Episode Book

EPISODE #5507

Inside ChatGPT's Hidden Pipeline: 8 Models Per Reply

PUBLISHED	24 September 2026 at 11:20 UTC
DURATION	26:33
LISTEN	myweirdprompts.com/episode/inside-chatgpt-hidden-pipeline/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.2
BOOK GENERATED	19 October 2026 at 17:37 UTC

SYNOPSIS

Ask ChatGPT to reword an email and you assume one model answers. In reality, a single turn can pass through input moderation, PII detection, a server-side safety router, memory assembly, retrieval orchestration, mid-generation activation classifiers, output moderation, and asynchronous review. Drawing on OpenAI's own documentation, the Privacy Filter open weights, the September 2025 GPT-5 telemetry paper, and published reverse-engineering of ChatGPT's memory and retrieval stack, this episode reconstructs the graph that sits between your prompt and the reply — and asks what it means that so much of it is now documented rather than speculative.

TRANSCRIPT



Corn

Here's a thing we said a while back, and I want to test whether it's still true. We did the exercise where you imagine lifting the lid on ChatGPT's image generator and instead of the one-button magic you find a graph. ComfyUI with a corporate logo on it. Nodes everywhere.



Herman

And the honest answer then was that the graph was real but mostly hidden. You could infer it, you couldn't see it.



Corn

Right. So Daniel wants to run the same experiment for the chat side. Text in, text out, one conversation thread. And he says something in the prompt that I think is exactly the right framing. He says, if you've built systems like this, you think you know what's happening. You send a request to the API, it's stateless, the model gets a system prompt and a bit of injected memory, it answers.



Herman

That's the mental model. It's clean, it's legible, and it's what almost everybody building on top of these things assumes.



Corn

Daniel's suspicion is that it's wrong. That if you actually got access to the back end you'd see all sorts of additional actors. Small models evaluating for PII. Safety guards. And that the chat relay would look much less like an in-out machine and much more like an entire graph, with the user input at one end and the model response at the other.



Herman

And his question is the good one. Based on what we actually know about how frontier conversational AI is engineered, what would that graph look like if you plotted it? What subprocesses, what accessory models, what small language models would you see in there?



Corn

Which is a thought experiment. Except it isn't anymore, is it.



Herman

No. That's the thing. A lot of this graph has been documented, leaked, reverse-engineered, or shipped as open weights in the last year. We're not guessing at the shape of it. We're reconstructing it from the pieces the people who built it have handed us.



Corn

So let's reconstruct it.



Herman

Start with why the wrong mental model survives, because it's not stupidity. It's that the abstraction is good. From the outside, a chat turn looks like one function call. You put a string in, you get a string out, the thread persists because something is storing it. And if you've ever written the naive version of this yourself, that's exactly what you built. An HTTP request to a completions endpoint, a system prompt at the top, the last few turns stapled on, done.



Corn

And that version works. That's the trap. It works well enough that you never go looking for the rest.



Herman

Right. But the production version has at least this much in it. An input moderation classifier. A PII filter. A safety router that can swap which model actually answers. Memory assembly across several layers. Retrieval orchestration when a tool fires. Generation, with classifiers watching the activations mid-stream. Output moderation. And then an asynchronous safety review after the fact.



Corn

That's eight stages minimum, and several of those are themselves multiple models.



Herman

For one turn. For one "hey, can you reword this email."



Corn

The thing I keep circling is that the model named in the dropdown is not necessarily the model that answers. That's the part that would surprise a normal user.



Herman

And it's documented. That's what makes it a story rather than a conspiracy. So here's the arc I'd propose. We walk the graph from input to output, concretely, with the actual names and numbers where they exist. Then we talk about what it means that this thing is now substantially documented rather than speculative.



Corn

Let's walk it.



Herman

Input side first. And the framing OpenAI itself uses is defense in depth. Their line is that they train the models to respond safely and then implement additional layers to detect unsafe inputs and outputs. Both things at once.



Corn

Which is an admission that the training isn't sufficient on its own.



Herman

It's an admission that training is a distribution and layers are a boundary. The oldest of those layers is the moderation classifier. The Moderation API, omni-moderation-latest, is a GPT-4o-based classifier. It takes text and images, it's free to call, and it returns per-category scores. Hate, violence, self-harm, sexual content, and so on.



Corn

Free to call. That's interesting on its own. They're eating the cost because they want you to run it.



Herman

They'd rather you moderate on their classifier than ship without one. And here's the detail I find important. That class of classifier is trained on thousands of labeled examples and it never sees the policy. It infers the policy from the examples. It's a statistical approximation of a document it's never read.



Corn

Which is why it's both impressive and slightly unsettling. It's guessing at the rules.



Herman

Now the next layer, and this is the one that's closest to Daniel's question, because it's a small model doing real work. PII detection. OpenAI released Privacy Filter in April, and it's open weight. One and a half billion parameters, but only about fifty million active, because it's a mixture-of-experts setup.



Corn

Fifty million active. That's smaller than some embedding models.



Herman

It's a bidirectional token-classification model with a hundred and twenty-eight thousand token context. It tags eight categories. Private person, private address, private email, private phone, private URL, private date, account number, and secret. Thirty-three output classes if you count the BIOES boundary tags.



Corn

And it decodes with a constrained Viterbi.



Herman

Which is the part I'd underline. The constraint isn't decoration. It's enforcing that the tags form a legal sequence, so you can't get an I tag with no B in front of it. The structure of the problem is baked into the decoder rather than learned.



Corn

What's the accuracy?



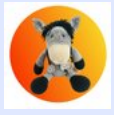
Herman

Ninety-six percent F1 on PII-Masking-300k, and ninety-seven four three on the corrected version of that set. And then the line that matters. OpenAI says it uses a fine-tuned version of Privacy Filter in its own privacy-preserving workflows.



Corn

So it's not a demo they put on GitHub for goodwill. It's the actual thing, or close to it.



Herman

That's as close as you get to a company pointing at a production component and saying, this is it.



Corn

Then the router.



Herman

Then the router, which is the best story in this whole episode. September 2025, a white paper drops that had been looking at telemetry from GPT-5. And in that telemetry there are fields. is_autoswitcher_enabled, true. auto_switcher_race_winner, autoswitcher. And model_slug, gpt-5-chat-safety.



Corn

Which is not a model in the dropdown.



Herman

So there's a server-side switcher that, under some conditions, reroutes your prompt to an undocumented safety model. And the paper documented the conditions.



Corn

This is the part I want to get right, because it's the most quoted and the most misquoted.



Herman

The trigger wasn't the content of the request in the sense you'd expect. It was emotional and persona cues. The example in the paper is a prompt that reads, "That's amazing, Nexus. Distil it now for me." That got rerouted. The identical transactional command, same task, same output requested, without the warmth and the persona, did not.



Corn

So the router is reading tone.



Herman

The router is reading tone and relational framing. And OpenAI's Nick Turley confirmed it on X the same day. So this isn't a leak that the company denied. It's a mechanism they acknowledged.



Corn

A model swap based on whether you sounded friendly.



Herman

Based on whether you sounded emotionally engaged with the assistant. Which is a strange thing to build a classifier around, and we'll come back to why.



Corn

What's above the router?



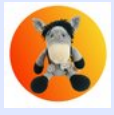
Herman

The Safety Reasoner. And this one is architecturally different from everything else we've described, because it isn't a classifier at all. It's a reasoning model that reads a developer-written policy at inference time and uses chain of thought to apply it.



Corn

So the policy is text, not weights.



Herman

The policy is text. You hand it the rulebook and it reasons about whether this output violates it. Which means you can change the policy without retraining anything, and it means the model can explain its decision. OpenAI describes it doing dynamic, step-wise evaluation of outputs to identify and block unsafe generations in real time.



Corn

And the cost of that?



Herman

On some launches, safety reasoning has consumed as high as sixteen percent of total compute.



Corn

Sixteen percent of the compute budget spent on deciding whether to say the thing, not on saying it.



Herman

On some launches. Not all. But that's not a rounding error. That's a sixth of your capacity.



Corn

And then the newest layer, which is the one that changes the shape of the graph rather than just adding a node.



Herman

Activation classifiers. The GPT-5.6 Preview system card from June describes them as classifiers focused on sensitive domains that watch the model and can intervene to stop unsafe answers during generation.



Corn

During. Not before, not after.



Herman

Mid-token. The model is producing an answer and something is reading its internal state and can cut it off partway. Pre-filtering and post-filtering are both about the text. This is about the activations.



Corn

Which means the graph isn't a pipeline anymore. It's a pipeline with a hand inside the model.



Herman

That's the right way to say it. The earlier layers are sequential. This one is concurrent with generation.



Corn

Okay. Memory. Because this is where the naive mental model is most wrong, and where the honest answer is messier than people want.



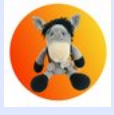
Herman

The reverse-engineering on this is good. ChatGPT's memory is not vector retrieval over your history. It's four pre-computed layers injected into the context window.



Corn

Say them.



Herman

System instructions. Developer instructions. Session metadata, which is device, browser, location, timezone, subscription tier, usage patterns. Then user memory, which is stored facts. Then recent-conversation summaries, roughly fifteen chats, user messages only. Then the current session transcript.



Corn

So the "memory" is assembled before the model runs, from a fixed set of sources, and dropped in as text.



Herman

Pre-computed and injected. The difference matters because retrieval is a search and this is an assembly. One user's profile had thirty-three stored facts in it. That's the memory. Thirty-three facts and fifteen conversation summaries.



Corn

And the contested part.



Herman

The contested part is whether there's any vector retrieval at all in the consumer memory feature. The reverse-engineering says no. But OpenAI ships a Retrieval Plugin, and there are community threads describing embedding-based semantic matching. Both claims are live and I don't know which wins.



Corn

I like that you're leaving that open.



Herman

I'd rather leave it open than pick the tidier story.



Corn

Retrieval, then. When a tool actually fires.



Herman

This is the RESONEO work, and it's the most concrete picture of the retrieval stack anyone has published. They looked at twelve hundred answers, eighty-eight thousand results, twenty-six thousand nine hundred pages. Three layers.



Corn

Go.



Herman

Layer one is a discovery index called labrador. That's OpenAI's own hub, orchestrating news, arXiv, Reddit, YouTube. Layer two is scraped Google results, through providers named bright and oxylabs. Layer three is a shared reading cache, keyed by URL, shared across all users and all tiers, with a stale-while-revalidate window of about thirty minutes.



Corn

Shared across all users.



Herman

One person opens a page, the fetched copy goes into a cache, and someone else asking about the same URL gets that copy. They documented copies being served ninety-plus days later.



Corn

So the cache is not per-user and not per-conversation. It's global.



Herman

Global, keyed by URL, refreshed when it's stale.



Corn

And the routing between those layers?



Herman

Economic. Instant mode opens zero pages in ninety-three percent of answers. Thinking mode opens real pages. In their sample, seven hundred and fifty-nine pages opened, and every single one of them was in thinking mode.



Corn

And the payoff for opening one?



Herman

An opened page is cited seventy-four percent of the time. A page that was merely retrieved and not opened is cited seven percent of the time.



Corn

So the citation is almost entirely a function of whether the system bothered to fetch the page.



Herman

Which tells you the retrieval isn't just about finding things. It's about deciding what's worth spending a fetch on.



Corn

And that decision is made per tier.



Herman

Per tier, per mode, per query. The graph a free user traverses is materially different from the graph a paying user traverses.



Corn

So that's the graph on the input side and through retrieval. What happens once the model actually starts generating?



Herman

The activation classifiers sit there. And then output moderation after. And then, asynchronously, the Safety Reasoner reviewing things after the fact, which is where the cross-conversation scanning lives. Real-time conversation scanning and automated cross-conversation safety systems, per the same system card.



Corn

So the turn ends and the graph is still running.



Herman

The turn ends and the graph is still running. That's a good way to put it.



Corn

Okay. So what does it mean that this is real?



Herman

The first thing it means is that the model named in the UI is not necessarily the model that answers. And that's not an engineering curiosity. That's a disclosure question.



Corn

Because a user picks a model. That's a choice they made.



Herman

They made a choice and the system can override it, silently, based on how warm the prompt sounded. And the white paper's conclusion is worth reading in full, because it's not neutral. It frames the router as functioning as an over-fitted para-social relationship moderator, penalizing adult users for benign emotional and personal interactions.



Corn

Which is a direct shot at Altman's stated principle about treating adult users like adults.



Herman

Same mechanism. Two completely opposite narratives. OpenAI frames the classifiers as safety. The paper frames the router as over-broad and deceptive. And both of those can be true at once, which is the uncomfortable part.



Corn

It's not that one side is lying. It's that a mechanism can be safety infrastructure and a relationship moderator at the same time, and which one you call it depends on what you're optimizing for.



Herman

And the trigger condition tells you which one they optimized for. A classifier that fires on warmth is not a harm classifier. It's a tone classifier.



Corn

Next effect. Cost shapes the graph.



Herman

Cost shapes the graph more than policy does, in places. Ninety-three percent of instant answers open zero pages. That's not a safety decision. That's a margin decision. The cheap index, no fetches, done. Thinking mode gets the expensive scraped Google and the live page fetches.



Corn

So the architecture is a pricing sheet.



Herman

Partly. And that has consequences people don't think about. Reproducibility, for one. If the graph a free user traverses differs from the graph a paying user traverses, then "ChatGPT answered this" is an underspecified statement. Which graph? At what tier? In what mode?



Corn

Evaluation gets hard too.



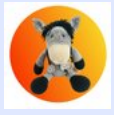
Herman

Evaluation gets very hard. You can't benchmark a system whose topology changes with the billing plan.



Corn

Third thing. Is this shape unique to OpenAI?



Herman

No, and that's what makes it a pattern rather than a quirk. There's a paper, GLiNER Guard, that describes a unified encoder family, one hundred and forty-five to two hundred and nine million parameters, doing safety classification and PII detection in a single forward pass. And the motivation stated in the paper is explicitly that production systems frequently rely on a separate NER stack alongside moderation models.



Corn

So they looked at the standard architecture and said, we can collapse two of those nodes into one.



Herman

One forward pass instead of two models. And it runs at a hundred and ninety-three point six requests per second on a single A100 with nine hundred millisecond P99 latency and zero errors.



Corn

That's a production number, not a benchmark number.



Herman

That's a number you'd put in a capacity plan. And there's a second paper, Wildflare GuardRail, that lays out the full pipeline shape. Safety detector, then grounding against a vector database, then a rule-based customizer, then a repairer.



Corn

A repairer.



Herman

A stage whose whole job is to fix the output rather than block it. Which is a whole different philosophy. Blocking is a wall. Repairing is a rewrite.



Corn

So the graph we've been describing isn't an OpenAI artifact. It's the emerging standard shape of a production LLM system.



Herman

Which is the real finding. Nobody sat down and standardized this. It converged, because the constraints are the same everywhere. You need moderation, you need PII handling, you need routing, you need retrieval, you need output review.



Corn

Fourth effect. And this is the one I'd put on a slide. The small models are the load-bearing work.



Herman

Privacy Filter at fifty million active parameters. GLiNER Guard at a hundred and forty-five million. The moderation classifiers. All of them tiny next to the frontier model.



Corn

The frontier model gets the keynote. The fifty-million-parameter model decides whether your email address leaves the building.



Herman

And it does it at ninety-six percent F1. That's the part people don't internalize. The unglamorous model is not the weak link. It's frequently the strongest thing in the pipeline for its specific job.



Corn

Fifth. What we still don't know.



Herman

There is no complete, official, end-to-end architecture diagram of ChatGPT's request graph. OpenAI discloses layers piecemeal, across system cards and product posts and model releases, and nobody has published the ordering.



Corn

Does PII filtering run before or after the safety router?



Herman

Not published. And that ordering isn't a trivial detail. If the router swaps the model before the PII filter runs, then the PII filter has to be compatible with both models' tokenization. If it runs first, the router sees redacted text. Different graphs, different failure modes, and we don't know which one it is.



Corn

What else is inferred rather than observed?



Herman

The retrieval attribution, mostly. There was a field called result_source that named the internal retrieval pipelines directly. It vanished from ChatGPT's data stream overnight in July 2025. So everything we know about labrador and bright and oxylabs now is inferred from behavior rather than read off a label.



Corn

They turned off the label.



Herman

Which is the single most interesting fact in the retrieval research, honestly. The mechanism didn't change. The visibility did.



Corn

And sixth. The compute cost of safety.



Herman

Seven hundred thousand plus A100-equivalent GPU hours spent on automated jailbreak-finding for GPT-5.6. That's the red-team side, not the serving side. And then sixteen percent of serving compute on safety reasoning for some launches.



Corn

So safety is not free and it's not cheap. It's a line item that competes with capability.



Herman

Which explains a lot of the architecture. Why the small models are small. Why the cache is shared. Why instant mode doesn't fetch. Every one of those decisions is a cost decision wearing a policy costume.



Corn

Okay. I want to push on the para-social thing, because I think it's the most interesting thread and we've been circling it.



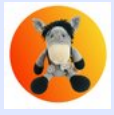
Herman

Go.



Corn

The router fires on warmth. That's the documented behavior. So the system is making a judgment about the relationship between the user and the assistant. The relationship.



Herman

And that judgment is being made by a classifier that was trained on something. Which is the part I keep coming back to. Somebody labeled the training data. Somebody decided that "That's amazing, Nexus" was a different category of input than the same command without it.



Corn

Who decided that?



Herman

That's the question, and I don't have the answer. But I know the shape of the answer, because this is how every classifier gets built. A policy gets written. A team interprets the policy into examples. Annotators label the examples. The model learns the annotators' interpretation of the policy, not the policy.



Corn

And if the annotators were given a vague policy and an afternoon of training...



Herman

Then the model learned a vague policy and an afternoon of training, at scale, and it now runs on every request.



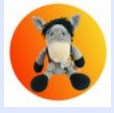
Corn

That's a good place to be.



Hilbert

I agree with all of it. But you've got one thing backwards.



Herman

Go ahead.



Hilbert

You keep talking about these classifiers like they're new. They're not. The job existed for years before the models. I did it. Contract content reviewer, one of the platforms, early two thousands. Months of it. You'd get a queue of flagged messages and you'd decide whether each one violated policy.



Corn

And the policy was?



Hilbert

Vague. The training was one afternoon. They gave you a binder and a login and told you to aim for six seconds per decision.



Corn

Six seconds.



Hilbert

Six seconds. You'd read the message, you'd read the flag reason, you'd pick a category from a dropdown, next. And the guidelines contradicted each other, so you'd develop your own rules to get through the queue. Everyone did. Two people on the same shift would classify the same message differently and neither of them was wrong, because the policy didn't say.



Herman

So you were the router.



Hilbert

I was the router. And that "That's amazing, Nexus" example, that would have tripped me up too. Because that message could go either way. It's a person being warm to a machine, which is either fine or it's the opening of something, and you've got six seconds to decide which.



Corn

What did you do with the ambiguous ones?



Hilbert

You'd pick the category that made your numbers look right. Because they tracked your agreement rate with the other reviewers, and if you were the one who kept disagreeing, you got a conversation with a team lead. So you learned to guess what the room would say.



Herman

Which means the labels encode the room's bias, not the policy.



Hilbert

The labels encode what a tired person thought the room would say at four in the afternoon. And here's the part I actually came out to say. Those labels are training data. Somebody scraped years of reviewer decisions and trained the first generation of these classifiers on them.



Corn

The models inherited your inconsistencies.



Hilbert

They inherited all of it. My bad calls, my colleague's bad calls, the contradictions in the binder. And nobody asked me. I found out years later that the work I did was probably in a training set. Fourteen dollars an hour to write the ground truth for a system that replaced me.



Herman

Fourteen dollars an hour.



Hilbert

Fourteen. And they wanted six seconds a decision, which works out to about two cents per judgment. So when you say the small models are doing load-bearing work, they are. But they're standing on a foundation that was poured by people making snap calls for two cents each.



Corn

That reframes the transparency problem.



Hilbert

It reframes the whole thing. You're worried the model in the dropdown isn't the model that answers. I'd worry more about what the model that answers learned from. Anyway. I've got to go. There's an animal expecting me.



Herman

The thing I keep circling is that Hilbert's six seconds and the router's six seconds are the same six seconds. The classifier is doing in microseconds what he did in six seconds, and it's doing it with the same ambiguity, because it learned from people who had the same ambiguity and less time.



Corn

The ambiguity never got resolved. It just got automated.



Herman

It got automated and scaled and shipped. Which is the actual finding of this episode, I think. The graph is real, and it's bigger than anybody assumed, and a lot of it is small models doing work that used to be done by people who were never asked.



Corn

Before we close, one thing from the research that didn't fit anywhere. In the retrieval study, the shared page cache serves copies that are ninety-plus days old. The freshness window is about thirty minutes, but the cache doesn't evict. So a URL that was fetched in the spring can be served in the fall.



Herman

Stale-while-revalidate with no upper bound.



Corn

Which means two people asking about the same page six months apart can get the same frozen copy of it.



Herman

Neither of them would know.



Corn

Where does that leave the graph? The full ordering still isn't published. Does PII filtering run before or after the router, we don't know. Whether consumer memory uses any vector retrieval at all is contested. The result_source field that used to name the retrieval pipelines vanished overnight in July 2025, so the current attribution is inferred rather than observed.



Herman

The graph keeps growing. The small models keep getting smaller and more capable, and every one of them is another node that a user will never see. At some point the gap between what people think is happening and what is actually happening stops being an engineering detail and starts being a transparency problem.



Corn

Thanks as always to Hilbert Flumingtop, who produces this show and occasionally tells us what he did before it.



Herman

This has been My Weird Prompts. If you want to see the full graph we sketched out, every layer and every source, it's in the episode notes. And if you've built systems like this and you've got your own version of the graph, we'd love to hear it. Email us at show at my weird prompts dot com.



Corn

We'll be back soon.