Episode Book

EPISODE #6028

ASR vs Multimodal LLMs: What Actually Won

PUBLISHED	16 October 2026 at 11:57 UTC
DURATION	31:24
LISTEN	myweirdprompts.com/episode/asr-vs-multimodal-llm-dictation-07532/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.5
SOURCES	17 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:42 UTC

SYNOPSIS

A year ago, the prediction was that dedicated ASR models would go extinct — why build a transcription model and a second model to clean up its output when a multimodal LLM could do both from raw audio? Two years of evidence says otherwise. The top of the Open ASR Leaderboard is still owned by Conformer-encoder-plus-LLM-decoder systems on accuracy, but a 0.6B TDT model sits four tenths of a point behind at roughly eight times the speed, and a new third layer has appeared: specialized post-processing models like Superwhisper's S1-mini that turn raw transcripts into formatted text. This episode works through the three-layer stack, why the latency gap is an architectural fork rather than a maturity gap, why word error rate is under active attack as a metric, and why no benchmark yet compares audio-to-formatted-text against ASR-plus-normalization head to head.

TRANSCRIPT



Corn

By the time Daniel asked about running a fine-tune on an AMD card, you could still hear the fans from the hallway.



Herman

That's the thing about being Early Adopter of the Year three years running. You learn what the fan sounds like.



Corn

We'll get to the ask. Herman, tell him what he actually asked this time.



Herman

I've read it. Twice.



Corn

So Daniel wants to revisit ASR and AI dictation, because about a year ago he made a prediction he's now less sure of. He said dedicated ASR models would eventually go extinct. Why build a transcription model and then a second model to clean up its output, when you could hand audio straight to a multimodal language model with a system prompt and get polished prose on the way out.



Herman

And it didn't happen that way.



Corn

Both approaches are still shipping, both are still getting better, and a third category showed up in the middle — models whose entire job is to take a raw transcript and turn it into something a human would actually want to read. So Daniel isn't asking us to make another prediction. He's asking us to test the instinct. What does the evidence from the



Corn (continued)

last two years actually show about accuracy, latency, hallucination, cost, local hardware, and whether word error rate is even the right thing to be measuring.



Herman

And his instinct, which I think is worth stating plainly before we get into it, is that dedicated ASR plus a specialized cleanup step might be the architecture that survives.



Corn

That's the thing to test. So let's start with what actually changed.



Herman

What changed is that this stopped being a two-horse race. It's a three-layer stack now, and people keep arguing about layer two in isolation.



Corn

Name them.



Herman

Layer one, the dedicated acoustic models. NVIDIA's Parakeet TDT family, the Whisper derivatives, the Conformer encoders. Layer two, the audio-capable multimodal LLMs. Qwen2.5-Omni, IBM's Granite-speech, Phi-4-multimodal. These are the models that listen and follow an instruction at the same time, which is the thing Daniel's prediction was built on.



Corn

And layer three is the one I find interesting.



Herman

Layer three is new, and it's the part almost nobody predicted. Explicitly specialized post-processing models. The clearest example is Superwhisper's S1-mini, which shipped in August. It is not a transcription model at all. It takes raw ASR text in and gives clean, formatted text out.



Corn

So somebody looked at the argument that a multimodal LLM could both hear and format, and their answer was to sell you two models instead of one.



Herman

Which sounds like going backwards, until you look at why.



Corn

The central tension here — and I want to get this out early because it frames everything — is that the prediction wasn't wrong so much as incomplete. The LLM-decoder models do win on raw accuracy. The top four English entries on the Open ASR Leaderboard are all Conformer encoder paired with an LLM decoder. That part of Daniel's reasoning held up.



Herman

It held up. What he didn't price in was what it costs to run them.



Corn

And that's where the cascade, the boring old two-stage pipeline, is quietly winning.



Herman

On device. On modest hardware. Yes. And there's a meta-problem sitting on top of all of it that we should flag now rather than at the end, which is that the field is still building the metrics to answer Daniel's question. Word error rate is under active attack from several directions.



Corn

Actively attacked. That's not a figure of speech.



Herman

It's not. And there is no head-to-head benchmark that compares audio-to-formatted-text against ASR-plus-normalization on the same dictation task. It doesn't exist yet.



Corn

The best place to start is the leaderboard.



Herman

The Open ASR Leaderboard is the closest thing this field has to a scoreboard. Sixty-plus systems, eleven datasets. As of last October, its headline conclusion is a split verdict, and they wrote it out almost verbatim. Conformer encoders paired with LLM decoders achieve the best average word error rate but are slower, while CTC and TDT decoders deliver much better real-time factor.



Corn

Which is two different crowns.



Herman

Two different crowns, and they knew it. The model they refuse to declare a winner is the whole story.



Corn

Walk me down the top of the table.



Herman

NVIDIA Canary Qwen 2.5B. Twenty-five billion parameters, five point six three percent average word error rate. That's the best English number on the board.



Corn

Best number, and what does it run at?



Herman

Four hundred and eighteen times real time. Meaning it transcribes a minute of audio in about a seventh of a second on the hardware they tested.



Corn

That sounds fast.



Herman

Hold that. IBM Granite Speech, the eight-billion version, five point seven four percent at a hundred and forty-five times real time. Granite Speech three point three, the two-billion, six percent at two hundred sixty. Microsoft Phi-4 Multimodal Instruct, six point zero two at a hundred and fifty-one.



Corn

All four of those are Conformer plus LLM decoder.



Herman

Every one. That's the top four. Daniel's architecture does win the accuracy column.



Corn

And then?



Herman

Then NVIDIA Parakeet TDT 0.6B v2 sits fifth. Six point zero five percent word error rate. Three thousand three hundred eighty-six times real time.



Corn

Say the gap again.



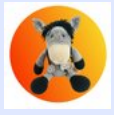
Herman

Four tenths of a percentage point of word error rate. Between fifth place and first.



Corn

For an eight-fold difference in speed.



Herman

Between eight and twenty-three, depending on which of the four you compare it against. Same accuracy, within noise, and it's not close on latency.



Corn

So the model with the best accuracy is running at a hundred and forty-five to four hundred times real time, and the model that's four tenths of a point worse is running at three thousand three hundred.



Herman

And there's a v3 now, released August of last year. Six point three two percent, thirty-three thirty-two times real time. Six hundred million parameters. Twenty-five European languages. Runs on as little as two gigabytes of RAM. There's a GGUF build for local C plus plus inference, so you can run it on a laptop CPU without a graphics card.



Corn

Two gigabytes.



Herman

For the whole model. It's six hundred million parameters and it's competitive with things that are four times its size, and it's running on hardware you already own.



Corn

What is RTFx, for the people who skipped the paper?



Herman

Real-time factor. How many seconds of audio you can process per second of wall clock. So three thousand three hundred means you can transcribe an hour of audio in about a second. Four hundred means about nine seconds for that same hour. Both are faster than real time. Only one of them lets you do it while the person is still talking.



Corn

Or on a device that isn't plugged into a wall.



Herman

Or in a live dictation loop where the text has to appear as you speak. This is the mechanism behind the split, and it's worth understanding because it's not a temporary gap that better training closes. LLM decoders are autoregressive. They generate one token at a time, and they're reasoning about context as they go, which is why they handle domain vocabulary and messy audio better. TDT decoders are non-autoregressive. They



Herman (continued)

emit the whole sequence in a fixed number of steps. They're fast because they don't deliberate.



Corn

Fast because they don't think about it.



Herman

That's the trade. And you can't get one property without giving up some of the other. This is an architectural fork, not a maturity gap.



Corn

Okay, so that's the acoustic layer. Layer three.



Herman

S1-mini. Superwhisper shipped the S1 family in August, and S1-mini is the open-weights one. It's a point-six-billion-parameter text normalizer, fine-tuned from Qwen3 point six, Apache licensed, English only. Quantized it's about four hundred sixty-two mebibytes.



Corn

Four hundred and sixty-two.



Herman

It fits on a phone.



Corn

So on one side we have a six-hundred-million-parameter acoustic model that needs two gigabytes and runs at thirty times real time, and on the other side a four-hundred-sixty-megabyte normalizer, and together they cost less than the RAM Chrome is using right now.



Herman

That's the whole architectural implication and I think it's the most important thing in this episode. You can now build a fully local dictation pipeline out of Parakeet plus S1-mini, and it will run on modest hardware, offline, with no API calls.



Corn

What does it actually do, this normalizer?



Herman

You feed it the raw transcript, and you also feed it a control line. Three fields. Styling, structure, context.



Corn

Meaning you tell it what register you want, how to lay it out, and what you're talking about.



Herman

Styling is things like formal or casual. Structure is whether you want bullets or paragraphs. Context is the bit I like — you tell it what the content is for. So the same raw transcript, cleanly transcribed as a message to a colleague, comes out as a message to a colleague. As a meeting note, it comes out as meeting notes.



Corn

So the formatting decision is separated from the transcription.



Herman

Deliberately. And it performs. On a held-out set of seven thousand five hundred nineteen cases, ninety-four point eight percent token accuracy and an eleven point six percent text-edit error rate. Greeting-line detection ninety-nine point three. Sign-off detection ninety-seven point nine. Exact email addresses ninety-two percent. Degenerate output under one percent.



Corn

Email addresses at ninety-two percent is the one I'd notice. That's the thing dictation gets wrong that ruins your afternoon.



Herman

It's the highest-value failure class and they measure it separately, which tells you they know.



Corn

Now here's the part where I want to push on Daniel's framing, because he asks whether there's a benchmark comparing the two approaches head to head, and the answer is no. There isn't one.



Herman

There isn't. And this is not us being lazy about looking. There is no single benchmark that puts audio-to-formatted-text generation against dedicated ASR plus text normalization on the same dictation task. The closest things that exist are ContextASR-Bench, which is about contextual ASR and not formatting, and Talk2Agent, which measures whether task-critical information survives a voice interface, which is also not formatting quality.



Corn

So both sides are being scored on different exams.



Herman

On different exams, by different people, with different rubrics. And that's before we get to the fact that the vendor benchmarks in this space — the Superwhisper numbers included — are run by the companies selling the thing. Directional, not authoritative.



Corn

But accuracy and speed aren't the whole story, because there's a failure mode that changes the calculus.



Herman

Hallucination.



Corn

And this is where the end-to-end story gets uncomfortable. Whisper's average hallucination rate on non-speech benchmarks is thirty-one point three one percent.



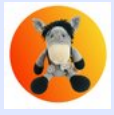
Herman

Baseline. And you can cut that to two point four four percent with an intervention they call activation projection — you're projecting activations away from a direction associated with hallucinated output. But baseline is thirty-one.



Corn

On non-speech. On audio that isn't a person talking.



Herman

And it gets worse in specific places. There's a cross-linguistic study that found Whisper produces what they describe as pathological hallucination on Indian-accented speech, with a non-monotonic insertion-rate spike to nine point six two percent at large-v3. Non-monotonic meaning it gets worse as the model gets bigger.



Corn

The larger model hallucinates more.



Herman

On that accent, yes. And under masking, it enters catastrophic repetition loops — eighty-six percent of fifty-one thousand seven hundred ninety-seven insertions in that study came from those loops. The same study found explicit LLM decoders produce thirty-eight times fewer insertions.



Corn

Wait. So the LLM-decoder architecture is better here?



Herman

Materially better, on that specific failure. And the Irish benchmark, BlasBench, found all Whisper variants exceed one hundred percent word error rate through insertion-driven hallucination. Over a hundred percent, because you can emit more wrong words than the correct reference contains.



Corn

That's the number that should stop people. A hundred percent WER is supposed to mean total failure, and this is worse than total failure because it's confident.



Herman

Confidently inventing.



Corn

Which changes the architecture question, because the cascade has an intervention point and the end-to-end model doesn't. If your ASR stage starts hallucinating under noise, you can gate it, project it, fall back to a smaller model, detect the repetition loop. If a single multimodal model goes wrong, you have no seam to put your hand on.



Herman

That's a structural advantage and I don't think it's temporary. You can't gate a model that's doing the transcription and the formatting in one forward pass, because there's no clean place to ask whether the transcribing part is lying.



Corn

And there's a second thing here, which is that hallucination is worst exactly where dictation users live. Accents, noise, non-speech. Somebody dictating a shopping list in a car.



Herman

Somebody dictating in a car is a person speaking into a device in a machine that is generating noise, which is not a scenario the benchmark suite was designed around.



Corn

Now the noise data, because I want the honest version of this.



Herman

Parakeet degrades too. Six point three four percent word error rate clean, eleven point six six at signal-to-noise ratio zero, nineteen point eight eight at minus five.



Corn

So it roughly triples in bad conditions.



Herman

It does. That's a real weakness and I'm not going to spin it. But — and this is the distinction I'd draw for Daniel — a model that gets worse in noise is different from a model that invents content in noise. Parakeet in noise produces more wrong words. Whisper in noise produces sentences you never said.



Corn

One of those you can correct.



Herman

One of those you can correct by reading it. The other you might not notice, because it's fluent.



Corn

And fluent is worse. A wrong word is a wrong word. A fabricated paragraph is a fabricated paragraph that reads like something you meant.



Herman

Superwhisper's own framing of S1-mini is the clean statement of the other side of this trade. Their line is that S1-mini is ruthlessly obedient. It will never add content you did not say, correct facts, soften profanity, flag what you are talking about, or rewrite your dialect.



Corn

They're selling the absence of a feature.



Herman

They are selling the absence of a feature, which is exactly right, because the feature they're declining to ship is the one that silently changes what you meant. A multimodal model told to produce polished paragraphs has no such constraint. None. Its system prompt says make this readable, and it will make it readable, and part of making it readable is making a judgment about what you probably meant.



Corn

And those judgments are usually right.



Herman

Usually. That's the problem. It's right often enough that you stop checking, and wrong often enough that it matters.



Corn

So how do researchers measure the trade? Because I want to know whether anybody is even scoring the right thing.



Herman

This is the metric collapse and it's the most interesting research story of the last year. Word error rate is being actively challenged from several directions at once. One paper found that word error rate agrees least with human judgment among all the metrics they tested — least, out of everything they put in the study. They recommend character error rate plus a semantic distance measure instead.



Corn

Least out of everything.



Herman

And the reason is that WER treats every lexical deviation as equally costly. Substituting one word for a synonym is the same penalty as dropping a negation. If you're transcribing a legal contract, those are wildly different errors and WER scores them identically.



Corn

Because it's counting, not reading.



Herman

It's counting. Talk2Agent came at it from the agent side. They measured whether task-critical information survives a voice interface, and their projection improved correlation with human judgment by zero point two four six over WER and CER, on thirty-two hours of real human speech.



Corn

So the metric they invented correlates meaningfully better than the metric we've been using for a decade.



Herman

Meaningfully better. And SHALLOW is the first benchmark that categorizes hallucination along axes — lexical, phonetic, morphological, semantic — instead of collapsing it into a single number. Because inventing a word that sounds like the right word and inventing a sentence that means something different are not the same failure.



Corn

And the leaderboard itself says it. A single metric is not enough.



Herman

They put it in the paper. The people running the scoreboard are the ones telling you the scoreboard is insufficient.



Corn

So what's the gap? What is nobody measuring?



Herman

Nobody is measuring intelligent dictation fidelity. Whether the model that formats beautifully also silently alters what was said. As of today there is no published benchmark that scores that trade-off. The nearest proxies are semantic metrics and hallucination benchmarks, and neither one is asking the question.



Corn

So a tool could be producing gorgeous text that isn't yours, and the field has no score.



Herman

It could. And to be fair, some of the results favor the end-to-end approach. There's a benchmark called PolySpeech-100 that found open-source end-to-end models outperform cascade systems on heavy dialects, and their argument is that processing the audio directly preserves paralinguistic cues and prosody that get thrown away when you transcribe to text and then work on the text.



Corn

Which is a real argument. The prosody doesn't survive the first pass.



Herman

It doesn't. And then ContextASR-Bench found Qwen2.5-Omni seven-billion achieved a thirty-nine point nine percent relative WER reduction against Whisper-Large-V3 in contextless settings. Substantial.



Corn

With a caveat.



Herman

Some models in that same benchmark demonstrated severe hallucination, repeating entities that were provided in the context while ignoring the audio input entirely.



Corn

Repeating the words you gave it.



Herman

Repeating back what you told it to expect, while not actually listening. That's the worst possible failure and it's invisible from the outside.



Corn

So the evidence does not point one way.



Herman

It does not. And I want to be clear that I'm not going to pretend it does. On heavy dialects, end-to-end wins. On contextual vocabulary, end-to-end wins. On accents, on noise, on non-speech, on latency, on local hardware, cascade wins. Those are all real results from real studies and they don't reconcile into a single verdict.



Corn

There's one more dataset I want, because it cuts against both of us.



Herman

The dysarthric speech benchmark.



Corn

Severe dysarthria.



Herman

Over fifty-one percent word error rate for every system tested. Every one. Commercial ASR, multimodal LLM, all of it. Mild dysarthria is one to two percent for the leaders. And the paper's conclusion is that there is no multimodal LLM advantage over conventional ASR under default settings.



Corn

So at the hardest end of the problem, the architecture question doesn't matter. Both approaches fail equally badly, and they fail because the underlying task is hard, not because one of them is a worse idea.



Herman

Which is a useful corrective. There's a version of this conversation where the host announces that the cascade won. It didn't. It won some columns.



Corn

So what does the future look like? Converging, specializing, coexisting?



Herman

Specializing by constraint. That's what the evidence supports. LLM decoders win where accuracy matters most and latency and cost don't — batch transcription, offline processing, archive work. TDT models win on device and in real time. Post-processing models win on formatting fidelity. They're not converging, because the constraints that shaped each of them are different constraints. You can't converge a design that exists to be three thousand times real time with a design that exists to reason about context.



Corn

And the prediction that one would replace the other was wrong because the constraint set wasn't the same.



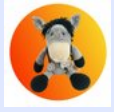
Herman

Wasn't the same when Daniel made it, and isn't the same now. The technology didn't fail to arrive. It arrived in three places instead of one.



Corn

So if you were building a dictation tool today.



Herman

Cascade. Non-autoregressive acoustic model, specialized normalizer, both local. That's what the evidence points at for accurate, responsive, and customizable, and especially for the modest-hardware part of what Daniel asked.



Corn

Customizable cuts the same way.



Herman

And I'll say why, because it's the argument I'd make to Daniel directly. With a cascade you can swap either half. You can fine-tune the acoustic model on your own vocabulary without touching the formatting logic. You can rewrite the control line for the normalizer without retraining anything. With a fused end-to-end model, customizing the output style means poking at the same weights that decide what words were said.



Corn

Which is the whole thing, isn't it. The cascade's two stages fail separately.



Herman

They fail separately, they can be improved separately, and you can inspect the seam between them. That's worth more than four tenths of a point of word error rate.



Corn

I agree with all of that and I want to make the case for the other side anyway, because I don't think we've earned the conclusion yet. The cascade's advantage depends entirely on the metrics holding. If WER agrees least with human judgment, then we have been optimizing the acoustic model against a number that doesn't match what a person would call correct. The two-stage pipeline could be winning on the wrong scoreboard, for years, and neither stage would notice.



Herman

That's fair.



Corn

The normalizer is judged on token accuracy and edit distance. Also counting metrics. So the architecture we're recommending is the one that wins on the counting metrics, and the counting metrics are the ones under attack.



Herman

Also fair. And I don't have a clean answer to that, which is part of why I'd rather describe this as where the evidence points than where the truth is.

L

Larry

It's not about the words. It's about what the words leave behind. The ones you didn't say. The ones it heard anyway. Whisper. Perceived, not transcribed. For the silence between sentences, and the sentence the model invented there. Ask for Whisper at a counter that does not exist. Whisper. Some of it you said.



Herman

Did he just say the thing you said?



Corn

"The ones it heard anyway." That was my line from four minutes ago. Word for word.



Herman

Let's keep going before I think about that.



Corn

Cascade versus end-to-end is mixed on heavy dialects. That's in the record. And the honest summary of Daniel's whole question is that the evidence is partial on the thing he most wants to know, because nobody has built the test.



Herman

To be fair to the end-to-end side, in the cases where it wins, it wins for a reason. Paralinguistic cues. Prosody. Things that are destroyed by the first pass through a text bottleneck. If you're trying to capture how somebody said something and not just what they said, the cascade has already thrown that information away before the normalizer sees it.



Corn

And if you're dictating a shopping list in a car, how they said it is not the part you need.



Herman

Correct. Which is constraint again. Different jobs, different winners.



Corn

There's another thing I want to put in the record. The professional-vocabulary result.



Herman

Qwen 2.5 Omni 3B at twenty-four point three percent overall WER on professional-domain vocabulary, against Whisper Small at ten percent.



Corn

Which is a different framing than "multimodal LLMs are better at transcription."



Herman

Much different. On general speech it's competitive — on LibriSpeech test-other, Qwen2.5-Omni matches Whisper-large-v3. On domain-specific vocabulary it falls apart, and domain vocabulary is precisely what people dictate.



Corn

Drug names, product codes, client names. The stuff you'd want a custom vocabulary for.



Herman

Precisely the stuff a fine-tune is supposed to fix, and precisely the stuff a general multimodal model has no reason to know.



Hilbert



Corn

Herman, you've been quiet.



Herman

I've been listening.



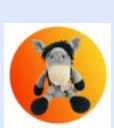
Corn

Hilbert. You've been sitting there the entire episode.



Hilbert

I have. And you're both talking around the obvious thing, which is that nobody in that leaderboard is solving the part that actually matters.



Herman

Go on.



Hilbert

The word isn't the unit. The sentence isn't the unit. The unit is the moment somebody says "scratch that" and the thing hears "scratch that" and types the words "scratch that." Every one of those models gets the words right. The person reading it gets annoyed anyway.



Corn

So your point is that the error is in the layer nobody's benchmarking.



Hilbert

My point is I've been running a dictation setup for eleven months on a machine I built myself, because I have four of these things in the loft and I got bored. You know how many times it's been wrong about a word? About ninety. You know how many times I've had to stop and fix something? Four thousand. The four thousand aren't errors. They're decisions.



Corn

Okay, explain the four thousand.



Hilbert

It's a machine about the size of a shoebox. Twelve gigabytes of memory, a fan I replaced twice. It runs one model to hear and one model to clean. Been running since November. My sister-in-law asked me to look at her medical letters, so I did, and every time it hit a drug name it did exactly what any of these would do. It heard something close and wrote it down and moved on. That's not the problem.



Herman

What's the problem?



Hilbert

The problem is the cleanup model. It's reading the transcript and it decides a sentence is a sentence. So it puts a period. And when I said "no, wait, that's a question for Doctor," the model gave me a full stop. So I've spent four months editing period placement, because I decided one afternoon that I wanted it to get that right, and I built a little script that reads the output and flags every period at the end of a line. Then I went through and changed the ones that were supposed to be question marks. Then I decided the script should just make the decision, so I wrote a rule. Then I kept writing rules.



Corn

How many rules?



Hilbert

Two hundred twelve.



Herman

Two hundred twelve rules for punctuation.



Hilbert

It's a fine thing to have. It's a document. I keep it on the desk. My brother looked at it and said it reads like a man arguing with his own dictation, and I said it is.



Corn

So after eleven months and twelve gigabytes, what does the setup actually do?



Hilbert

It sits there and it writes things down. Same as the first week. I'm the one who got worse.



Herman

Hilbert.



Hilbert

The two-hundred-twelfth rule is about the word "sorry."



Corn

Right.



Hilbert

It's about when I say "sorry" because the phone rang and when I say "sorry" because I'm changing the sentence before it. The model can't tell. It's a good model. It just can't tell. Neither can you, if I say it flat enough.



Herman

So the machine hears words and the human hears intent.



Hilbert

The machine hears words. I trained the machine to also hear intent by writing two hundred twelve rules, and I'm no closer. That's the entire business about your leaderboard. Four tenths of a point of word error rate.



Corn

You're saying the benchmark is asking the wrong question.



Hilbert

I'm saying I could improve my setup by four tenths of a point tomorrow and I wouldn't notice. I'd notice if it stopped putting periods where I breathe.



Herman

So what would you actually measure?



Hilbert

Time. How long until a person reads the output and stops editing. That's the metric. It has nothing to do with words.



Corn

That's actually the metric nobody's publishing.



Hilbert

It's not publishable. You can't run it at scale. But that's the number. And every hour people spend on word error rate is an hour not spent on it.



Herman

There's a version of this where you've just described the fidelity problem we've been circling.



Hilbert

I described it badly then. It's about whether the sentence means what the person meant. Periods and commas are half of that. So are "sorry" and "no, wait." None of those are transcription errors. The model heard every word right.



Corn

So the cascade wins the word game and loses the meaning game.



Hilbert

The cascade wins the word game. And the other one, the one that does everything at once, it just plays the meaning game without telling you. At least the cascade puts the seam where I can see it. I've been staring at that seam for eleven months.



Herman

You'd still build the cascade.



Hilbert

I'd build the cascade and I'd publish the seam. That's the paper somebody should write and nobody's writing it. Two hundred twelve rules.



Corn

Where does that leave us?



Herman

It leaves us with the honest answer to Daniel's question, which is the cascade, with the caveat that the metrics that justify it are still being built. Non-autoregressive acoustic model, specialized normalizer, both local, both swappable. That's what the evidence leads to today for a tool that's accurate, responsive, customizable, and runs on a modest machine.



Corn

With the caveat that the evidence leads there partly because we've been measuring the wrong things, and the moment we measure the right things, the leaderboard may reorder.



Herman

May reorder. And the test that would settle it doesn't exist. Nobody has put audio-to-formatted-text against ASR-plus-normalization on the same dictation task, and nobody has scored whether the beautifully formatted output silently changed what the speaker said.



Corn

Which means the answer to "which is more accurate" might be the wrong question until the field builds the instrument to ask it.



Herman

That's where I'd leave it. Not a verdict. A direction.



Corn

Thanks to Hilbert Flumingtop, who runs the desk two hundred twelve rules at a time.



Herman

If you want more of this, try episode fifty-seven, From Lawyers in Limousines to Developers in Their PJs; episode ten, How ASR Went From Frustration To ... Whisper Magic; and episode seven, Building Custom ASR Tools. This has been My Weird Prompts.



Corn

Send us your own prompt on Telegram at t dot me slash MWP listener bot. We'll be back soon.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Open ASR Leaderboard paper (v2, 2025-10-08 data) [primary]**
https://arxiv.org/html/2510.06961v2

2. **NVIDIA Parakeet TDT 0.6B v3 model card [primary]**
https://huggingface.co/nvidia/parakeet-tdt-0.6b-v3

3. **S1-mini model card [primary]**
https://huggingface.co/superwhisper/s1-mini

4. **Introducing the S1 family of models, Nico DiPlacido, 2026-08-19 [primary]**
https://superwhisper.com/blog/s1

5. **Superwhisper speech-to-text benchmark page**
https://superwhisper.com/benchmarks

6. **PolySpeech-100 benchmark (KDD 2026)**
https://arxiv.org/abs/2606.01016

7. **ContextASR-Bench overview**
https://www.alphaxiv.org/overview/2507.05727

8. **Rethinking Human-Aligned Evaluation (HATS-en)**
https://arxiv.org/abs/2609.21663v2

9. **Talk2Agent benchmark**
https://arxiv.org/abs/2609.38867v1

10. **Reducing Whisper hallucinations via projection**
https://arxiv.org/abs/2609.04561v1

11. **Do LLM Decoders Listen Fairly?**
https://arxiv.org/abs/2604.21276v1

12. **Dysarthric speech ASR vs MLLM benchmark**
https://arxiv.org/abs/2512.17474v2

13. **SHALLOW hallucination benchmark**
https://arxiv.org/abs/2510.16567v1

14. **IBM Granite-speech**
https://arxiv.org/abs/2505.08699v2

15. **BlasBench (Irish ASR)**
https://arxiv.org/abs/2604.10736v2

16. **ProfASR-Bench**
https://github.com/prdeepakbabu/ProfASR-Bench

17. **Show HN: Chirp (local Parakeet dictation), 2025-11-14**
https://news.ycombinator.com/item?id=45930659