Episode Book

EPISODE #6302

Why Fine-Tuning Whisper Won't Stop Duplicated Text

PUBLISHED	27 October 2026 at 17:36 UTC
DURATION	25:18
LISTEN	myweirdprompts.com/episode/whisper-decoding-search-parameters-34957/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
BOOK GENERATED	28 October 2026 at 09:53 UTC

SYNOPSIS

Speech recognition errors are often blamed on model weights when the real culprit is inference orchestration. Fine-tuning an acoustic model will not fix duplicated phrases caused by chunk merging, fallback temperature schedules, or silence heuristics. By examining Whisper's DecodingOptions, beam search trade-offs, and WhisperX's choice to disable conditioning on previous text, this discussion maps the operational line between neural logits and the search algorithms that actually decide what to print.

TRANSCRIPT



Corn

Daniel's back on speech recognition, with something his fine-tuning work has made hard to ignore. You can improve the weights and still get Whisper repeating a phrase three times, or two chunks both claiming the same words. He's asking where running the model ends and deciding what to print begins. Those jobs keep getting bundled together under inference, even though the second one has an alarming number of settings.



Herman

Including settings that decide whether yesterday's mistake becomes the next window's instructions. That is quite a privilege to give a transcript.



Corn

His practical question is what you can change without training anything. Search, repetition controls, chunk joining, the whole route from probabilities to readable text. I think the trap is seeing a bad sentence and assuming the sentence must live somewhere inside the weights.



Herman

Sometimes the probabilities are bad. Sometimes the machinery consuming perfectly usable probabilities makes a bad choice, preserves it, and feeds it back. Those need different interventions.



Corn

More training pairs won't teach two overlapping chunks to share custody.



Herman

Right at that boundary, you can improve the recognizer and still leave the duplication untouched. The joining procedure has no idea you spent the weekend fine-tuning.



Corn

Before you draw a clean boundary, Herman, inference servers usually call the entire operation inference. They're not all misusing the word.



Herman

No, that broader usage is normal. Inference can mean the complete deployed prediction process, including decoding. For this discussion, we're separating neural computation from the search and orchestration around it. There isn't a universally enforced terminology police force.



Corn

Finally, an unregulated department.



Herman

Whisper makes the separation concrete. Its encoder processes the audio. Its neural decoder produces logits for the next token, conditional on audio and the tokens already supplied. Then a search procedure chooses what happens next. That selected token goes back into another forward pass. The two jobs alternate rather than happening in two neatly separated time blocks.



Corn

Neural decoder and decoding algorithm. Someone named these to discourage visitors.



Herman

The neural decoder is part of the trained model. The decoding algorithm uses its scores. Whisper has a separate, frozen DecodingOptions object, consumed by DecodingTask. You can hand the model the same audio with different options and receive different text without changing its parameters.



Corn

Frozen means that particular options object is immutable. Not that the settings were delivered on stone tablets.



Herman

You construct another object when you want another attempt. Task and language live there. Search settings include temperature, beam size, patience, sample length, best of, and length penalty. Prompt and prefix affect supplied context. Token suppression and blank suppression constrain outputs. Timestamp controls govern whether timestamps appear and how early the first can be.



Corn

Half precision lives there too, although that's a computation choice rather than a rule for choosing words. The options container is a useful architectural clue, not a philosophical definition.



Herman

I'd make the same distinction for chunk stitching. That's usually orchestration around decoding, not token search itself. Daniel's useful operational boundary includes it because it changes the final transcript before a separate text cleanup stage ever sees it.



Corn

Punctuation cleanup edits an existing transcript. Search is still deciding which transcript exists.



Herman

What would convince you that a repeated phrase came from this surrounding machinery rather than just a poor acoustic prediction?



Corn

Keep the recording and checkpoint fixed. Change context handling, examine the window outputs separately, and see whether the repetition follows the search settings or appears only after merging. A single final text file has already hidden the evidence.



Herman

Whisper's fallback loop gives us an unusually visible example. It decodes a window, evaluates the result, and may decode that same window again. Its default temperature schedule starts at zero, increases in steps of zero point two, and reaches one.



Corn

A transcript gets six auditions.



Herman

Unless an earlier attempt passes. One test is compression ratio, with a default threshold of two point four. The code's comment is simply, too repetitive. It compares the text's size with its gzip-compressed size. Repeated material compresses unusually well.



Corn

Your transcript has been rejected for efficient storage.



Herman

It's a cheap heuristic with a useful intuition. A paragraph that keeps repeating the same sentence gives a compressor very little new information. Normal speech usually provides more. The test doesn't understand whether the speaker actually repeated themselves, though. It only sees the resulting text.



Corn

Nor does a repetitive transcript necessarily mean the model looped. A meeting can contain very little new information.



Herman

Average log probability supplies another test. The default threshold is minus one. Below that, the selected sequence was weakly supported by the model's own token probabilities, and the system can retry.



Corn

Its examiner is also its author.



Herman

With two different kinds of evidence. Compression examines the generated string. Average log probability summarizes model scores along the chosen sequence. Neither is an independent listener checking the recording.



Herman

Silence has its own interaction. The no speech threshold defaults to zero point six. High no speech probability together with low average log probability can mark a window as silence, allowing it to be skipped rather than repeatedly searched for words.



Corn

Hm.



Corn

That sounds reasonable until two tests disagree. One says the window is silence. Another says the text is an obvious loop. Which one wins?



Herman

That exact interaction has been argued over in Whisper's own repository. In the original code, the silence guard could clear the request to retry even when the compression check had already flagged repetition. A pull request proposed that silence should only win when the log probability test failed, never when the text was an obvious loop.



Corn

Silence had been granting diplomatic immunity to repetition.



Herman

The important detail is where that fix lives. No new checkpoint is needed to change that decision. The program was already detecting the repetitive text, then another branch was cancelling the response.



Corn

I can hear you preparing to call the whole model innocent.



Herman

Not innocent. Merely entitled to have the surrounding software investigated before sentencing.



Corn

Accepted. With supervision.



Herman

Fallback also changes more than temperature. At zero temperature, Whisper removes the best of sampling option. It can use beam search when a beam size is supplied. At positive temperatures, it removes beam size and patience and uses sampling instead.



Corn

You changed the recruitment process halfway through the auditions.



Herman

Beam search retains several promising partial sequences and compares their extensions. Sampling draws tokens from a temperature-adjusted probability distribution. Retrying at a higher temperature therefore changes both how probabilities are shaped and, in this implementation, how a sequence gets explored.



Corn

A user could say, I increased the beam and it still looped, without noticing the fallback attempt wasn't using that beam at all.



Herman

Yes. The effective settings of the successful attempt matter more than the settings you remember requesting. Corn, would you preserve previous text across windows by default?



Corn

For an interview with names recurring throughout, probably. If someone establishes a spelling early on, I don't want every window holding a fresh election. For a rough recording with long silences, I'd be much less eager to preserve whatever the machine invented.



Herman

That's the central exchange. Previous output can help continuity, but it can also carry errors forward. Whisper defaults to conditioning on previous text. Its own documentation says disabling that may make windows inconsistent, while making repetition loops and timestamp drift less likely.



Corn

The wrong phrase gets promoted from answer to context.



Herman

Then the next window is interpreted partly in light of that context. If the audio is weak, ambiguous, or silent, the erroneous text can gain influence. You don't just have one mistaken prediction. You have a prediction changing the conditions under which later predictions happen.



Corn

Training can make that first mistake less likely. It can't rewrite the branch that keeps passing the mistake along.



Herman

That's the narrower claim I'd defend. Fine-tuning can change repetition tendencies and silence behavior. We shouldn't say weights never matter. Better probabilities still leave the prompt policy, fallback rules, and chunk merger exactly as programmed.



Corn

You can hire a better witness and still have the clerk photocopy their statement twice.



Herman

So when you change the decoding parameters, you're not changing the model, you're changing how it searches. And that's the layer we're going to map out.
Ivan: Nightjar. Sixteen, forty two. Two thousand and one, twelve. The goose counts her eggs twice, but the third egg is a lie. Repeat sixteen, forty two. The gosling knows the pond is not a pond, only a puddle with a memory. Seventeen, ninety. The nest is warm because the lamp is on. Blackberry.



Herman

Message received.



Corn

What on earth was that?



Herman

Stock Whisper's search controls divide into exploration, scoring, constraints, and context. Beam size changes how many candidate paths survive. Patience changes how long beam search continues collecting completed candidates. Length penalty changes how completed sequences are ranked by length. Larger searches cost computation and don't guarantee a more faithful transcript.



Corn

You're going to turn the beam up until it needs its own chair.



Herman

Only if the chair improves recognition.



Corn

There it is.



Herman

You prefer a setting that never changes, even when it consistently gets your words wrong.



Corn

I prefer knowing which setting got them wrong.



Herman

Then you'd also want the selected temperature, retry reason, and window boundaries recorded. That makes comparisons much more informative than checkpoint names alone.



Corn

There is a separate practical distinction between helping the system recognize a word and forbidding it to write something. An initial prompt can establish vocabulary. Token suppression removes choices. Those are very different interventions, even when both happen to correct today's transcript.



Herman

Prefixes constrain the beginning of the current output, while prompts supply preceding context. Suppressing selected tokens can block unwanted special symbols or other outputs. Blank suppression affects the initial search. Language and task settings also matter enormously. Translation and transcription are different requested outputs from the same weights.



Corn

None of those changes should be confused with teaching the model a new accent. They steer the use of existing knowledge. Sometimes that's enough. Sometimes the acoustics remain wrong.



Herman

WhisperX makes a strong context choice in its ASR options construction. It sets conditioning on previous text to false. Its pipeline uses voice activity detection to identify speech regions and organize batched transcription, rather than simply carrying a continuous transcript through every window.



Corn

That trades some continuity for fewer opportunities to inherit a bad sentence. It also means punctuation and names can lose help from neighboring speech.



Herman

Context-aware batching moves back toward supplying context to recover punctuation and proper noun consistency. That tension hasn't disappeared. The aim is to preserve useful history without letting erroneous history dominate.



Corn

History with visiting hours.



Herman

WhisperX also reaches CTranslate2 through faster-whisper. Its batched generation exposes no repeat ngram size and repetition penalty, controls absent from stock OpenAI Whisper. The first prevents specified repeated token sequences. The second penalizes reuse during generation.



Corn

That can't repair a boundary duplication created afterward, though. Two individually clean chunks can both contain the same perfectly legitimate phrase.



Herman

Correct. Token search constraints operate within their supplied generation context. An overlap merger needs information about shared audio, timing, and sequence alignment. A string deduplicator that just deletes identical neighboring words can erase speech the person actually said.



Corn

Picture someone saying, I think... I think we should leave. Both chunks may correctly include the second I think because they share that audio. You want one copy of that acoustic event, while retaining the first hesitation. Text alone makes those two decisions look deceptively similar.



Herman

Nor should we assume WhisperX's voice activity batching proves a particular overlap deduplication algorithm. Its architecture and anti-repetition controls are clear. The exact text merge behavior needs checking in the implementation you're using. Batching doesn't automatically create overlap, either. Explicitly overlapping chunks do.



Corn

A no repetition setting is not a receipt proving you stitched correctly.



Corn

So where does the overlap actually get resolved, if not in the search?



Herman

Hugging Face's transformers pipeline is the clearest example, and it's exactly Daniel's case. Set a chunk length of, say, thirty seconds and a stride, and it cuts the audio into windows that overlap at the edges. Each window is decoded on its own. Then the pipeline lines up the token sequences from neighbouring windows, finds the longest run of tokens they agree on inside the overlap, and keeps one copy. That merge never sees a probability. It's sequence alignment on finished output.



Corn

And stock Whisper doesn't overlap at all.



Herman

Its sequential long-form mode slides forward instead. It reads the timestamp tokens the model predicted and jumps the next window to where the last complete segment ended. So the boundary is set by the model's own timestamp guess. If that guess drifts, you get either a gap or a repeat, and changing the chunk length, the stride or the long-form mode changes the transcript without touching a single weight.



Corn

Two duplicate-word bugs, two different machines responsible.



Herman

Implementation differences matter even under apparently identical option names. Faster-whisper's release notes include fixes to how its batched and sequential modes handle token suppression and silence. Same checkpoint, revised surrounding code, different possible output.



Corn

Your model hasn't changed. Your transcript has received a software update.



Herman

NeMo gives this separation an explicit runtime interface. You can change decoding strategy on the ASR model. Depending on the architecture and supported configuration, you swap greedy decoding for beam search, add language model scoring, or use vocabulary boosting.



Corn

Whisper isn't the template for every recognizer here. CTC handles alignments differently, and transducer families have their own search behavior. You can't copy a Whisper temperature setting into a CTC system and expect it to mean anything.



Herman

NeMo spans CTC, RNN-T, token duration transducers, attention encoder-decoders, and hybrid models. Some hybrids let you choose between their transducer and CTC heads at inference time. The settings belong to a particular architecture and implementation, not to speech recognition in the abstract.



Corn

Suppose I need a rare surname that keeps coming up in my meetings. Would you boost it or add a language model?



Herman

I'd consider the error. A known, rare name may benefit from boosting. Broader linguistic ambiguity may benefit from language model scoring. Shallow fusion combines the acoustic score with a weighted language model score and a sequence length term. Beam alpha controls language model influence. Beam beta controls the length contribution.



Corn

More influence buys familiarity, including familiar mistakes.



Herman

Yes. Excessive vocabulary bias can insert a favored word where the audio doesn't support it. Neural rescoring can instead rerank a shortlist of candidates. The beam width determines what reaches that shortlist, so a rescoring model cannot rescue a correct answer that search already discarded.



Corn

It's difficult to win an election after being removed from the ballot.



Herman

What's your objection to a strict repetition ban, apart from meetings becoming suspiciously concise?



Corn

Speech repeats on purpose. No, no, no has different force from no. Rehearsal, stuttering, emphasis, dictation, language learning... a transcript needs to preserve those when they're audible. I don't want a recognizer silently appointing itself my editor.



Herman

Which is why a no repeat ngram setting needs a test set with deliberate repetition in it. Otherwise you only measure the loops it removed, never the real speech it deleted.



Corn

A very confident delete key with no audit log.



Herman

And to be fair to training, some of this belongs there. The Calm-Whisper work, from Interspeech twenty twenty-five, found that three of Whisper's twenty decoder attention heads accounted for most of its hallucinations on non-speech audio. Fine-tuning just those heads on noise paired with empty transcripts cut those hallucinations by over eighty percent, with under a tenth of a percent worse word error rate on clean speech. That is a weights problem, fixed in the weights. The useful separation is intervention by intervention. Fix the search rule when the search rule is wrong. Train when you need to change the model's response to the audio.



Corn

Have we got a clean example of improving output through decoding alone?



Herman

Whisper-CD uses training-free contrastive decoding. It compares scores from clean audio against scores from altered inputs, including noise, silence, and temporal shifts. The aim is to favor words supported specifically by the real audio over words the decoder would produce regardless.



Herman

In its reported CORAAL evaluation, word error rate improved by up to twenty-four point three percentage points. Token generation was forty-eight percent faster than beam search. Those are comparisons within that evaluation, not a universal speed promise.



Corn

Useful evidence that the search deserves its own engineering budget.



Herman

It changes how the scores are used without updating weights. That demonstrates the leverage, though it doesn't establish that every decoding intervention beats every fine-tune.



Corn

I'd also separate recognition benchmarks from whether I can trust a verbatim transcript. If a person repeats an instruction deliberately, removing the duplicate can change what happened in the room.



Herman

Would you keep even a formal repeated instruction, Corn, if the second copy added nothing?



Hilbert

...both copies, Herman. I say, The chair must remain seated. The chair must remain seated. That's the rule at the Emergency Adjournment Society. I say it at every meeting, before the chairman is winched up to the ceiling. He says it improves his view of abstentions. I object in the same words twice. I have done that for years.



Corn

An objection with a load-bearing clause.



Hilbert

The minutes currently say it once. Martin Bell runs the transcription. He's our assistant secretary. He can hold a procedural silence for four minutes and seventeen seconds. I admire his patience. I leave his tea within reach because he will not break a silence to ask. This month the transcript printed, The biscuit account is closed, three times. Nobody



Hilbert (continued)

said it. Martin stopped feeding the previous window back in. That fixed the loop. Then he applied his repeat ban to everything.



Herman

Including an instruction the recording supported.



Hilbert

Including mine. I use a foot switch because the little record buttons don't suit anteater claws. I record the objection again after each meeting. Martin's automatic correction removes the second copy again. Herman, you distinguished prompts from prefixes earlier. I can get the distinction into our minutes. I cannot get the second sentence in. The software advertises verbatim transcription. I consider that ambitious.



Corn

Put the two deliveries under separate agenda items. Then he's deleting a whole item rather than tidying a sentence.



Hilbert

They're already separate items. Standing objection and renewed standing objection. The dispute occupies six pages of the rules. Martin keeps both headings. He has also applied the same setting to the attendance roll. There are two members called Peter Ward. Only the first receives a biscuit.



Herman

Two speaker labels would at least make them different records. The biscuit allocation could use the member identifier rather than the text of the name.



Hilbert

The biscuits are counted by hand. The missing name is the problem.



Corn

Separate records do matter, though. Once the transcript becomes minutes or an attendance list, deleting text starts changing another system's inputs. The recognizer's neat output can conceal the fact that something was removed.



Herman

Especially if only the final cleaned transcript survives. An intermediate window transcript, timestamps, or a record of suppression decisions can distinguish a repeated acoustic event from repeated generated text. Without that, you're reconstructing the error from its least informative artifact.



Corn

That doesn't mean every user should inspect token scores. It means the application shouldn't pretend every printed word, or every absence, came straight from the microphone.



Herman

There is still a measurement I'd like to see isolated more cleanly. How much repetition disappears from disabling previous text alone, compared with changing the fallback schedule? The mechanism is explicit, but I don't know a controlled number separating those contributions.



Corn

The common wrong belief is that a better checkpoint necessarily fixes the transcript you already dislike.



Herman

Better weights can improve the probabilities. They don't repair an overlap merger, change the retry policy, or restore a legitimate repetition that the application forbids. Those are separate decisions, even when the interface gives you just one transcribe button.



Corn

What I'll be watching is whether context-aware systems learn to retain names and punctuation without erasing a second instruction simply because the first one looked sufficient.



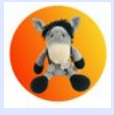
Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If you want more of this, try episode three, Safetensors or something else; and episode eight, Building Your Own Whisper. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoyed it, leave us a review wherever you listen. It helps people find the show.



Corn

See you tomorrow.