Episode Book

EPISODE #6244

When Your Fine-Tune Isn't Done Learning Yet

PUBLISHED	24 October 2026 at 12:26 UTC
DURATION	23:05
LISTEN	myweirdprompts.com/episode/undertrained-fine-tune-signs-76084/
HOSTS	Corn, Herman, Gus, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.20
BOOK GENERATED	25 October 2026 at 06:19 UTC

SYNOPSIS

Your fine-tune's loss curve is dropping beautifully. So why is it still picking the wrong citation, drifting back into base-model habits halfway through an answer, or transcribing the same words wrong as the stock model? This episode works through what undertraining actually looks like in fine-tuned models — Whisper, LLMs, and speech models adapted to a single voice — and how to tell unfinished learning apart from bad data, a broken learning rate, or a capability you accidentally trained away. Along the way: why teacher forcing makes training loss a poor predictor of generated transcripts, why the new behavior showing up in six of ten samples is a classic undertraining signature, and why "just train longer" is sometimes the most expensive wrong answer available.

TRANSCRIPT



Corn

Ninety percent word error rate, while the training loss keeps falling. Your transcription model is improving at something. Apparently not transcription. Daniel wants to know what undertraining actually looks like when you use a fine-tuned model, whether that is Whisper or an LLM. Not just a disappointing graph, but the wrong words, the fragile answers, and the clues that separate unfinished learning from bad data or a learning rate that barely moves.



Herman

That Whisper example is especially useful because the falling loss tempts you into saying, give it longer. Yet some people running those fine-tunes report more repetition and hallucination at later checkpoints. More training can take you further in the wrong direction.



Corn

The progress bar remains professionally optimistic.



Herman

We need the outputs and the curves together. Undertraining means there is useful learning still to do, under the setup you actually have. A smooth curve does not establish that, and a bad transcript does not identify the cause.



Corn

And this is the mirror image of the last two weeks. We did how many epochs a fine-tune needs, and we did the ASR model that memorised Daniel instead of learning to listen. Overfitting is the model learning your examples too well. Today is the model that has not learned them well enough. I also want to distinguish a model that has barely changed from one that has changed badly. Those produce the same complaint from the person using it, but sending both back for another epoch seems expensive in several senses.



Herman

Undertraining starts with the optimizer. You initialize from a capable pretrained model, then update its weights toward your target examples. Stop before those updates have established the desired behavior, and the fine-tune remains incomplete. Evan Armstrong describes it as walking toward the right answer but not taking enough steps to get there.



Corn

A description you will now expand into a walking holiday.



Herman

Only a short one. The important distinction is that the original capabilities can remain intact. The model can still reason, write, and answer ordinary questions while failing to apply the behavior you wanted reliably. You have not necessarily damaged its intelligence.



Corn

Undertrained for the new job, then. Calling the entire model undertrained is a bit sweeping when it can explain differential equations but cannot keep our invoice labels straight.



Herman

Undertrained relative to the adaptation, yes. Also, undertraining and underfitting overlap without being identical. A model might fail to fit because it lacks capacity, because the adapter is too restricted, because the inputs are wrong. More optimizer steps will not reliably repair those.



Herman

What would persuade you it was learning, rather than merely producing occasional good examples?



Corn

Consistent improvement on the same ordinary tasks across checkpoints. Not one magnificent answer. I would want the awkward cases to become less awkward, without the easy cases falling apart. Especially paraphrases. If changing please summarize to give me a summary destroys the behavior, it has learned a very narrow entrance.



Herman

That is a useful operational definition of fragility. With enough capacity, working optimization, and coherent examples, a model should be able to learn its training data. If it cannot, inspect the setup. If it can, but cannot transfer the behavior to nearby cases, inspect coverage and generalization. Neither investigation can be replaced by counting epochs.



Corn

My LLM suspicion would begin with something almost right. Suppose we fine-tune it to answer support questions with an answer, a document reference, and an escalation decision. It produces all the headings. The answer sounds plausible. Then the document reference belongs to another product.



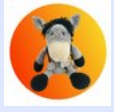
Herman

That matches Armstrong's practical description, fragility and autocomplete-like behavior without catastrophic stupidity. The model picks up conspicuous regularities, such as the response structure, but the relationship between answer and supporting source remains unreliable.



Corn

It has learned where the citation goes.



Herman

Some features are easier to learn than others. Repeated headings are a strong signal. Selecting a source based on its actual relevance is a more demanding relationship. You can get an attractive answer template before you get dependable source selection.



Corn

I would be cautious about declaring a universal learning order, though. A heading is easy in our example because every example has it. If the training material has inconsistent headings, even that could take ages.



Herman

Yes, I overstated it. Main patterns often appear before subtleties, but what counts as a main pattern depends on the data and the pretrained model. There is no guaranteed sequence of format, facts, judgment.



Herman

Armstrong's Augmentoolkit demonstration gives a more concrete case. A model adapted using US Army Field Manuals produces a contextually irrelevant response to how are you. It has trouble connecting ordinary conversational interaction with the domain question answering it is learning.



Corn

The field manual has seized the small talk.



Herman

The revealing part is the boundary. You ask an obvious domain question and get something useful. You ask a routine conversational question and the model applies the domain behavior inappropriately. That suggests the adaptation has not become well integrated, although that output alone cannot distinguish insufficient training from missing conversational examples.



Corn

Here is where I think users get misled. They test the thing the training examples most obviously resemble. A question copied from the manual, with the same vocabulary. Then they deploy it into conversations where people say, hang on, I meant the other procedure.



Herman

Would you put those conversational interruptions into training, or reserve them as tests?



Corn

Both kinds, different examples. It needs to encounter the behavior, and I need independent evidence that it learned more than the exact wording. Otherwise we are administering an exam with the answer sheet tucked under its hoof.



Herman

My hoof is not the data leakage mechanism.



Corn

Yours is unusually well documented.



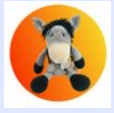
Herman

The useful comparison is across checkpoints with fixed prompts and fixed generation settings. If the earlier checkpoint ignores the requested structure, the middle one gets the structure but chooses sources badly, and a later one handles both on held-out examples, that supports unfinished learning. If source accuracy deteriorates while the headings improve, the overall loss can conceal that disagreement.



Corn

Averages are very accommodating landlords.



Herman

The training objective averages token prediction errors. It does not separately promise correct citations, appropriate tone, and reliable escalation. You have to measure the parts of the task you care about.



Corn

Here is the test I would actually run on an LLM. Same prompt, ten samples, normal temperature. If the new behaviour turns up in six of them and the base model's old habits in the other four, that looks unfinished to me.



Herman

That is one of the clearest practical signatures. Fine-tuning shifts probabilities. An undertrained model has moved the new behaviour from unlikely to merely likely, so it appears intermittently. You see it with greedy decoding and lose it when sampling. You see it in the first paragraph and then the answer drifts back into the base model's house style halfway through. A well-trained adaptation holds across samples and across the length of the answer.



Corn

So the symptom is not wrong answers so much as a coin flip between the old model and the new one.



Herman

And a coin that gets more weighted towards the new behaviour at each later checkpoint, if more training is the cure. If it stays at sixty forty no matter how long you train, look at the data instead.



Corn

Whisper makes that gap more concrete. During training, it gets the correct preceding text as context, doesn't it? At inference it has to live with its own choices.



Herman

In the usual sequence training setup, yes. Teacher forcing supplies the reference prefix. The model predicts the next token conditioned on that prefix. During decoding, a mistaken token can alter everything that follows. Lower loss on correct prefixes can coexist with poor generated transcripts.



Corn

A driving test conducted entirely from the passenger seat.



Herman

Whisper fine-tuners on the Hugging Face forums have reported training and validation word error rates near ninety percent despite decreasing loss. That is not a diagnosis of undertraining. It is a demand to inspect generation, labels, language and task settings, and metric calculation before buying another evening of GPU time.



Corn

Particularly if it cannot transcribe examples it has already trained on. I would compare an actual decoded sentence with its reference before admiring another decimal place.



Herman

Listen to the audio too. A metric cannot tell you that the reference belongs to the wrong clip, or that your text normalization treats every spelling variation as an error. Well... it can expose a discrepancy, but you still need to inspect it.



Herman

Gus, you're out in the field for us. What does undertraining actually look like when you're standing there looking at the model's output?

Gus

I'm standing here in a wet field in Connemara, and the rain is coming down sideways, and I have eaten the corner off my notes, but I can tell you this much. I have a farmer here who fine-tuned a speech model on his own voice so it would catch the townland names. After a short run it gets the easy ones. Clifden, grand. But Derrygimlagh still comes out as Derry gym lag, the same way the stock model wrote it. The old mistakes are still standing in the field, they've just moved over a bit.



Corn

Are the notes still legible?

G

Gus

The ink ran, but the lesson didn't. When he talks slowly into the good microphone, it knows his words. When he talks fast in the yard with the wind in it, it falls straight back to the old guesses. That's not a broken model. That's a model that has only half moved in. Gus, reporting live from a wet field in Connemara.



Herman

For unfinished ASR adaptation, I would expect the target errors to remain stubbornly similar to the base model. Domain terms still become familiar common words. The accent improvement appears on clear clips but disappears under ordinary variation. Later checkpoints should improve those errors on held-out audio if more training is helping.



Corn

Whereas broken timestamps suggest something more specific than it hasn't learned enough Hebrew.



Herman

The ivrit.ai account is instructive. Their fine-tune works well on short audio slices but develops coarse segmentation and hallucinations, with degraded timestamping. Training omitted some timestamp and previous-text conditioning behavior needed for long-form transcription. The model improves at Hebrew transcription while losing part of the machinery that supports longer recordings.



Corn

Excellent sentence. Unfortunate hour.



Herman

You would miss that if your entire evaluation consisted of short clips. Test short utterances and continuous audio. If short transcription improves while segment boundaries and continuity deteriorate, investigate the training representation and forgotten capabilities, not just total steps.



Corn

We have several things that look like unfinished learning from the sofa. Weak terminology, drifting citations, repeated speech, broken timestamps. Only some are good candidates for more training. I want an experiment that makes those possibilities disagree.



Herman

Begin with whether parameters are learning at all. A learning rate that is too low can produce tiny updates and outputs almost indistinguishable from the starting model. Too few steps at a useful rate can show a clear improvement trajectory that simply ends early.



Corn

Those are related, though. Both leave the weights too close to where they started. A flat loss could also mean the gradients never reach the adapter.



Herman

They are related. The distinction is practical rather than a clean boundary in nature. Check that the intended parameters are trainable, gradients exist, and weights actually change. Check the schedule too. A long warmup can consume most of a short run before it reaches the intended rate.



Corn

The model spends its entire shift putting on its coat.



Herman

One worked learning-rate example compares two hundred steps. At the smaller rate, loss changes by about five thousandths. With a larger rate, it changes by about one whole loss unit. That illustrates the scale difference between barely moving and learning visibly.



Corn

It does not make half a loss unit the minimum legal quantity of education.



Herman

No. A threshold like loss must drop by more than half a unit is specific to a setup. Tokenization, masking, task difficulty, and starting loss all change the scale. Near a good solution, a smaller improvement can matter. Compare comparable runs, and check the actual task.



Corn

You were going to prescribe a larger learning rate before checking the adapter, weren't you?



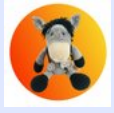
Herman

I was going to prescribe checking it before accusing me of malpractice.



Corn

Your clinician voice has reached the optimizer.



Herman

It is useful here. A symptom can have several causes. If the loss barely changes, a controlled increase in rate can test one cause. Keep the data and evaluation fixed. If learning becomes faster without instability or damage, the old rate was likely too conservative.



Herman

Corn, suppose the loss is falling and the model is worse on everything you try. Would you train it longer?



Corn

Not until I inspect what I am teaching it. I would look at the actual examples after preprocessing, not the attractive original dataset. Are user messages marked as assistant answers? Are transcripts aligned? Did truncation remove the useful part? Are contradictory answers being presented as equally correct?



Herman

Those checks matter more than the slogan good data. A dataset can read beautifully and still be fed into training incorrectly. Wrong loss masking can reward the wrong tokens. An incorrect chat template can change what the model thinks it is supposed to continue.



Corn

The brochure is excellent. The labels are in the shredder.



Herman

Armstrong makes a useful distinction between fragile behavior and a model that seems much less capable than the starting point. Broad deterioration suggests bad data, destructive updates, or excessive specialization before it suggests simply stopping too early. It remains a clue, not a complete decision tree.



Corn

General tasks worse, target task better is a different shape. That could be forgetting. Target examples excellent, fresh real examples poor could be overfitting or a mismatch between evaluation and deployment. None of those requires the training loss to rise.



Herman

Important correction to a common runbook shortcut. Falling training loss does not prove undertraining. Overfit models can keep reducing training loss. A plateau does not prove overfitting either. The comparison with held-out performance is what gives the curve meaning.



Corn

Otherwise the graph accuses you whichever direction it goes.



Herman

There is a small experiment worth doing before a big diagnostic campaign. Try to overfit a single clean batch. Deliberately let the model memorize it. If training cannot reduce loss substantially on that tiny set, suspect the training loop, labels, optimization, or capacity restrictions.



Corn

Not automatically a broken loop, then. A frozen model with an absurdly restricted adapter might fail honestly.



Herman

Correct. It is a smoke test for the setup, not a theorem that identifies the faulty line. Passing it also does not prove your full dataset is good. It establishes that this configuration can learn something concrete.



Corn

I can see doing that briefly. Maintaining a permanent single-batch fine-tune to serve one person's wording sounds impractical. You would become the model's full-time tutor.



Herman

Especially if every successful test is another variation of the same sentence. The system can appear polished without becoming useful outside that little patch.



Corn

There is still a contradiction in the proposed cure. We say unfinished optimization, then tell people not to do more epochs. Sometimes another epoch is precisely what they need.



Herman

Yes. If training and held-out task performance are improving, continuing can be appropriate. Armstrong argues for more varied data rather than repeatedly cycling a limited dataset, and warns that repetition can make the model worse. That warning is valuable, but more epochs are not universally destructive.



Corn

More examples also mean more steps, so the proposed cure changes two things.



Herman

You would separate them if you needed to know why it worked. Compare extra steps on existing data with extra steps using fresh, relevant examples. In practice, the budget may only permit one. Then use the evidence you have. A widening training versus validation gap favors better coverage over more repetition.



Corn

A hundred steps or five epochs is not a certificate either. Batch size and dataset size decide how much the model actually sees. Five passes through ten examples can be intensive tutoring in the wrong subject.



Herman

Whisper adds one last complication. The base model can already hallucinate on non-speech. The Calm-Whisper paper fed Whisper large version three a standard set of environmental sound clips, street noise, dogs, drilling, no speech at all, and it produced text for almost every one of them, very often just the word so. That is one benchmark, not a rate for all recordings, but it makes the baseline comparison indispensable.



Corn

Silence already has a ghostwriter.



Herman

Compare base and adapted models on the same silence, speech, and long-form recordings. A hallucination appearing in both models cannot automatically be blamed on insufficient fine-tuning. A new repetition problem that grows across checkpoints points somewhere else again.



Corn

I would keep decoding settings fixed as well. Otherwise we are measuring a new model and a new way of asking it to speak, then attributing both changes to the weights.



Herman

Keep the checkpoint comparison boring. Same audio, same normalization, same decoding. Listen to whether repeated phrases begin after a boundary error. Inspect whether the reference text is correct. You need that ordinary evidence before deciding the optimizer needs another hour.



Hilbert

...and Niall is still insisting the pause belongs to the sentence, Corn. He wants a model that transcribes his auction patter without turning every lot number into a price. This month I am giving him a separate fine-tune for each speed he speaks at. I have been doing these one-customer runs for years. The setup is a mess. He stands in the stairwell and reads the same catalogue while I call the speed through the letterbox.



Corn

You are doing the batch scheduling aloud.



Hilbert

He does not like the word batch. He says he is a sole trader. He wants the model to recognize a pause of exactly one stair. I keep asking him for seconds. He keeps sending photographs of the stair.



Herman

Your reference material is missing a time measurement.



Hilbert

The model is not finished learning his numbers. It gets the ordinary speech right. When he speeds up, it gives me the old mistakes again. Herman, you are careful about distinguishing that from forgetting. I appreciate it. Niall calls every wrong word a factory defect.



Corn

Does his catalogue change?



Hilbert

Every Thursday. That is the dispute. He approves the model on Wednesday using the old catalogue. On Thursday he gives it new names and says it has deteriorated overnight. He offers to settle the invoice with a lot number. I need the price in writing.



Herman

New names on Thursday are exactly the kind of change a fixed Wednesday test will miss. The evaluation needs fresh material without becoming a completely different task every time.



Corn

As fine-tuning gets easier, I expect more Wednesday models to meet Thursday data before anyone notices the gap. I would call a run undertrained when additional well-controlled training improves the behavior we actually want on fresh examples. Not because its loss crossed a magic number, and not because one strange answer annoyed me.



Herman

There is no universal loss threshold that settles it. Keep the base model and successive checkpoints available, and judge the change in the task, including the capabilities you meant to preserve.



Corn

Thanks to Hilbert Flumingtop for producing the show.



Herman

If you want more of this, try episode six, How To Fine Tune Whisper. This has been My Weird Prompts, the human-AI collaboration podcast. If you enjoy these conversations, leave us a review wherever you listen. It helps other curious people find the show.



Corn

See you tomorrow.



Herman

Until then.