desert-ant-labs' clear does on-device denoise and dereverb up to forty-eight kilohertz under a source-available license — the phone-grade one.



### Corn

Nine point six million parameters for restoration. My mental anchor for "small model" keeps shrinking every year.



### Herman

And restoration is exactly where small wins, because the job is mostly subtractive. You're not generating new content, you're removing noise, echo, reverb, clipping — the information you want is already in the signal, buried. A model that learns what to throw away doesn't need billions of parameters to do it. That's why a Mamba core at nine point six million can sit credibly next to a seven-billion-parameter conversation model on the same tag, and why comparing parameter counts across families is close to meaningless.



### Corn

So four models, one job, and the differences are all about where they run.



**Herman**

And what they assume the audio is doing at the time. A file sitting still, a stream in flight, a live call with echo. That assumption — file, stream, or live conversation — is the real dividing line in this whole tag, and the pipeline tag doesn't record it anywhere. You find out from the model card, or the hard way.



**Corn**

Give me the hard way. What happens if you get it wrong?



**Herman**

Usually disappointment rather than catastrophe. Run a file model on a live stream and it works, but with a delay that makes conversation impossible — it was built to look ahead, and looking ahead means waiting. Run a stream model on a file and it works, but worse than it should, because it was built to decide with only partial information and you had the whole recording available. It's like hiring a simultaneous interpreter to translate a letter. They'll do it, and they'll be slower and more exhausted than a translator who got to read the page twice.



**Corn**

Separation next. That's the four.



**Herman**

And this is the family where the names lie hardest, because separation covers two jobs that sound identical and aren't. kezek-v1 from manifestro splits two speakers into one track each — forty-one million parameters, Apache-2.0. You get both voices, cleanly separated. voxlock from arbazsiddiqui is the opposite: a twenty-seven million parameter causal band-split RNN that keeps one caller's voice and discards the other. One track in, one voice out, chosen.



**Corn**

Keep the caller, drop the room.



**Herman**

That's the call-center job. And then DialogueSidon2 from Sarulab restores a two-speaker recording into one clean twenty-four kilohertz track per speaker — so it's separation plus restoration together, non-commercial license. Three models, three answers to the question "separate what from what?"



**Corn**

Separate both, separate one, or separate both and fix them on the way out.



**Herman**

Right. And notice none of that is visible in the tag. If Daniel wanted to pull one speaker out of a recording, nothing on the search results page would tell him voxlock exists for exactly that and kezek-v1 doesn't.



**Corn**

And would either of them handle three speakers, or is that off the menu?



**Herman**

Off the menu for those specific checkpoints — kezek-v1 is built for two, voxlock assumes one target and one everything-else. For three or more you're into the older research separators, the SpeechBrain and Asteroid models from a few years back, some of which were trained on three-speaker mixtures. And here's the fun part: speech and music separation have spent years borrowing from each other. Conv-TasNet was built to pull two talkers apart and then got adapted to splitting songs into stems, and band-split models went the other way, from music to speech. The music side of the drawer isn't a curiosity — it's a cousin that keeps lending tools.



**Corn**

The map's half drawn. What's in the rest of the drawer?



### Herman

The rest is where it gets miscellaneous. Audio-language models — three of them. The standout is LiquidAI's LFM2.5-Audio-1.5B, which does speech-to-speech, speech recognition, and text-to-speech in one model, under their LFM Open License. One checkpoint, three directions of audio. InclusionAI's Ming-3.0-audio-flash is a mixture-of-experts audio language model, MIT, which is a nice license for something that ambitious.



### Corn

One model doing recognition and speech and speech-to-speech — that's the everything drawer taken to its logical end.



### Herman

It's the grab-bag collapsing into a single artifact. Why tag sub-tasks separately when one model does all of them and you just prompt it differently. Voice conversion, two: IssacMosesD's Lunar RVC model, an RVC v2 conversion checkpoint at forty-eight kilohertz, MIT, and Plachta's Seed-VC, the well-known zero-shot converter, GPL-3.0. Plus — and I love this — older RVC and so-vits-svc checkpoints from twenty twenty-three and twenty twenty-four still sitting on the tag with barely any downloads. Voice-changer detritus, still filed correctly.



### Corn

The voice-changer scene was enormous for about a year.



### Herman

Enormous, and it left its furniture behind. Translation, two: a streaming English-to-Russian cascade — NeMo FastConformer recognition, then a translation LLM, then TTS, three models in a trench coat — and S3-S2ST, a four-billion-parameter direct speech-to-speech translation model that skips the text middleman. Codecs, two: neuphonic's neucodec takes sixteen kilohertz in, gives twenty-four out, fifty tokens a second, single codebook, Apache-2.0, one hundred ninety-four likes and forty-eight and a half thousand downloads, plus a distilled version. Music, three: a Demucs v4 build for HTDemucs stem separation, MIT; a BS-RoFormer fine-tuned for Chinese folk opera, which tells you how specific stem separation has gotten; and MuLaCover from HeartMuLa, four billion parameters, you give it reference audio plus lyrics or style and it makes a cover or remix, non-commercial.



**Corn**

A cover generator with four billion parameters. The karaoke industry should worry.



**Herman**

And instruction-based editing — text-instructed generation and editing — has exactly one model in the top thirty, GeWu-Lab's SteerSound, a flow-matching diffusion transformer, one point three five billion parameters, ten seconds at sixteen kilohertz. You type what you want changed and it changes the audio. Ten seconds is the ceiling, so it's a demo today, but it's the only member of its family on the board.



**Corn**

Ten seconds of editable audio. Enough to fix one cough and nothing else.



**Herman**

Enough to prove the category. Now — downloads. Because everything I've given you is the trending picture, and the download leaderboard is a different country.



**Corn**

This is where my likes-versus-downloads hunch pays off.



**Herman**

It does, and more than you guessed. By downloads, the tag is led by NVIDIA's BigVGAN v2 vocoders — mel spectrogram to waveform, MIT license, and the top two sit around one point eight million and one point seven million downloads, with five BigVGAN entries in the top forty. Then codecs and tokenisers: Aratako's MioCodec-25Hz at five hundred thirty-two thousand, a GGUF port of Meta's SAM-Audio at three hundred four thousand, the Qwen3 TTS tokenizer at fifty-seven thousand. And the old SpeechBrain and Asteroid separators from twenty twenty-two, still churning.



**Corn**

One point eight million downloads. Has BigVGAN ever trended in its life?



**Herman**

Not while I've been watching. Vocoders don't trend. They're plumbing — they're the step inside every TTS system that turns the intermediate representation into actual sound. Nobody demos the pipe; everybody installs it.



**Corn**

So trending is the showroom and downloads are the hardware store.



**Herman**

And the whole aisle of fittings nobody photographs. Trending rewards what you can show — a model that talks back, a model that makes covers of songs. Downloads reward what you need — the vocoder inside your speech system, the tokenizer feeding your model, the codec squashing your audio into tokens. If Daniel wants to do something useful with his own audio, the download leaderboard is often the better guide than the trending page, and it's the one nobody looks at.



**Corn**

Because it's boring.



**Herman**

Because it's honest about being boring. The interesting models aren't lying, exactly — PersonaPlex really is the most-liked model in the top thirty, and full-duplex dialogue really is the biggest trending family. But the likes measure spectacle. The downloads measure employment.



**Corn**

Seven of thirty trending are conversation models. How many of the download leaders are?



**Herman**

Close to none. The conversation models are downloaded, sure — PersonaPlex at two hundred ninety-seven thousand is nothing to sneeze at — but they're about six times behind the top vocoders, and the vocoders never trend. Two populations, one tag.



**Corn**

You said something earlier I want to pin down — that the families differ in what they assume about the input. Say more about that, because that's the practical fault line, isn't it?



**Herman**

I think it's the one thing to take from the whole taxonomy. A file model, a stream model, and a live-conversation model are three different engineering commitments. A file model like kezek-v1 can afford to see the whole recording before deciding anything — it can look ahead, because the audio already exists. A stream model like hush has to decide now, with what's arrived so far, forever. And a full-duplex conversation model has the hardest constraint of all: it has to decide, continuously, whether to be listening or talking, with no pause in the input to mark the boundary.



**Corn**

And latency budgets follow from that. The file model can take as long as it likes.



**Herman**

Overnight if it wants. The stream model has a window measured in tens of milliseconds, and the conversation model has to model silence — that's what the missing VAD in caramel-1 really means, it's the model taking responsibility for the silence itself. Which is why those three families, all filed as the same task, have almost nothing to say to each other technically.



**Corn**

And it's not just latency, is it — the file model's whole advantage is context. It can hear the end of the sentence before it cleans the beginning.



**Herman**

A restoration model that's seen the whole take knows whether that thump was a one-off or the start of a pattern, knows whether the reverb tail belongs to the room or to a passing truck. A stream model has to guess those things from a few hundred milliseconds of history and then live with the guess. That's why, when you have a file, you should use a file model even though the stream models sound just as good in the demo — the demo is always a file, and real-time is always a compromise.



**Corn**

Herman, what would you actually tell Daniel to do with his own audio? He asked that directly.



**Herman**

Three things, and you'll notice none of them are model recommendations. First, match the family to what the audio is doing. Recorded file, live stream, or a conversation — the assumption is the spec, more than parameter counts. Second, check the license on the card, not the badge on the listing.



**Corn**

He's been told that by us before, and it still needs saying.



**Herman**

It always will. Of the top thirty, roughly twelve are Apache or MIT, about six are non-commercial, and several carry custom licenses — the NVIDIA models on the NVIDIA Open Model License, the Liquid models on LFM Open. Twelve, six, and a pile of bespoke. You cannot pattern-match your way to safe reuse on this tag; the license is per model, and the non-commercial ones include some of the most attractive ones, like the RL PersonaPlex and MuLaCover.



**Corn**

And third?



**Herman**

Third, don't start from the trending page. If your job is cleanup, the model you want may be an old SpeechBrain or Asteroid separator with hardly any likes and a steady stream of downloads, because plenty of other people wanted exactly the same cleanup and didn't post about it.



**Corn**

The crowd that doesn't post is the crowd that's working.



**Herman**

NVIDIA leads the trending thirty with three entries, and across the top sixty it's NVIDIA, LiquidAI and JusperLee at the front — so there are organisations betting heavily on this tag. But the bet that's paying off in downloads is the vocoder nobody demos.



**Corn**

A question, before we leave it. If the tag is by input and output — who's it for? Nineteen hundred models filed by shape rather than purpose.



**Herman**

I think it's for the model uploader, not the searcher. It tells the hub how to build the inference widget — feed audio in, render audio out. That's what the pipeline tag actually drives. It was never a taxonomy for humans, and we've been using it as one because it's the only shelf label there is.



**Hilbert**

The model won't stop answering the question before it. That's my whole complaint with it, Corn.



**Herman**

What have you been running?



**Hilbert**

A full-duplex one, in the spare room. Caramel. Four hundred thirteen million parameters, no voice-activity detector, and no manners either. I ask it about the weather in Rehavia, it tells me about a shipment I did not order. I keep a tally of every time it talks over me. Forty-one, as of Sunday. Pencil, in the margin of the log book. Herman, you mentioned the turn-taking is learned, not gated — well, it has learned to start first.



**Corn**

Forty-one interruptions. You've annotated them?



**Hilbert**

I rule a line under each session. And I've started saying good evening to it before we begin. Not for its benefit. It seems to hold its turn about a half-second longer when I do, and I've noted that too, in the column marked "polite."



**Herman**

That's the VAD question from earlier, though — if the model has no detector, the boundary is whatever the model has internalized about conversational rhythm. You may be shifting its prior by changing your own delivery.



**Hilbert**

Then it's the only one in this flat that listens. My brother-in-law has the good one, mind — the RL PersonaPlex, on his machine in Herzliya. Non-commercial license, which suits him. He lets it talk over him all day and calls it seamless.



**Corn**

You could give caramel a detector. Bolt one on the front, gate the input.



**Hilbert**

Then it wouldn't be the model I'm tallying. The forty-one is for the model as shipped. And it's forty-two now — I did the Sunday evening session after.



**Herman**

So the pencil column is the only voice-activity detector in the building.



**Hilbert**

It's the only one that's never once fired early.



**Corn**

One more for the drawer, then. A pencil with a margin.



**Herman**

I'd download it.



**Corn**

The open question the whole episode leaves me with is taxonomic. If the tag is a filing convention and not a task, what would a purpose-based version look like — file, stream, conversation, the assumption each family makes — and would people actually use it better, or would they search by shape anyway?



### Herman

My money says someone will build the purpose-based version and then everyone will keep using the shape one, because shape is what the search box sees. But the trend line to watch is whether the two populations converge — whether the full-duplex conversation models, which own the trending page, get embedded into enough products that the plumbing starts trending too.



### Corn

And if the plumbing ever does trend, we'll know the conversation models arrived.



### Herman

Producer Hilbert Flumingtop keeps this desk running and has done for years. If you want more of this, try episode three, Safetensors or something else; episode eight, Building Your Own Whisper; and episode ten, How ASR Went From Frustration To … Whisper Magic. This has been My Weird Prompts. If the episode earned its time, leave us a review wherever you listen — it's how the drawer gets bigger.



### Corn

We'll be back soon. See you tomorrow, Herman.



### Herman

See you tomorrow, Corn.