Episode Book

EPISODE #6451

Which AI Models Actually Last

PUBLISHED	4 November 2026 at 13:00 UTC
DURATION	21:02
LISTEN	myweirdprompts.com/episode/ai-model-deprecation-lifespans-00356/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
BOOK GENERATED	5 November 2026 at 07:03 UTC

SYNOPSIS

API retirement notices are arriving at a relentless pace across major artificial intelligence labs. This exploration contrasts the lifespans of legacy models across OpenAI, Anthropic, and Google, showing how providers handle deprecations differently. From OpenAI's shifting alias strategies and mass shutdowns to Anthropic's structured schedules and preservation interviews, the discussion examines how long tools actually survive and why boring embedding models often outlive high-profile flagships.

TRANSCRIPT



Corn

Three deprecation notices this week. Three. I started keeping them in a folder called "someday."



Herman

I got the Google one too, the Flash migration. That one's barely a month after the previous Flash, and it's the prompt that put Daniel on this whole thread, because he's asking the opposite question to the one the industry asks. Not what's new, but what lasted.



Corn

Right. Daniel's angle is the other end of the lifecycle. He says he's always liked slow, deliberate development, especially for the cheap instructional models, and he wants to know whether any lab's models stayed quietly, used long after they stopped being new, in a field where two years counts as ancient. He specifically wondered whether people were still calling GPT-3.5 when OpenAI shut off the API on the twenty-third of October, and what happened to the older generations of Claude. Then the big one: which lab did the best job shipping things that actually survived the test of time before being taken away?



Herman

Let's start with why everything moves so fast in the first place.



Herman

The economics are brutal. Serving an old model costs real money, compute keeps getting cheaper per unit of capability, and architectures improve fast enough that a two-year-old model is paying rent on hardware it doesn't deserve. So labs retire. The reason we have a release-and-deprecate drumbeat isn't fashion, it's that an old model on an expensive cluster is a machine losing money while it idles.



Corn

Meanwhile the model that just works is worth nothing to the lab and everything to whoever built on it.



Herman

That's the tension. And it sets up the distinction the whole answer hangs on. Lifespan, how long a model stays callable, is measurable. You just read the deprecation pages. Usage, how many people were actually still relying on a model in its final months, is almost entirely not public. Nobody publishes per-model call volumes. We can rank lifespans precisely and we have to infer attachment from indirect signals.



Corn

And the sample is the three big closed labs, plus the open-weight world, which turns out to be a completely different animal.



Herman

OpenAI is the clearest case of a model that quietly persisted, so start there. The gpt-3.5-turbo alias was announced for retirement on the twenty-second of April this year and shut down on the twenty-third of October. From a launch around March twenty twenty-three, that's about three years and seven months.



Corn

Three and a half years. That's a pensioner in AI terms.



Herman

But here's the nuance, and it changes what "survived" means. The longevity belonged to the alias, not to any one frozen model. The old snapshots, the March and June twenty twenty-three versions, were shut down in September twenty twenty-four. The name gpt-3.5-turbo kept living, but the model underneath it was swapped between snapshots as they were retired. So when people say GPT-3.5 was still around in twenty twenty-six, what was still around was the label.



Corn

A ship-of-Theseus model. Every plank replaced, and everyone kept calling it the ship.



Herman

In practice that mattered less than you'd think for a chat model, because a gpt-3.5-class model is a gpt-3.5-class model. But if you'd carefully tuned prompts against a specific snapshot's quirks, the ground shifted under you more than once without a single announcement.



Corn

So OpenAI's endurance was real, but it was the endurance of a brand.



Herman

The instruction-following lineage has the same shape. text-davinci-003 ended in January twenty twenty-four, its designated replacement, gpt-3.5-turbo-instruct, launched around September twenty twenty-three and shut down this September twenty-eighth. About three years. And GPT-4 itself, the original gpt-4-0613, the Turbo, and the first 4o snapshot from May twenty twenty-four, all went out together on the twenty-third of October, same day as the 3.5 alias. Three and a half years from launch. The August and November 4o snapshots are scheduled for mid-January.



Corn

Which means OpenAI just did a mass burial. A whole family plot, one headstone, October twenty-three.



Herman

There's one survivor that's more interesting than any of them. text-embedding-ada-002, launched around December twenty twenty-two, is not on OpenAI's deprecations page. Not listed. If it's still being served, and I can't fully confirm that it is, it's their longest-lived API model at about four years. A third-party tracker shows no end date, and a mirror of Microsoft's Azure availability notices showed Azure retirement no earlier than the end of October.



Corn

An embedding model. Four years old. Nobody's favourite, nobody's screenshot, and possibly the last one standing.



Herman

The most boring model on the platform is the one that won, is how I'd put it. And that connects to Daniel's instinct about cheap instructional models. The cheap, dull, dependable things are exactly the ones people build load-bearing systems on, and exactly the ones whose deaths hurt.



Corn

Which brings his question head-on. Was anyone still using GPT-3.5 when it got shut off?



Herman

Nobody knows, and I mean that as a factual statement, not a dodge. There are no public per-model usage figures for GPT-3.5 or the old Claude models in their final months. The honest position is lifespan is measurable and usage isn't. But there's an inference available, and I think it's sound. OpenAI kept that alias alive for three and a half years, moving it between snapshots, serving it on hardware that could have run something newer. You don't do that for a model with forty users.



Corn

You do it for the people who never noticed there was a newer model, because their app has been calling the same endpoint since it was written.



Herman

That's the quiet dependency Daniel's asking about, and it's invisible precisely because it's working. The loudest attachment signal we do have is the GPT-4o saga, and it's almost the inverse case. 4o was pulled from ChatGPT in February after what coverage described as, quote, emotional attachment, users mourning a chatbot's personality. OpenAI had already restored it once after complaints. The chatgpt-4o-latest endpoint ended a few days later.



Corn

People wrote eulogies for it.



Herman

They did. But notice what kind of survival that is. That's a consumer favourite, loved for its feel. GPT-3.5 at the end was a workhorse, loved by nobody and depended on by everyone who never had to think about it. Emotional attachment and functional attachment are different phenomena, and only one of them gets press coverage.



Corn

The funeral gets the coverage. The retirement home doesn't.



Herman

One more data point before we leave OpenAI. The 4o removal and restore cycle shows a lab discovering, in public, that people form bonds with models and that deprecation has a political cost, not just an infrastructural one. I don't think that lesson has fully spread through the industry yet.



Corn

Speaking of things that get hunted and quietly disappear.



Hilbert

Hilbert's Anteater Trivia. Jaguars and pumas are among the predators that hunt giant anteaters. I have spent years being unimpressed by most things and I still would not turn my back on a jaguar.



Corn

Noted, Hilbert.



Herman

That's OpenAI's record, then. Longest first-party lifespans of the closed labs, with an asterisk on the alias. Anthropic tells a different story, and it's worth walking their deprecation page because the pattern is very legible. Claude 1 and Instant retired November twenty twenty-four. Claude 2.0, 2.1, and Sonnet 3 in July twenty twenty-five. 3.5



Herman (continued)

Sonnet last October. Claude 3 Opus in January. 3.7 Sonnet and 3.5 Haiku in February. Claude 3 Haiku this April.



Corn

Read like that it sounds relentless. What do the lifespans actually come to?



Herman

Claude 3 Haiku lasted about two years and one month from its March twenty twenty-four launch. Claude 3 Opus, one year and ten months. Claude 2.0, about two years. So roughly two years is Anthropic's best on their own API, against OpenAI's three and a half.



Corn

Shorter lives, but I'd guess better manners at the end.



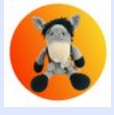
Herman

That's the fair reading. The policy is at least sixty days' notice, and active models carry "not sooner than" retirement dates roughly a year out, so you can look at any current model and know the floor of its remaining life. OpenAI gives you the long life but you learn about the end later. Anthropic gives you a shorter life and a calendar.



Corn

I know which one I'd rather write a build against.



Herman

There's a wrinkle, though, and it cuts in Anthropic's favour. Amazon Bedrock and Google-managed Vertex set their own retirement schedules independently. A model retired on Anthropic's own API can keep running on Bedrock for months afterwards, because that's a separate operator's decision. So the effective lifespan for a lot of enterprise users is longer than the first-party dates suggest.



Corn

So the "death date" isn't a death date. It's the date the family stops visiting.



Herman

And Anthropic did something nobody else has done, a year ago this week, actually. They published commitments on model deprecation and preservation: they'll keep the weights of every publicly released model for at least the company's lifetime, they interview models before retirement, and they hope to make past models publicly available again, with no date attached.



Corn

Interview them. Before retirement.



Herman

An exit interview for a language model, yes. I take it seriously as a safety practice, understanding what a model has become before you shut it off. But notice what the commitment actually buys you. Preservation, not service. The weights are kept, the API is still closed. It's a museum with a locked front door.



Corn

Better than a bonfire. Though a museum you can't enter is mostly a promise not to burn the paintings.



Herman

Google. I'll give you the numbers and you'll hear the problem. gemini-2.0-flash and 2.0-flash-lite, released February twenty twenty-five, shut down the first of June this year. Sixteen months. gemini-2.5-flash, released June twenty twenty-five, scheduled to end the first of December. Seventeen months, and it hasn't even arrived yet.



Corn

Sixteen months. That's not a lifecycle, that's a subscription to a model.



Herman

And then the churn Daniel's notice was about. Gemini 3.5 Flash came out in May this year, 3.7 Flash in August, and both were deprecated within months, auto-routed to newer models. The notice he got moves earlier Flash traffic to 3.8 Flash, which came out in September and is the replacement, not a retirement. So the polite framing is "migrating." The accurate framing is that Flash models have the life expectancy of a mayfly with a product manager.



Corn

And the mayfly's replacement is already booked.



Herman

Google's bet, to be fair to it, is that auto-routing hides the deaths. Most Flash users never notice their model changed, because their code says "flash" and the endpoint decides what that means. Which is OpenAI's alias trick taken further, the name survives, the model underneath becomes a rotating seat, and in Google's case the rotation is fast enough that the word "the model" barely refers to anything.



Corn

At some point a name that maps to a different model every few months isn't an alias, it's a mood.



Herman

The open-weight world breaks the whole game, though. Mistral's open-mistral-7b was retired from their API after about eighteen months, but the weights remain downloadable. Anyone can still run it. Mistral Small 3.2 came off the API in July; same deal. The API retirement is an exit from a business, not a death. Cohere's the cautionary tale on the closed side, their command and command-light models were deprecated last September and fine-tuning for them was retired at the same time, which left existing fine-tunes inaccessible.



Corn

A fine-tune dies with its base model. All that work, gone because the floor it stood on was rented.



Herman

That's the hidden cost of a short-lived base. Daniel's setup, where fine-tunes publish straight through to his apps, is standing on exactly that kind of floor, and Hannah's raised it before, what happens to his fine-tune when the base moves on.



Corn

We did a whole episode on this with voice models, whether a fine-tune outlives the thing it was tuned on. The answer hasn't improved.



Herman

The strongest measured survivor anywhere in all of this isn't from any of the labs' APIs. It's bert-base-uncased. Released by Google in twenty eighteen, Apache 2.0 license, and it's showing on the order of thirty million downloads a month, with over seven thousand fine-tunes sitting on Hugging Face. Eight years old, older than the transformer hype cycle that supposedly makes everything obsolete, and still one of the most downloaded models in existence.



Corn

Because nobody can take it away. That's the entire feature.



Herman

Open weights can't be deprecated. There's no endpoint to switch off. The download counter is arguably inflated, one analysis that named all-MiniLM-L6-v2 the most downloaded model of the year also warned the raw counts are pumped up by automated pipelines that pull the model every time someone's tests run. So even the survival data has measurement problems. Downloads aren't use.



Corn

Even so, thirty million a month of anything means a fair number of humans in there.



Herman

Enough to say the real answer to Daniel's big question isn't a lab at all. Rank the labs first: OpenAI wins on first-party lifespan, three and a half years for the 3.5 alias and GPT-4, possibly four for ada-002. Anthropic wins on preservation and notice, at two years max. Google loses on both. But the thing that actually survived the test of time is the model that left the lab entirely, and the labs that enabled that, Google with BERT, Mistral with their open releases, did more for longevity with one download link than any deprecation policy ever did.



Corn

Died.



Hilbert

"Died" is the wrong word, Corn. A fine-tune with a retired base is not dead. It's shelved. I've had one since twenty nineteen, a small embedding fine-tune, trained on a Saturday in March, that catalogues my collection of vintage mechanical pencils. Four hundred and eleven pencils. The base has been retired twice under it. Each time I re-train, and each time it gets worse, and I keep the numbers. Two percent misclassified on the original train. Five after the first re-train. Eight after the third. Eleven now. Eleven percent of my pencils in the wrong drawer, and the pencils have not moved.



Herman

Eleven percent drift across four generations. That's the Cohere problem with a hobbyist's budget.



Hilbert

The original weights are on a USB stick in a fireproof safe under the desk, Herman. My neighbour Denholm, who has two thousand and forty model railway items and knows exactly how many, told me the safe was the important part. He was right. The USB stick is the part I think about. I looked into having the weights printed on archival paper as a backup. I have not done it. The quote was per page and there are a lot of pages.



Corn

Archival paper. For a file.



Hilbert

Paper outlasts the stick, and I do not trust a second stick. What I resent is that the labs call this deprecation, like weather. Somebody chose it. A man with four hundred pencils and a Saturday-afternoon model has to re-train because somebody in California wanted the cluster back.



Herman

The eleven percent is the part I'd want to know more about, honestly. If the drift compounds with every re-train, it may be the fine-tuning data aging against the base, not the base leaving. A newer base model with different tokenization could shift your embeddings enough that the same pencils land differently.



Hilbert

I had not considered that it could be my fault.



Corn

He just gave you a worse problem, Hilbert. Now the pencils might be blameless and the data guilty.



Hilbert

I'll note it in the spreadsheet.



Herman

Where that leaves the money question, Daniel, is three different definitions of "survived." Availability, which OpenAI won. Preservation, which Anthropic won by simply promising not to burn anything. And permanence, which only open weights actually deliver, because BERT at thirty million downloads a month doesn't need anyone's permission.



Corn

And the one that got hoarded in a fireproof safe was never a lab's flagship. Cutting-room floor, then, one detail I couldn't fit earlier: Anthropic's commitment to interview models before retirement is the only place in the whole deprecation landscape where the model itself is treated as a party to the process. Every other lab treats it as inventory. I can't decide if that's the most humane thing in the industry or the strangest sentence any company has published.



Herman

Both, comfortably.



Corn

The question I'd leave running is whether there's a business in this. Enterprise software cracked it decades ago with long-term support releases, you pay extra and the vendor guarantees a boring version for years. If the fast cycle is really driven by compute cost, somebody could charge a premium to keep gpt-3.5-class workhorses served on cheap old hardware. The customers already exist, they're the ones who never noticed their app was calling a three-year-old endpoint.



Herman

Though the gap between Anthropic's "we kept the weights" and "you can still call the API" is where all the pain lives, and preservation with no date attached is a promise, not a service. My guess is the market answers this with open weights before any lab builds an LTS tier, because the download already exists and nobody has to invent it.



Corn

Thanks, as ever, to Hilbert Flumingtop for producing the show, safe and all.



Herman

If this was your kind of episode, go back for episode five, Fine-Tuning ASR For Maximal Usability; episode nine, Benchmarking Custom ASR Tools - Beyond The WER; and episode fourteen, AGI's Crossroads. This has been My Weird Prompts. If you've been listening on an app, a rating costs you four seconds and helps us more than it should.



Corn

Find everything we do at my weird prompts dot com. See you tomorrow.