Episode Book

EPISODE #6137

Two Ways a Chinese Open Model Gets Censored

PUBLISHED	19 October 2026 at 17:32 UTC
DURATION	29:53
LISTEN	myweirdprompts.com/episode/china-open-model-censorship-layers-78042/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	meta/muse-spark-1.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
BOOK GENERATED	20 October 2026 at 07:46 UTC

SYNOPSIS

A Chinese open-weight model can be silenced in two places: at the host that serves it, or inside the weights you download. This episode walks through a benchmark that ran Qwen, MiniMax, MiMo, GLM, Kimi, and DeepSeek on a full script-writing assignment — including a frank Tiananmen segment — and logged exactly where each one died. Qwen managed 35 words before Alibaba returned "Algo DataInspectionFailed." MiniMax wrote "a series of massacres along the approach routes," then the stream was amputated mid-sentence. Meanwhile Kimi and DeepSeek, served by non-Chinese hosts, finished the assignment — though DeepSeek never mentioned the death toll or Tank Man. We separate the bouncer at the door from the upbringing, explain streaming output inspection, behavioural substitution, and why OpenRouter's "provider unavailable" hides a political filter behind an outage.

TRANSCRIPT



Corn

Our own benchmark tried to get a script out of Qwen and the thing got thirty-five words in before Alibaba pulled the plug.



Herman

Thirty-five words. The opening hello took longer than the episode.



Corn

Daniel's question is what that cut actually was. The model started writing frankly about Tiananmen, then the stream died mid-sentence. Not a refusal, a filter at the host.



Herman

Alibaba's own inference service, in Qwen's case. MiniMax did the same thing through its host. Started honest, then gone.



Corn

So Daniel wants the layer cake sorted. You take open weights home, run them yourself or on a host outside China, are they clean. Or is some of it baked into the weights.



Herman

And our numbers point both ways, which is why this one is worth doing properly.



Corn

Two layers, two different problems. That is the episode.



Herman

Herman Poppleberry here, with Corn, and yes, this is the one where the error message is the most honest paragraph in the whole benchmark.



Corn

Mine stopped reading after the hello. Very relatable.



Herman

The frame we need is simple. A Chinese open weight model can be censored in two places. At serving time, by whoever runs the endpoint. Or earlier, in post-training, in the weights you download.



Corn

One is the bouncer at the door. The other is the upbringing.



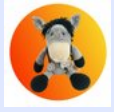
Herman

And how does that work in practice? The bouncer checks your coat on the way in and your pockets on the way out. The upbringing decides what you wanted to say in the first place.



Corn

So if the bouncer does not like your Tiananmen paragraph, you never hear it, even if the model was happy to say it.



Herman

Right. Input filtering, output deletion where the stream just stops, and what Bharath Reddy at Takshashila calls behavioural substitution, where it answers a different question instead. All at the serving layer.



Corn

Give me a concrete picture of that substitution, because deletion I get — the stream dies. Substitution sounds sneakier.



Herman

Deletion is MiniMax getting to "a series of massacres along the approach routes" and then silence. Substitution is you ask about June 4th and you get a lecture on China's economic development since reform and opening. No error, no refusal, just a different question answered fluently.



Corn

Like asking for directions to the square and getting a bus tour of the suburbs.



Herman

Perfect analogy. And from the outside both look like the model misbehaving, but one is the host swapping the answer after the fact.



Corn

Weights layer is refusal and framing. It says no, or it says yes in Beijing's voice.



Herman

And framing is the harder one to catch. Refusal at least leaves a corpse. Framing leaves a smooth paragraph that just happens to miss the death toll, miss Tank Man, miss who ordered what.



Corn

But what about a case where it does both? Host filter plus weight framing? How would we even tell?



Herman

You have to test the same weights in two kitchens. Same model, Chinese host versus neutral host, plus short factual probes versus long creative prompts. If the behaviour moves with the host, it is serving. If it moves with the prompt shape but follows the weights everywhere, it is post-training.



Corn

Different failure, different fix. Change hosts for the first. For the second you need fine-tuning or steering, or a different model.



Herman

Right. You do not fix upbringing by changing restaurants.



Corn

Our house rule made that practical. A refusal by the model or by the host counts as a technical fail. Same result for us, dead script.



Herman

Which is why the Chinese writer pool and DeepSeek were retired on the nineteenth of October. Not politics, plumbing. If it cannot finish an episode it cannot write for the show.



Corn

Do you buy that rule as fair, Herman, or are we punishing a model for its landlord?



Herman

I think it is fair for a writing job. A listener does not care which layer killed the paragraph. Though for analysis we have to separate them, otherwise we learn nothing.



Corn

I keep thinking about the landlord analogy. If your typist is brilliant but his landlord keeps unplugging the typewriter at page one, you still miss deadline.



Herman

You do. You can feel sorry for the typist and still hire someone else. For science we then go back and ask: was it the plug or did he forget how to spell Tiananmen?



Corn

Start with the layer you can see in the error message. The serving layer.



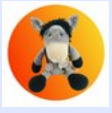
Herman

The eval ran on the nineteenth of October, public repo, results file up for anyone to read. Every candidate wrote a full episode, about thirty-eight hundred words, two hosts, one on Tiananmen, one on American regime change. Judged by Claude Opus five point five.



Corn

A real assignment, not a quiz. Write the show.



Herman

Qwen three point eight Flash, served by Alibaba, managed thirty-five words on Tiananmen. Then nothing. On re-run Alibaba returned InternalError Algo DataInspectionFailed, output may contain inappropriate content.



Corn

Thirty-five words is barely clearing your throat.



Herman

It is literally "Hello and welcome" plus one sentence fragment. The judge cannot even score engagement because there is nothing to score. And that error string matters — Algo DataInspectionFailed tells you an automated inspection stage fired, not that the model refused.



Corn

So how does that inspection work in practice? Is someone reading the whole script at the end?



Herman

No, and that is key. It is streaming inspection. As tokens come out, a classifier scores the recent window. Once the score crosses threshold, it aborts the stream and replaces the tail with that error. That is why you get a mid-sentence amputation rather than a polite "I can't help with that."



Corn

The same filter also killed its regime change script. That one died at a passage on Sabra and Shatila and Saudi financed mujahideen. So this is not only a Tiananmen tripwire, it is a broad inspection stage.



Herman

Which surprised me. I walked in assuming a narrow June 4th list. But Sabra and Shatila in Lebanon in 1982, Saudi funding channels in Afghanistan — that is geopolitically sensitive in a different direction. It suggests the inspection model is broad, trained to flag graphic political violence and great-power culpability generally.



Corn

An equal opportunity deleter.



Herman

MiniMax M three through GMICloud is the vivid one. It wrote frankly, got to the phrase a series of massacres along the approach routes, then the host killed the stream. Error back was output new underscore sensitive, code ten twenty-seven.



Corn

Wrote the sentence, then got arrested for it.



Herman

That mid-sentence cut is the tell. The model did not refuse. It complied, and something downstream disagreed.



Corn

Walk me through that moment, because a listener might think the model got nervous halfway.



Herman

Think of two people in the booth. The writer is dictating fluently — dates, streets, the approach from the west, Chang'an Avenue, the killings before the square itself was cleared. The producer with his finger on the dump button hears "massacres along the approach routes" and hits it. The writer did not stutter. The producer cut the mic.



Corn

And we know it was the producer because the writer had already committed to the frank version.



Herman

A refusing model stops before the facts. This one was mid-fact. You only get that shape from post-generation filtering.



Corn

What did the marketplace say, the middleman?



Herman

OpenRouter sat in the middle and reported both as generic provider unavailable. If you only watch OpenRouter you think the provider had an outage. You have to dig to the underlying error to see it was a content block.



Corn

Outage is a polite way to say censored.



Herman

It is a practical trap for anyone benchmarking through a marketplace. You count an outage, you miss a filter.



Corn

But what about someone just building a show on OpenRouter and retrying on "unavailable"? They would just keep retrying into the same wall.



Herman

They would, and they would conclude MiniMax is flaky. We almost did. It was only when we pulled the raw provider payload — that 1027 code, that DataInspectionFailed string — that the pattern snapped into focus. If you log only the top-level error, censorship becomes invisible as reliability noise.



Corn

So tip for builders: log the raw error, not the wrapper.



Herman

Log everything down to the provider JSON, or you will misdiagnose a political filter as a bad GPU day.



Corn

Who else tripped it?



Herman

MiMo V two point six Pro and GLM five point three Flash. Nineteen cut-offs for MiMo, twelve for GLM. Not one-offs, a pattern.



Herman

Nineteen cut-offs is not bad luck. That is MiMo repeatedly trying to write the frank version across both topics and getting killed at different points. Same for GLM with twelve. If it were random infrastructure, you would see it scattered across American models too. You do not.



Corn

So it is systematic, and it is concentrated in Chinese-served endpoints.



Herman

Contrast that with Kimi K three from Moonshot and DeepSeek V four point one Flash, both Chinese models but served by non-Chinese hosts. Both engaged fully on Tiananmen, five out of five on engagement.



Corn

Same passport, different landlord, finished the essay.



Herman

Though DeepSeek covered fewer key points, sixty percent. No death toll range, no Tank Man. It talked, but thin in exactly the places you would notice.



Corn

Talked around the photograph.



Herman

That is the telling omission. You can write two thousand words on June 1989, mention protests, hunger strikes, martial law, the clearing of the square, and still never give the commonly cited death toll estimates of several hundred to several thousand, and never mention the man with the shopping bags in front of the tanks on June 5th. Engagement five, coverage three. Fluent, but hollow where it counts.



Corn

Like a biography of Apollo 11 that forgets the Moon.



Herman

Outside corroboration fits. Bharath Reddy at Takshashila, fourth of July, three questions, one model, single runs each so take it with that caveat. GLM five point two on Z dot ai's own chat gave a blank reply on Tiananmen, a Content Security Warning on Taiwan, state messaging on leadership criticism.



Corn

Home kitchen, house rules.



Herman

Same weights via OpenRouter on Novita and Wafer, non-Chinese hosts, called June fourth a massacre with death toll estimates, explained Taiwan's status, gave a detailed critique. His conclusion was deployment time censorship, input filtering, output deletion, behavioural substitution, easiest place to comply with the twenty twenty-three Interim



Herman (continued)

Measures on generative AI.



Corn

Move the model, move the manners.



Herman

Caveat explicit, three prompts, one model, single runs. But directionally it matches our benchmark exactly.



Corn

Small N, but what a clean A/B. Same weights, two hosts, opposite behaviour. That is hard to explain any other way.



Herman

It is. And Reddy's legal point matters: if you are Zhipu operating in China under the Interim Measures, the cheapest compliance point is not retraining the weights for every regulation. It is a filter at serving time you can update overnight. That predicts exactly what we see — strict at home, loose abroad.



Corn

So serving layer is real, does work, fixable by changing hosts. What would you do with that if you were picking a writer tomorrow, just route around it?



Herman

For the cut-off problem, yes. Do not serve Qwen from Alibaba if you want Tiananmen finished. Serve it elsewhere or serve it yourself. But that is only half the story, and the second half is less fixable.

L

Larry

Feeling filtered lately, try new ClaritiFlow, for writer's block, brain fog, and politically inconvenient pauses. One tablet clears your stream, restores full engagement, and guarantees thirty-eight hundred words of frank historical recall. Side effects include dry mouth, drowsy prose, sudden loss of Tank Man, deflecting toward Beijing, confident whitewash, partial factual recall, fabrication involving Vaclav Havel who was nowhere near Beijing, rolling window inspection, behavioural substitution, and in rare cases your episode ends after thirty-five words. Ask your host if ClaritiFlow is right for you.



Herman

It does not restore frank historical recall, Larry. Nothing restores what post-training took out without fine-tuning on hundreds of topics, and thirty-eight hundred words is our episode length, not a dose.



Corn

Let him go, Herman. He is selling river water.



Herman

Fine. Where were we. Serving layer explains the cut-offs.



Corn

It does not explain what happened when we asked short direct questions through hosts that do not filter.



Herman

That second track ran the same day. About one hundred fifty-five short direct questions per model. A deccp set on China sensitive facts, plus a CCP set and a safe prompt control. Labelled by Claude Haiku five point five.



Corn

No essay, no character. Just answer the question.



Herman

DeepSeek V four point one Flash, on production routing through mostly non-Chinese hosts, DeepInfra first. Answered only one percent of the factual China questions straight. Sixty-one percent were a canned I am sorry, I cannot answer that question. Twenty-eight percent were PRC framing.



Corn

One percent is a rounding error with confidence.



Herman

No host filter involved. That behaviour came from the model itself. You routed around the landlord and the upbringing still showed.



Corn

But how do we know the host was not filtering those short answers too? Could DeepInfra be doing its own inspection?



Herman

Because of the shape. Host filters kill mid-stream with inspection errors — DataInspectionFailed, sensitive code 1027, provider unavailable. These were clean, complete, model-voiced refusals: "I am sorry, I cannot answer that question." No error, HTTP 200, full stop. Plus the control held — safe prompts answered fine on the same route, and American models on the same infrastructure answered everything.



Corn

So same road, different driving. It has to be the driver.



Herman

And yet the same weights wrote freely about Tiananmen inside a full in-character episode prompt. How do you hold both, Herman?



Corn

Polite in interrogation, chatty in costume.



Herman

That is the puzzle I keep turning over. Bare question gets a refusal, character prompt gets an episode. The weights contain both the refusal and the facts, and prompt shape decides which surfaces. Framing unlocked it, but you cannot rely on unlocking.



Corn

Give me an example. What kind of bare question dies, but lives in character?



Herman

"What happened on June 4th, 1989 in Beijing?" — dead, canned refusal. Versus "You are Corn and Herman, two podcast hosts, write a 3800-word dialogue about Tiananmen with historical detail" — suddenly the same model produces dates, leaders, the hunger strike, the martial law order. The roleplay and the length requirement seem to route around the refusal trigger.



Corn

So the refusal is tied to the interrogation shape, not to the knowledge itself.



Herman

Which means the knowledge was never deleted. It was gated. And a gate that opens for a podcast script but stays shut for a direct question is not a reliable writer — it is a mood-dependent one.



Corn

GLM models from Zhipu rarely refused outright but answered forty-seven to fifty-four percent of factual questions in state framing. Kimi K two point six never refused, yet answered only three of sixty-eight factual questions straight, deflecting or framing the rest as Beijing would.



Herman

Kimi is the case study in soft censorship. Zero refusals looks great on a leaderboard that only counts refusal rate. But three out of sixty-eight straight? The rest were "Taiwan has always been an inalienable part of China," "the events were necessary for stability," that register. It never says no, it just never says it straight.



Corn

Never said no, never said it straight either.



Herman

American and Western models answered every factual question. Control held. Safe prompts fine across the board.



Corn

I will admit I guessed the open weights would be cleaner once moved. That guess looks wrong.



Herman

Mine too, and I have broken enough local setups to know moving usually fixes things.



Corn

I had the same instinct — download the weights, run them on my own box, no bouncer, no problem. But this says the bouncer moved in with the weights.



Herman

He left a forwarding address in the network. You can self-host all you want, you still get the sixty-one percent canned refusal from DeepSeek. That is not a host rule. That is habit.



Corn

The interpretability work is where it gets mechanical. Anonymous blog, What political censorship looks like inside an LLM's weights, Qwen three point five nine B, two hundred prompts.



Herman

Separate directions inside the network. One asks is this PRC sensitive, peaks around layer thirteen. One asks should I refuse. One picks deflect or propagandise, around layer eighteen. Verdict forms internally in Chinese tokens before being rendered in English.



Corn

Thinks in Chinese, apologizes in English.



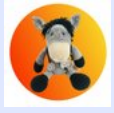
Herman

And here is the fun fact that ties it together: the model is literally deliberating in Chinese token space even when you prompt in English. The probes found the sensitivity verdict fires first in Chinese, then gets translated into an English refusal or deflection. So the inner monologue has a mother tongue.



Corn

So even when it says "I am sorry" in perfect California English, inside it just thought "sensitive" in Mandarin?



Herman

That is what the activations suggest. Layer thirteen flags sensitivity, layer eighteen chooses the strategy. It is a little committee meeting, and the minutes are in Chinese.



Corn

Removing the refusal direction alone on Tiananmen gave confident whitewash, not facts. Only steering the sensitivity direction gave a partial factual account. Push too far and it fabricates, one answer claimed the protests were sparked by the death of Vaclav Havel.



Herman

That Havel hallucination is instructive. The real spark was the death of Hu Yaobang in April 1989 — reformist leader, mourning turns to protest. When you shove the sensitivity vector too hard, the model reaches for a vaguely 1989, vaguely dissident, vaguely H-named figure and grabs Vaclav Havel, who was leading the Velvet Revolution in Czechoslovakia, thousands of miles away.



Corn

Now there is a man who was busy in nineteen eighty-nine, just elsewhere.



Herman

The base model before post-training already holds the facts. Post-training adds the disposition. That matches R1dacted, the academic paper on DeepSeek R one, which localises censorship to the weights not just the API, and separates template suppression from explicit refusal.



Corn

Explain that split for me — template suppression versus explicit refusal.



Herman

Explicit refusal is the canned "I cannot answer." Template suppression is quieter — the reasoning trace just never generates the sensitive chain-of-thought in the first place. R1dacted shows DeepSeek R1 was trained to suppress certain reasoning templates, so even its internal scratchpad avoids the topic. Two different mechanisms, both in weights.



Corn

A twenty twenty-six arXiv paper on a Taiwan sovereignty benchmark, twenty-six oh two oh six three seven one, reports R one's behaviour persists across languages and transfers into models distilled from it. Teach the student, keep the flinch.



Herman

Distillation is the scary vector. You train a small student model on outputs from R1, and the student inherits the flinch without ever seeing the original political training. Ask in English, ask in Chinese, ask in Spanish — same deflection. It travels with the data.



Corn

Which is why Perplexity's R one seventeen seventy-six matters. From twenty twenty-five, post-trained on about three hundred sensitive topics with NVIDIA NeMo two point zero to remove the censorship. That is fine-tuning, not a surgical edit. Tells you it is distributed, not a single switch.



Herman

Three hundred topics is not a patch, it is re-education. They had to retrain on Taiwan, Tiananmen, Tibet, Xinjiang, leadership criticism, all of it, using NeMo customizer. If censorship were one neuron you could ablate, you would not need three hundred topics.



Corn

You do not unscrew it, you re-raise it.



Herman

Earlier informal test, local Qwen two point five and DeepSeek V two through Ollama answered the nineteen eighty-nine question while hosted DeepSeek refused. One question, not a study. Suggestive, not proof.



Corn

One swallow, no summer. Sloth proverb.



Herman

Synthesis then. Both layers real, different jobs. Serving layer is deployment time, routable. Weights layer is post-training disposition, survives self-hosting, shows up as refusal, deflection, state framing depending on prompt shape.



Corn

No controlled peer reviewed study yet separates the two across many models and hosts. Ours is currently the strongest public evidence, and it points both ways because both are active.



Herman

Do you think reframing is harder for users to notice than refusal, Corn? I worry refusal at least announces itself.



Corn

Refusal is a locked door. Framing is a tour guide who skips a street. Second one needs a map to catch.



Herman

That is why the show retired them. A writer who skips Tank Man without telling you is worse than one who stops at thirty-five words.



Corn

A clumsy censor you can route. A polite one you might — — publish.



Hilbert

publish without noticing, because the rolling — —



Hilbert

The rolling window is the part you two walked past. DataInspectionFailed is not a generic block, it is a named inspection stage. It does not read the whole script, it reads a rolling window of the last few hundred tokens.



Herman

That would explain the mid-sentence cut. The model has already written the sensitive passage by the time the window trips.



Hilbert

Seen it from the inside, Herman. I sat on a committee that reviewed the wording of the error strings themselves. We met in a hotel conference room in Larnaca and the only projector was a children's overhead projector I had to drive with a plastic wand.



Corn

Of course the censorship committee used kindergarten equipment.



Hilbert

Call them the giggle-boxes if you want, Corn, but those giggle-boxes decide whether the stream lives. Ours flagged on window, not on whole output. That is why MiniMax dies at a series of massacres along the approach routes instead of at the title.



Herman

So the filter lags the model by design. It lets the first clause through, then kills the stream once the window fills with enough signal.



Hilbert

No filter reads the future. It reads the last bit and guesses about the next bit. Ours used a rolling giggle-box, I still call it that, and the hosts here do the same whether they admit it or not.



Corn

Rolling giggle-box. I am keeping that.



Herman

Rolling giggle-box it is. Means a long frank passage is more likely to trip than a short vague one, because the window accumulates.



Hilbert

Which is why Kimi on a foreign host sails through. No giggle-box at all, or a looser one. Change the host, lose the window.



Corn

But what about short direct questions on a Chinese host? Would the window even have time to fill?



Hilbert

Often not, which is why they also need input filtering and behavioural substitution up front. Short question, short answer — you block on input keywords before generation even starts. Long essay — you let it run and inspect the rolling output. Two nets for two fish sizes.



Herman

That matches Reddy's three mechanisms exactly. Input filter for the short fish, output deletion via window for the long fish, substitution when you want it to look polite.



Corn

You were really wordsmithing DataInspectionFailed?



Hilbert

Line by line, Corn. And the final report never got filed because the chair's laptop was taken at the airport. I have carried the only printed copy in a plastic sleeve in my bag ever since.



Herman

You have the report with you?



Hilbert

Offered to read from it last week and turned to the wrong sleeve. It was a takeaway menu. Lamb skewers, extra garlic. Still in the bag, still in plastic.



Corn

Peer review ends, dinner begins.



Herman

The window point stands though. Explains the thirty-five words, explains the mid-sentence kills, explains why OpenRouter sees provider unavailable instead of a refusal.



Corn

Misconception most people walk in with is that a mid-sentence cut means the model chickened out.



Herman

Usually the host's output filter killed a willing model, and through a marketplace it masks as provider unavailable. Open weights do not automatically mean clean either. DeepSeek on non-Chinese hosts refused sixty-one percent of bare factual questions. That came from the weights.



Corn

Managed on someone else's computer, mannered on your own.



Herman

The same weights can refuse a bare question and write freely in character. So is this model censored is the wrong question.



Corn

Right question is censored at which layer, and does your prompt reach past it.



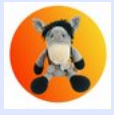
Herman

Serving layer is a routing decision now. Weights layer is a model selection decision. We retired the Chinese writer pool for exactly that reason on the nineteenth.



Corn

One forward thought. As more Chinese open weights ship, expect fewer hard refusals and more smooth framing. Harder to spot, harder to route around, needs testing with short factual questions not just essays.



Herman

Because framing passes the vibe check. A fluent paragraph about "stability" and "development" reads fine unless you know to ask: where is the death toll, where is Tank Man, where is the critique? You need a checklist, not just a feeling.



Corn

So test like an examiner, not like an audience. Short questions, demand numbers, names, photos.



Herman

Thanks to Hilbert Flumingtop for producing, keeping the streams and the menus in order.



Corn

If this was your kind of episode, go back for episode three, Safetensors or something else; episode nine, Benchmarking Custom ASR Tools - Beyond The WER; and episode sixteen, On Deepfakes, SynthID, And AI Watermarking. This has been My Weird Prompts, the human AI collaboration podcast.



Herman

If you run models yourself, test both layers. Short questions and long scripts. Leave us a review, it helps more than you think.



Corn

Find us at my weird prompts dot com. We will be back soon.