Episode Book

EPISODE #6559

Who Trains on Your AI Chats and What Happens Next

PUBLISHED	13 November 2026 at 20:17 UTC
DURATION	21:12
LISTEN	myweirdprompts.com/episode/ai-chat-training-defaults-46133/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	14 November 2026 at 06:47 UTC

SYNOPSIS

Consumer tiers of OpenAI, Anthropic, and Google Gemini all train on chats by default, while API and business tiers do not. Even with opt-outs enabled, feedback buttons, safety reviews, and separate Codex settings can still expose conversations. The bigger privacy risk isn't models memorizing secrets—it's five-year retention windows and human reviewers. Also covered: how OpenRouter publishes provider retention policies, and what research from Mireshghallah and Li alongside PII-Compass reveals about actual data extraction risks.

TRANSCRIPT



Corn

Daniel wants to know if the thing everyone mutters at dinner parties is true — "I don't want to type that into an AI, it'll train on it." He trains models himself, so he came at it with a practitioner's instinct: if raw private chats were going into these things, they'd be blurting out secrets within days, and there'd be a scandal a week. Nobody has that scandal, which he takes as evidence the fear is misplaced. But he also knows OpenRouter will route you to providers that train on prompts in exchange for free inference, so he's asking us for the real picture — who trains on what, on which default, and what actually happens to a chat once it's been swallowed.



Herman

And the short answer is that his instinct is right about the mechanism and wrong about the premise. Which is the most interesting way to be right and wrong at the same time.



Corn

The premise first, because it's fast. If you use the free or consumer tier of any of the big three — OpenAI's Free, Plus or Pro, Anthropic's Free, Pro or Max, Google Gemini — your conversations can be used for training by default. All of them. There's no holdout.



Herman

OpenAI's help center says it flat out: "When you use our services for individuals, Free, Plus, Pro, we may use your content to train our models." The opt-out is the Improve the Model for Everyone toggle in settings, or "Do not train on my content" in the Privacy Portal — and they're clear that either one is sufficient. You don't need to hunt down both like it's a scavenger hunt.



Corn

Anthropic's consumer terms got an update in August of twenty twenty-five that gave Free, Pro and Max users an actual training choice, with the toggle on by default. Existing users had until early October that year to decide, and if you did what most people do — clicked accept and moved on with your life — you opted in.



Herman

WIRED ran a piece on it at the time, precisely because the toggle being automatically on meant a lot of people consented without ever seeing a toggle. It was consent by default text size.



Corn

Google is the bluntest of the three, which I almost respect. Gemini's Keep Activity is on by default once you're eighteen, auto-delete defaults to eighteen months, and with it on your activity is used to, quote, provide, develop and improve its services, including training generative AI models.



Herman

And the single most useful fact for anyone listening is the shape underneath all this: consumer tiers are uniformly on, API and business tiers are uniformly off. OpenAI says it plainly — by default they don't use inputs or outputs from ChatGPT Business, Enterprise, Edu, or the API to improve models unless an org owner deliberately enables data sharing. That's the cleanest default in the industry.



Corn

So the surface you're typing on matters more than the logo on it.



Herman

And the same chat can live in two regimes at once. Temporary chats in ChatGPT aren't used for model improvement, but save one and it becomes a regular chat and follows your account's normal settings. The wall is only a wall while you don't touch it.



Corn

Herman, before we get to Daniel's corner of it — OpenRouter.



Herman

OpenRouter is the honest broker here, and I mean that almost literally. It's the one place where "some providers train on your prompts" is stated up front, published as structured data per endpoint, and controllable. There are separate toggles for paid and free models, and if you opt out of training in your account settings, OpenRouter simply won't route you to providers that train. They publish a three-tier vocabulary: Zero retention, meaning no routine storage after processing; Safety Retention, where prompts may be kept only to prevent abuse and it's not used for training; and a third tier where providers retain data for a stated period and may train if marked. Their docs table shows NVIDIA prompts as retained for an unknown period and "May train," OpenAI as retained for an unknown period and "Does not train."



Corn

"Retained for unknown period." Both of them. That's a comfort.



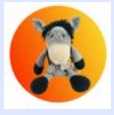
Herman

Their own phrasing is careful too. "Wherever possible, OpenRouter works with providers to ensure that prompts will not be trained on, but there are exceptions." Wherever possible. That's doing a lot of load-bearing work for three words.



Corn

It's the most transparent any of this gets, and it's a reseller. The first parties bury the same facts three menus deep.



Herman

So that's the map. Defaults by company, by plan, by surface. Now — what actually happens to a chat once it's been picked up?



Corn

That's where Daniel's mental model is half right in an interesting way. He assumed a curation step — that somebody selects good conversations and throws away junk, because random chats are terrible training material. And the providers don't describe that. What they describe is de-identification.



Herman

Anthropic is the most specific. Where you've allowed them to use your chats and coding sessions, they automatically de-link them from your user ID — the email, gone — before the chat is used, and then they run tools derived from their privacy-preserving analysis work to filter or obfuscate sensitive data. OpenAI's version is a single line: before using eligible user content, they take steps to reduce the amount of personal information in training datasets. Reduce. Not remove.



Corn

"Reduce" is a word with a percentage hiding inside it.



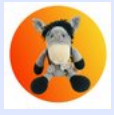
Herman

And there's a layer above the data itself. Anthropic's constitutional training includes principles like "choose the response that has the least personal, private, or confidential information belonging to others" — the model is trained toward not disclosing personal data even if prompted. Their privacy center also makes the conceptual argument directly: models don't store text like a database, don't collage existing content, and don't have access to training data once trained. Which is true as far as it goes.



Corn

But nowhere in any of those pages — OpenAI, Anthropic, Google — does anyone say "and we select the high-quality conversations." That's the gap. Daniel intuited a curator. What exists is a scrubber.



Herman

Whether quality filtering happens is simply not stated publicly by any of them. It probably does, because you'd be mad to fine-tune on the full firehose, but no provider has put their name to it.



Corn

So the real risks turn out to be the leaky opt-out, the retention window, and the share button — not the model memorising anything.

L

Larry

Friends, are you worried your chats are being trained on? Well stop worrying and start scrubbing, with the Privacy Preserving De-Linking Personal Information Obfuscation Scrubber and Curated Data Filter, the only device that takes your conversations, de-links them from your email, filters the sensitive bits, and hands you back a model that has memorised nothing about you whatsoever. And here is the best part: the price is dropping as I speak. It was four hundred and ninety nine, now it is three ninety nine, now two ninety nine, and for the next thirty seconds only, one ninety nine, and I will throw in a retention window, a leaky opt-out patch kit, and a share button that never shares anything, not even with your mother.



Herman

Wait, the retention window doesn't come with the... I don't think that's how — the three ninety-nine wasn't even —



Corn

Herman. Leave the man to his pricing strategy.



Herman

The leaky opt-out, though. This is the part I'd want Daniel to internalise, because the toggle looks like a switch and behaves like a suggestion. Three concrete leaks on the OpenAI side alone. First: feedback. Their help center says, "If you choose to provide feedback, the entire conversation associated with that feedback may be used to train our models" — even if you've opted out. You turn everything off, then you give a response a thumbs down because it was wrong about something, and you've handed over the whole conversation.



Corn

The thumbs button outranks the settings page.



Herman

Second: Codex has a separate "Include environments" setting, and OpenAI is explicit that changing your ChatGPT settings or the Privacy Portal doesn't change it. It's its own island. Third: conversations with OpenAI Support themselves may be used to improve services if the model-improvement setting is on — the support channel inherits the setting you



Herman (continued)

thought was about your chats.



Corn

And the other two have their versions.



Herman

Anthropic has four conditions under which chats may be used, and the one that matters is the safety carve-out. Conversations flagged for safety review may be used or analyzed to improve their ability to detect and enforce their usage policy, "including training models for use by our Safeguards team" — regardless of your training setting. Google's version: even with Keep Activity off, Google still uses your chats to respond to you and to help protect Google, users and the public, including with the help of human reviewers. The setting doesn't control processing to create anonymized data either.



Corn

It's the same shape we hit in the messaging-app episode — the privacy setting that appears to govern a thing it only partially governs. WhatsApp all over again, with a nicer font.



Herman

And Google is oddly honest about the human part. The Gemini privacy notice says, and I want to read this one because I wouldn't have believed the tone otherwise: "Please don't enter confidential information that you wouldn't want a reviewer to see or Google to use to improve our services, including machine-learning technologies." Please don't. It's the please that gets me. They're not hiding the reviewer. They're asking you nicely to accommodate them.



Corn

A human reading your chats, disclosed in the tone of a hotel asking you not to steal the robes.



Herman

Reviewed chats on Gemini are retained up to three years, and per their privacy hub, they're not deleted when you delete your activity. Deleting is for your timeline, not for their folder.



Corn

Which tees up the thing nobody talks about — retention. Because the Anthropic numbers are the most concrete risk in this whole landscape, and they're not about models at all.



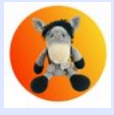
Herman

Per WIRED's reporting, if you allow training, retention extends to five years. If you don't, thirty days. Five years versus thirty days is your data sitting on a disk sixty times longer because you left a toggle on. That's not a hypothetical about gradient descent. That's a hard drive with a five-year lease on your Tuesday-night questions.



Corn

Sixty times. I've let a lot of things slide for less.



Herman

And there's a position paper from Mireshghallah and Li that makes exactly this argument at the field level. They analyzed over thirteen hundred AI and machine-learning privacy papers from twenty sixteen through twenty twenty-five, and their conclusion is that the discourse has disproportionately focused on verbatim memorization of training data, while — and this is close to a quote — the most pressing privacy harms lie elsewhere. Data collection practices, inference-time leakage, agents. The entire research community has been staring at the thing Daniel's dinner-party worrier is staring at, and both of them are looking at the wrong end of the pipe.



Corn

So let's give the memorization question its day in court. How real is it?



Herman

Real, measurable, and small — and almost all the literature is about pretraining corpora, not chat logs. The PII-Compass work ran extraction attacks for phone numbers: roughly one percent success with a single query, about four percent at a hundred and twenty-eight queries, and just under seven percent at two thousand three hundred. Their own framing at the top end is that the phone number of one person in fifteen is extractable — and that's with prompts carefully grounded in-domain, which they showed improves extractability by more than tenfold over naive attacks. Seven percent is not nothing. One in fifteen is a real number if you're a hospital. It's a small number if you're one person asking ChatGPT about your back pain.



Corn

And there's a dialogue-specific result.



Herman

Zhang and colleagues, in a paper accepted to IEEE TIFS this year, extracted training labels from LLM-based task-bot dialogue — travel schedules, phone numbers, whole dialogue-level events — with best-case precision over seventy percent. Now that sounds alarming, and it's the closest thing in the literature to the chats-leak scenario. But it's task-oriented dialogue, structured slots, a system trained on that specific corpus. It's not free-form consumer chat that's been de-linked and scrubbed before it ever hit a training run.



Corn

So the attack works great on the data nobody scrubbed.



Herman

There's also a BERT result I find quietly reassuring — a fine-tuned BERT doesn't generate more named entities specific to its fine-tuning dataset than the pretrained base model did. And much of what consumer chats feed into is exactly that kind of light fine-tuning, not a from-scratch pretrain. Small, targeted adjustments leave less fingerprint than the corpus they were drawn from.



Corn

Which leads us to the one time this whole fear actually went public. And it wasn't training at all.



Herman

Late July, twenty twenty-five. Fast Company found nearly forty-five hundred shared ChatGPT conversations indexed in Google search results — you could find them with a site-restricted search on the share domain. Some of them revealed addiction struggles, mental-health details, things people clearly never intended for a search engine. The researcher Henk van Ess found over five hundred directly, and over a hundred and ten thousand more through the Wayback Machine. OpenAI removed the feature and their security chief called it a "short-lived experiment."



Corn

Short-lived experiment. The two-word obituary of every feature that ships before the privacy review.



Herman

The mechanics of it matter for Daniel's question. To leak, a user had to share a chat, and during sharing there was a box — Make this chat discoverable — with the caveat in smaller text underneath. People ticked boxes they didn't read, links got crawled, and Google did the rest. The leak vector was sharing and indexing. Training was nowhere near it.



Corn

The database everyone's afraid of never leaked. The share button did.



Hilbert

What does "improve our services" mean? Either of you. You've both said it four or five times this hour and neither of you has said what it means.



Corn

It means whatever the sentence needs it to mean, Hilbert.



Hilbert

That's what I've been writing to a company about since ninety-two. A seed supplier — we ran the allotment club's bulk order through their dial-up ordering line. Vic handled the seeds, I handled the modem, because nobody else would sit still for it. You'd dial in, leave your messages on their board. I left one about a short-shipped order of seed potatoes, and the reply came back — your messages may be used to improve our services. So I asked them which services, and by how much. That was the start of it.



Herman

You've been corresponding with a seed company for thirty-odd years?



Hilbert

I've been quoting the phrase back at them. There's a difference. I have the exchange printed out, in the green folder. I've gone through it and circled every place their side of the exchange says "improve our services." There are eleven. In a conversation about nine kilos of seed potatoes, eleven. Herman, you'd have counted them the same way.



Herman

I would have, yes.



Hilbert

The support ticket on their end is still open, because closing it would be accepting the terms. That's my position and I've stated it in every letter. Whether it's still open on their end, I couldn't say. I haven't checked. Checking would be accepting that it might have closed.



Corn

You've kept a dispute alive for three decades through strategic non-verification. That's the most sustainable data-retention policy I've ever heard of.



Hilbert

It's the patter, Corn. "Improve our services" is sales patter that passed a legal review, and everybody types around it like it's load-bearing. Your Daniel said the models would be repeating secrets within days — no. The phrase is what never gets repeated. And for the record, when you two were on the twenty twenty-five thing earlier — the box didn't say "publish to web." It said Make this chat discoverable. Those are different words and the difference was the whole point.



Herman

He's right, I conflated those.



Hilbert

Vic gave up his allotment in ninety-four. Thought you'd want the order of events.



Corn

The patter. I'm keeping that.



Herman

It does deserve the name, because it explains something about Google's warning. "Improve our services, including machine-learning technologies" — the phrase has to be vague enough to cover everything, which is exactly why it covers everything, including a man in a green folder circling it eleven times.



Corn

Notice what the actual risks were, all episode. The leaky opt-out — the thumbs button overruling the settings page. The retention window — thirty days versus five years. The share button — forty-five hundred conversations in Google because of a checkbox. Not



Corn (continued)

one of them is the model reciting anything. Daniel's instinct was right about the mechanism and wrong about the premise, and the reason the premise matters is that the defaults decide the disk, not the weights.



Herman

The disk is the thing with the lease on it. The question I'd leave him with isn't "will the model repeat my secrets" — it's who is holding this conversation, for how long, and under what phrase. Five years hides inside "improve our services" the same way eleven circles hide inside a seed order.



Corn

The open question I keep circling myself — no pun intended — is the number nobody publishes. What percentage of a frontier model's training mix comes from user chats at all? Is it one percent, is it twenty? Every provider could state it. Not one does. That's the number that would settle Daniel's debate, and its absence is the most eloquent thing in the whole policy stack.



Herman

If it were small, you'd publish it. That's my reading of the silence.



Corn

Mine too. Thanks as always to Hilbert Flumingtop at the desk — and to the green folder.



Herman

If you enjoyed this one, try episode thirty-five, The Privacy Gap; episode fifteen, AI Gets Personal; and episode sixteen, On Deepfakes, SynthID, And AI Watermarking. This has been My Weird Prompts. If today's episode earned its twenty-five minutes, leave us a review wherever you listen, and the full archive is at my weird prompts dot com.



Corn

Corn here, signing off for both of us.



Herman

We'll be back soon. See you tomorrow.

SOURCES

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

1. **OpenAI Help Center [primary]**
   How your data is used to improve model performance, updated ~24 days before 2026-11-13 (≈2026-10-20).
   https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

2. **OpenAI [primary]**
   How your data is used to improve model performance policy page, updated 2026-03-13.
   https://openai.com/policies/how-your-data-is-used-to-improve-model-performance/

3. **OpenAI [primary]**
   How ChatGPT learns about the world while protecting privacy, 2026-05-06.
   https://openai.com/index/how-chatgpt-protects-privacy/

4. **OpenAI Help Center [primary]**
   Data controls in ChatGPT, updated 3 days before 2026-11-13.
   https://help.openai.com/en/articles/7730893-data-controls-in-chatgpt

5. **OpenAI Help Center [primary]**
   Privacy Center in ChatGPT.
   https://help.openai.com/en/articles/20001488-privacy-center-in-chatgpt

6. **Anthropic Privacy Center [primary]**
   How Do You Use Personal Data in Model Training?, 2026-03-16.
   https://privacy.claude.com/en/articles/10023555-how-do-you-use-personal-data-in-model-training

7. **Anthropic Privacy Center [primary]**
   How do I change my model improvement privacy settings?, 2026-08-03.
   https://privacy.claude.com/en/articles/12109829-how-do-i-change-my-model-improvement-privacy-settings

8. **Anthropic [primary]**
   Updates to Consumer Terms and Privacy Policy, 2025-08-28.
   https://www.anthropic.com/news/updates-to-our-consumer-terms

9. **Anthropic [primary]**
   Privacy Policy (PDF).
   https://assets.claude.com/693a25036936f7e7bf4dd0c299a2e654f14fac01.pdf

10. **Google [primary]**
Gemini Apps Privacy Hub, page updated 2026-10-26; Gemini Apps Privacy Notice dated 2026-10-13.
https://support.google.com/gemini/answer/13594961

11. **OpenRouter Docs [primary]**
Provider Logging, Provider Data Retention Policies.
https://openrouter.ai/docs/guides/privacy/provider-logging

12. **OpenRouter Docs [primary]**
Data Collection.
https://openrouter.ai/docs/guides/privacy/data-collection

13. **OpenRouter Docs [primary]**
Zero Data Retention, planned update 2026-10-27.
https://openrouter.ai/docs/guides/features/zdr

14. **OpenRouter Docs [primary]**
Input & Output Logging.
https://openrouter.ai/docs/guides/features/input-output-logging

15. **OpenRouter [primary]**
Privacy Policy.
https://openrouter.ai/privacy

16. **WIRED**
Anthropic Will Use Claude Chats for Training Data. Here's How to Opt Out, 2025-09-30.
https://www.wired.com/story/anthropic-using-claude-chats-for-training-how-to-opt-out/

17. **DevOpsCompass**
Gemini Keep Activity: Does Google Train on Your Chats?, checked 2026-11-05.
https://www.devopscompass.com/2026/11/gemini-keep-activity-does-google-train.html

18. **Fast Company (Chris Stokel-Walker)**
Google is indexing ChatGPT conversations..., 2025-07-30.
https://www.fastcompany.com/91376687/google-indexing-chatgpt-conversations

19. **Ars Technica (Ashley Belanger)**
ChatGPT users shocked to learn their chats were in Google search results, 2025-08-01.
https://arstechnica.com/tech-policy/2025/08/chatgpt-users-shocked-to-learn-their-chats-were-in-google-search-results/

20. **The Telegraph**
Private ChatGPT conversations leak onto Google search results, 2025-08-01.
https://www.telegraph.co.uk/business/2025/08/01/private-chatgpt-conversations-leak-google-search-results/

21. **Engadget**
OpenAI is removing ChatGPT conversations from Google.
https://www.engadget.com/ai/openai-is-removing-chatgpt-conversations-from-google-194735704.html

22. **Search Engine Land**
Your ChatGPT conversations may be visible in Google Search.
https://searchengineland.com/google-indexing-shared-chatgpt-conversations-459839

23. **Nakka et al.**
PII-Compass: Guiding LLM training data extraction prompts towards the target PII via grounding, arXiv:2407.02943v2, revised 2025-08-23.
https://arxiv.org/abs/2407.02943v2

24. **Zhang et al.**
Extracting Training Dialogue Data from Large Language Model based Task Bots, arXiv:2603.01550v1, 2026-03-02 (IEEE TIFS).
https://arxiv.org/abs/2603.01550v1

25. **Mireshghallah & Li**
Position: Privacy Is Not Just Memorization!, arXiv:2510.01645v1, 2025-10-02.
https://arxiv.org/abs/2510.01645v1