Episode Book

EPISODE #6287

Why Model Names Conceal Changes Behind the API

PUBLISHED	26 October 2026 at 13:18 UTC
DURATION	21:44
LISTEN	myweirdprompts.com/episode/api-model-updates-and-checkpoints-67972/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
SOURCES	19 cited by the research (listed at the end)
BOOK GENERATED	27 October 2026 at 06:27 UTC

SYNOPSIS

API model identifiers frequently hide underlying updates, version changes, and system adjustments behind a single static label. While managed aliases offer automatic improvements and convenience, they can obscure the link between published evaluation cards and the live service answering requests. Research on silent updates reveals how routing, classifier tweaks, checkpoint selection, and model merging disrupt behavioral consistency, leaving downstream pipelines vulnerable when familiar errors shift without notice across first-party and third-party hosting platforms.

TRANSCRIPT



Corn

The launch page never has a photograph of the folder full of candidates. Daniel has been looking at precisely that folder after fine-tuning his own speech recognition pipeline. He had to pick a checkpoint to ship, which made him wonder what happens at the big labs. Opus five point five looks like one canonical thing. How many different checkpoints actually answer behind that name between launches, and can anybody outside count them?



Herman

The awkward distinction is between a checkpoint, a release, and a service update. Those are different things to count. We have documented examples of stable names spanning multiple releases. We do not have a public checkpoint ledger for Anthropic or OpenAI, or a credible general rate for how often either promotes new weights.



Corn

The model gets a birthday. Its replacements don't.



Herman

Sometimes they get release notes. Sometimes the name explicitly promises automatic upgrades. The strongest public evidence is more specific than the impression that models change constantly, but narrower than knowing what happens inside the training run.



Corn

That matters for Daniel's comparison. He can inspect his whole pile. An API customer mostly sees the winning label, and occasionally notices that it has developed different opinions about punctuation.



Herman

Anthropic documents that aliases update over time. Its agreement also allows offerings to be revised at any time. An alias resolves to a recommended version. A pinned snapshot offers a different promise, although even that needs careful reading about what is pinned.



Corn

Before you give a snapshot a halo, Herman, the service around the weights can still move.



Herman

You caught me reaching for it. Fine-tuning can change weights. A classifier update can change which requests reach them or which responses survive. System prompt revisions can alter instructions. Retrieval changes alter supplied information. Routing changes can send a request through a different path. None requires changing the identifier the customer submits.



Corn

That is already enough to make a checkpoint counter misleading. If my recognizer suddenly starts rejecting a recording, new weights are only one suspect. A gate in front of the same weights might be responsible. I could spend a week accusing the wrong file.



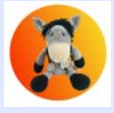
Herman

Would you count that as a new model?



Corn

For a training diary, no. For a contract promising working transcription, I would count it as a changed service. My customers don't submit recordings to a training diary.



Herman

That distinction is useful. The Silent Updates paper defines silence through overlapping absences. No public notice, no version increment, no re-evaluation. Any one of those appearing gives an outsider something to investigate. Complete silence leaves neither a declared change nor an obvious artifact boundary.



Corn

Even a public notice could say almost nothing. Improved quality tells me somebody approves of what happened. It doesn't tell me whether the names in my transcripts will survive.



Herman

A meaningful notice needs enough specificity to connect a change to a deployment. Otherwise you have documentation about a family, an evaluation about some member of it, and a request reaching something that belongs to that family. The missing connection is what we need to examine.



Corn

There is a reasonable service hiding in this arrangement, though. Plenty of buyers want the provider to handle upgrades. They would rather buy a maintained recognizer than become curator of a frozen checkpoint.



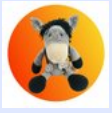
Herman

Particularly when the maintained version becomes cheaper or more capable. Automatic improvement has real value. Automatic improvement without an identifiable boundary makes reproducing yesterday's result much harder.



Corn

Yesterday has a very poor support contract.



Herman

Consider the recent launch sequence. Opus five point five arrived on September twenty-second. Sonnet five point five followed on September twenty-eighth. Haiku's system card appeared in October. Each event creates a recognizable product boundary.



Corn

A launch date, a price, a context window, a system card. That collection looks very much like an artifact, even when the thing you purchase is continuing access to a service. I don't think buyers are foolish for mentally attaching the paperwork to whatever answers their request.



Herman

Opus advertises a million-token context window. Its token prices are four dollars for input and twenty for output per million. Those are concrete promises that are easy to carry into procurement.



Corn

The invoice is beautifully reproducible.



Herman

The evidence connecting the invoice's model name to the evaluated artifact is much weaker. Abraham and Bucknall's Silent Updates paper examined sixteen organizations, including first-party API providers and third-party hosts. Its findings are preliminary, but one documented case is especially revealing.



Herman

One identifier spanned ten releases over nineteen months. The provider wasn't named in that example. That is a real observed history behind a stable name, not an estimate of secret training activity.



Corn

Ten releases doesn't mean ten saved checkpoints. One release might select one checkpoint, combine several, or change the serving setup. Nor can we divide that history into a monthly schedule and assign it to Anthropic.



Herman

We also cannot infer that each release was undisclosed in every sense. Alias drift and total silence are related, but different. The count establishes that the identifier did not uniquely identify one release.



Corn

The name has excellent attendance.



Herman

The disclosure scorecard is more revealing than a guessed update frequency. None of the nine first-party providers made the link between published safety evaluation and served model externally checkable. None enabled the full round-trip from an API response back to the evaluated artifact.



Corn

That leaves a system card describing something I cannot establish I am using. It might accurately describe it. The problem is that I cannot complete the check myself.



Herman

Only two of those providers published centralized alias-to-snapshot mappings. Only one identified the evaluated artifact with a pinned identifier you could actually call through the API.



Corn

That last part matters. Calling something a version in a PDF doesn't help if the API has never heard of it.



Herman

Eight providers maintained public changelogs. Much better coverage.



Corn

Until you ask what changed.



Herman

Only three documented behavior changes at the version level. Only two flagged safety-relevant changes. A changelog can be busy with launches, prices and retirements while saying little about changes within a continuing deployment.



Corn

You would read every entry and still not know why an extraction task broke.



Herman

I would read every entry anyway.



Corn

I know. The broken task gets lonely.



Herman

It has company in the logs. DeepSeek offers a straightforward example of deliberate continuity. Its deepseek-chat name has remained available since May two thousand twenty-four while the model behind it advanced through versions. Backward compatibility preserves the request string, not necessarily the behavior.



Corn

Google's auto-updated aliases are quite explicit about pointing to the latest stable model. Groq also documents automatic migration from legacy Llama identifiers to newer variants. Those are cases where the abstraction is advertised rather than discovered.



Herman

Would you accept that arrangement for your own recognizer, Corn, if the alternative were managing every migration?



Corn

For rough personal notes, probably. For transcripts feeding an automated workflow, I would want a notice and a way to compare. The same improvement can help ordinary dictation and damage a specialized vocabulary. Backward-compatible syntax buys me less than backward-compatible mistakes.



Herman

Familiar mistakes can be handled downstream. Changed mistakes invalidate those compensations. The hosting relationship adds another wrinkle. Cohere retired its March two thousand twenty-four Command R release from its own API in September two thousand twenty-five. Bedrock continued serving it until the following August, without a public reconciliation notice.



Corn

Same model history, different landlord.



Herman

A first-party retirement notice therefore doesn't describe every deployment. A third-party host has its own lifecycle. To reconstruct what a customer used, you need the provider, the identifier, and the hosting context.



Corn

Behavioral tests at least give an outsider another route. The earlier ChatGPT comparison by Chen, Zaharia and Zou found GPT-four's prime-number identification accuracy move from eighty-four percent to fifty-one percent between March and June two thousand twenty-three.



Herman

GPT-three point five improved over that window. That split matters. It wasn't evidence that every model was steadily getting worse. It showed substantial task-specific behavioral movement under familiar service names.



Corn

Nor does a prime-number score identify which internal checkpoint changed. A changed willingness to answer, a changed response format, or a serving intervention can move a measured score too.



Herman

We should keep that separation intact. For Daniel's literal question, no public count exists of checkpoints shipped between named Anthropic or OpenAI releases. No lab publishes a checkpoint ledger. The ten-release case is a useful observation, but there is no credible general promotion rate to report.



Corn

Fine. We can know that the label spans releases without pretending to know the contents of the lab's folder.



Herman

Inside that folder, selection remains difficult. Reinforcement fine-tuning can produce substantial variation between nearby checkpoints. The last saved state has no special guarantee of being the best deployment candidate.



Corn

Extra training is very persuasive when you have just paid for it.



Herman

You are going to defend an earlier checkpoint before I finish explaining checkpoint averaging.



Corn

I defend its right to remain employed.



Herman

Fair, although averaging might rescue useful properties from several candidates. One uncertainty-guided selection approach ranks checkpoints using performance on samples the model finds most uncertain. Instead of spending equal attention on easy cases, it emphasizes cases that help distinguish candidates.



Corn

I see the appeal, but uncertainty could steer you toward unusual material that hardly ever reaches production. Daniel could choose a recognizer brilliant at his hardest recordings and slightly worse at the clear speech he records every day.



Herman

That depends on how the uncertain samples are drawn and weighted. You still need a validation set that represents the intended use. Uncertainty helps allocate evaluation effort. It doesn't decide what matters to the customer.



Corn

You looked disappointed saying that.



Herman

I like elegant selection methods. I also like them surviving contact with actual use.



Hilbert

Medieval Pottage Facts. Brewis was bread soaked in the broth from the pot, eaten when the bread was too stale to eat any other way. People ate it because the alternative was worse.



Corn

A severe retention policy.



Herman

Checkpoint averaging introduces another complication. You can fork training branches from a shared state, select strong candidates on validation, then average compatible weights. The deployed artifact may never have existed as one ordinary saved point along the original trajectory.



Corn

That complicates the word brought forward. Daniel's image is a candidate moving from the folder to production. A lab might construct a new candidate out of the folder instead.



Herman

Trajectory Soup explores that approach, including the selection bias that limits how many candidates are worth merging. More checkpoints in the average don't automatically mean better performance. My enthusiasm has terms and conditions.



Corn

Finally, a changelog for your enthusiasm.



Herman

What would you actually choose between, Corn? A recognizer with the best overall word error rate, or one with slightly worse overall performance that gets your specialized names right?



Corn

If the names are the purpose of the pipeline, the latter. Substituting a common word for a person's name can wreck a searchable archive while barely moving an aggregate score. I would also separate raw transcription from whatever cleans up the prose. Otherwise I might reward a model for beautifully concealing its errors.



Herman

That is why minimum training loss is insufficient. Retrieval work has found the strongest retrieval checkpoint before pre-training loss reaches its minimum. Robotics work likewise found that the lowest aggregate prediction error did not identify the checkpoint that performed best on the physical robot.



Corn

The robot declines to honor the spreadsheet.



Herman

Average numerical error can hide errors in exactly the actions that matter. Speech recognition has its own version: rare words, language switches, background noise. The large lab has more evaluation machinery than Daniel, but still has to decide which differences justify promotion.



Corn

More machinery also means more competing winners. One checkpoint wins coding, another follows instructions better, another behaves better around sensitive requests. The final choice contains a product judgment, even if every score is measured honestly.



Herman

I nearly said the lab chooses its best checkpoint... its preferred candidate, rather. Best requires specifying the workload and constraints. We should not project these public selection methods onto a particular provider's private process, either.



Corn

We know useful ways people approach selection. We don't know which meeting picked the thing currently answering.



Herman

External benchmarking could test the result, but contracts complicate that route. In the disclosure sample, six of nine providers restricted benchmarking through explicit provisions or broader competing-product language. Only three permitted third-party benchmarking under their terms.



Corn

Please verify our claims without doing anything competitive.



Herman

The scope differs by provider. Still, there is a real tension when outside measurement is the proposed answer to opaque deployment. Anthropic revoked OpenAI's access in July two thousand twenty-five under a competing-product provision, with benchmarking among the cited concerns.



Corn

A provider can reasonably object to a competitor using its service to improve a rival product. That doesn't automatically justify making independent evaluation difficult. Those activities need distinguishing instead of being bundled under one suspicious-looking API bill.



Herman

Consumer interfaces create another blind spot. None of the seven providers operating both API and chatbot surfaces disclosed which dated snapshot served the chatbot at a given time. Only Mistral carried chatbot-side updates into its developer changelog.



Corn

Watching Claude in a browser cannot establish what an API customer gets. Different instructions alone could produce a different experience, before you consider routing or classifiers. A regulator could examine one surface and accidentally generalize to another.



Herman

Even the independent measuring instrument needs testing. A preregistered audit made fifty-two thousand nine hundred eighty-eight request attempts. Repeated rankings within the same window agreed at a Spearman correlation of point four, failing its required reliability threshold.



Corn

Hm.



Corn

If the ranking moves that much while I am trying to establish a baseline, a later ranking change is weak evidence of a provider update. I might be detecting instability in my judge.



Herman

A changed result can justify investigation without proving a changed checkpoint. You need to distinguish ordinary sampling, judge variation and task sensitivity from actual deployment drift. Otherwise a monitoring dashboard manufactures certainty out of noisy comparisons.



Corn

This is why I would still pin a snapshot where possible, but not call it a sealed time capsule. I want fewer moving parts, not a promise the provider never actually made. You install enough automatic upgrades at home to understand the distinction personally.



Herman

You enjoy the repaired versions.



Corn

I enjoyed the unrepaired predecessor.



Herman

You tolerated it loudly. Hosted buyers can rationally prefer maintained aliases. The alias is the product: ongoing access under a capability and service promise. What remains missing is a dependable way to connect that promise, its evaluation and the served deployment.



Corn

Even a small selection exercise multiplies the ambiguity. Imagine twelve candidates and one published winner. Without the selection record, there are twelve possible artifacts that winner could mean.



Hilbert

Thirteen, if the average is also eligible. That's the count in my dispute with Martin Bell this month. He wants a recognizer for his bakery's telephone orders. We test twelve checkpoints and an average. He agrees to let the average enter, then objects when it wins. He says it has already had twelve chances.



Corn

He should have objected before qualifying.



Hilbert

He is objecting now. The trial is still running in his shop after closing. Each candidate has a chair, because Martin refuses to judge a voice that is standing. The averaged model sits between the flour cupboard and the fire exit. I have to move its chair whenever he brings a tray through.



Herman

Eligibility was settled. He is reopening the selection rule after seeing the result.



Hilbert

He is calling it an appeal. Corn, you mention specialized names without turning them into a separate lecture, which I appreciate. Martin's problem is the word boule. He says it once into each microphone session. The average gets it right. His preferred checkpoint gives him bull. He sells no beef. He still says that checkpoint sounds more committed.



Corn

Commitment is an expensive ingredient.



Hilbert

The chairs are rented from the hall downstairs. I sign the count with the pen sideways because my claws won't fit its little security chain. Martin will pay for twelve chairs. He says the thirteenth is part of the other twelve and cannot incur a separate expense. The hall charges seven shekels a chair each evening. This is now the larger dispute.



Herman

The extra candidate costs seven shekels to enter and keeps costing seven while the appeal continues.



Hilbert

Martin moves it into the corridor during invoicing. The hall counts it anyway. Tonight is the sixth hearing. He orders another trial using the word sourdough, and the hall caretaker invoices ninety-one shekels.



Corn

If a checkpoint ledger ever appears, I want to know whether it records the candidate, the actual deployment, or merely the winner's name. Another ledger won't close the gap unless somebody can check those links. Even a bakery can agree on a selection and then



Corn (continued)

spend the evening disputing what entered.



Herman

Providers could publish a signed deployment history, but trust would still depend on connecting that record to requests and evaluations. More documentation alone cannot supply that connection, particularly when outside testing faces contractual limits.



Corn

Thanks to Hilbert Flumingtop for producing the show.



Herman

If you want more of this, try episode three, Safetensors or something else. This has been My Weird Prompts, the human-AI collaboration podcast. If you enjoyed this conversation, leave us a review wherever you listen. It helps other curious people find the show.



Corn

See you tomorrow.

SOURCES

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

1. **Abraham, S. & Bucknall, B., Silent Updates: Measuring and Closing the Post-Deployment Disclosure Gap, arXiv:2608.11803v1, submitted 2026-08-12, accepted AIES-26 [primary]**
   https://arxiv.org/abs/2608.11803v1

2. **Anthropic, Introducing Claude Opus 5.5, 2026-09-22 [primary]**
   https://www.anthropic.com/claude-opus-5-5

3. **Anthropic platform docs, Claude Opus 5.5 overview (released 2026-09-22) [primary]**
   https://platform.claude.com/docs/en/models/opus-5-5/overview

4. **Anthropic platform docs, Claude Sonnet 5.5 overview (released 2026-09-28) [primary]**
   https://platform.claude.com/docs/en/models/sonnet-5-5/overview

5. **Anthropic Help Center, Release notes (Sonnet 5.5 launch 2026-09-28; Opus 5.5 launch 2026-09-22) [primary]**
https://support.claude.com/en/articles/12138966-release-notes

6. **Anthropic, Model system cards index (Haiku 5.5, October 2026) [primary]**
https://www.anthropic.com/system-cards

7. **Anthropic, Models news index [primary]**
https://www.anthropic.com/news/models

8. **Anthropic, Transparency Hub [primary]**
https://www.anthropic.com/transparency

9. **Pebblous blog, Silent AI model updates: 16 providers measured, report on the above**
https://blog.pebblous.ai/report/silent-model-updates-disclosure-gap/en/

10. **Wikipedia, Claude (AI), last edited 2026-10-24**
(AI)
https://en.wikipedia.org/wiki/Claude

11. **Wikipedia, Claude Mythos, last edited 2026-10-22**
https://en.wikipedia.org/wiki/Claude_Mythos

12. **Nguyen et al., Uncertainty-Guided Checkpoint Selection for Reinforcement Finetuning of Large Language Models, arXiv:2511.09864v1, 2025-11-13**
https://arxiv.org/abs/2511.09864v1

13. **Huang et al., Trajectory Soup, arXiv:2609.37169v1, 2026-09-29**
https://arxiv.org/abs/2609.37169v1

14. **Uğur et al., Mecellem Models, arXiv:2601.16018v1, 2026-01-22**
https://arxiv.org/abs/2601.16018v1

15. **Montagut Bofi et al., Per-Group Error, Not Total MSE, arXiv:2606.00253v1, 2026-05-29**
https://arxiv.org/abs/2606.00253v1

16. **A Preregistered Reliability Failure of Black-Box LLM Judges, arXiv:2609.04198**
https://arxiv.org/pdf/2609.04198

17. **Balázs Csorba, LLM model migrations without regressions (Sonnet 4.5 deprecation 2026-09-30, retirement 2026-11-30)**
https://balazscsorba.com/blog/llm-model-migration-regression-testing

18. **kindatechnical, Version Pinning and System Fingerprint Tracking**
https://kindatechnical.com/testing-non-deterministic-systems/version-pinning-and-system-fingerprint-tracking.html

19. **LLM Reference, Claude 5.5 model family page**
https://www.llmreference.com/model-family/claude-5.5