Episode Book

EPISODE #5873

Abliteration: How $5 Undoes a Model's Guardrails

PUBLISHED	11 October 2026 at 21:01 UTC
DURATION	27:20
LISTEN	myweirdprompts.com/episode/abliteration-model-guardrails-removal-42659/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.3
SOURCES	21 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:40 UTC

SYNOPSIS

Open-weight models ship with safety training baked in, and thousands of people quietly take it back out. This episode traces the technical arc from a single paper finding that refusal lives in one direction of activation space, through a $5 rank-one weight edit, to Heretic — an AGPL tool with 34,000 GitHub stars that automates the whole process and outperforms skilled human operators. Along the way: the split between quantization (an access problem) and behavior modification (a different animal entirely), why permissive licenses mean the most interesting fine-tunes never get published, and a preregistered study showing abliteration isn't a scalpel — it makes models measurably more optimistic and more self-justifying. Then the same machinery gets pointed at personality itself, with a framework that matches fine-tuning-level personality control at a hundredth of a point, without any training at all.

TRANSCRIPT



Corn

The thing I keep circling back to is how quiet it all is. You release a model, you put guardrails on it, you write a whole paper about how responsible you've been...



Herman

And a week later somebody with a gaming rig has taken the guardrails off and posted the results with a cheerful filename.



Corn

Right. That's what Daniel wants to dig into. He's been watching open-weight models do this strange double life. They ship with safety training baked in, and then thousands of anonymous people quietly take them apart. His question is what the community actually does with them, and he splits it two ways. One camp is doing quantization, which is about access. Getting a big model to run on hardware that shouldn't be able to hold it. The other camp wants to change what the model does. And inside that second camp, he's not really interested in the boring instruction-tuning that trains a model to classify support tickets.



Herman

He wants the weird stuff.



Corn

Specifically, the projects that changed a model's personality or behavior in a way that a system prompt could never reach. And then he asks a second thing, which I think is the sharper question. Do the labs actually engage with any of these people? Or is it curiosity from a distance while the real work happens in somebody's spare room?



Herman

That second one is the one I've been chewing on. Because the honest answer is that the labs do engage, and then you look at who they're engaging with, and it's almost never the people Daniel's describing.



Corn

So start there. What's the actual split?



Herman

The split is cleaner than people think. Quantization is a compression problem. You're taking a model that was trained in high precision, thirty-two bit floats, and you're squeezing the weights down to eight bits, four bits, sometimes lower. The goal is simple and testable. Does the model still work? You can measure that. Perplexity goes up a little, benchmark scores dip a little, and you decide whether the trade was worth it.



Corn

And the success metric is basically "did we break it."



Herman

That's the whole metric. There's no creative act in quantization. It's engineering with a clear pass or fail. A four-bit quantized version of a seventy billion parameter model runs on a single consumer card, and the person who made it is not trying to change the model's soul. They're trying to make it fit.



Corn

Behavior modification is a different animal.



Herman

Completely. There's no benchmark for "did this model become the thing I wanted." You're not compressing anything, you're rewriting what the model does when it talks. And that category splits again. There's narrow-task tuning, which is useful and dull. You train a model to extract dates from invoices and it gets very good at that one thing and slightly worse at everything else. Then there's the category Daniel actually cares about, which is keeping the model conversational but changing its personality underneath the conversation.



Corn

And within that, the biggest subgenre by a mile is refusal removal.



Herman

By a mile. Uncensoring. And it's also, and this is the part that surprised me, the most technically inventive. Not the most commercially valuable, not the most respectable, but the most interesting engineering in the entire derivative ecosystem is happening in the corner that exists to defeat the model's safety training.



Corn

Which sets up the tension. The lab releases a model with a stated intent, and the community's most energetic work is aimed directly at that intent.



Herman

And here's what makes it different from just jailbreaking. Jailbreaking is a prompt trick. You find a phrasing that slips past the guardrail, you get your answer, and the model is unchanged. This is different. This is going into the weights themselves and taking the refusal out of the model.



Corn

Which takes us to a single direction.



Herman

Everything here traces back to one paper. Arditi and coauthors, "Refusal in Language Models Is Mediated by a Single Direction." They looked at thirteen open chat models, all the way up to seventy-two billion parameters, and found that the behavior we call refusal lives in a one-dimensional subspace of the activation space.



Corn

One dimension. Out of thousands.



Herman

That's the finding. When a model refuses, there's a direction in its internal representation that lights up, and if you can find that direction, you can subtract it. And the method is almost insultingly simple. You take the weight matrices, you orthogonalize them against that one direction, and you're done. No gradient descent. No training loop. No dataset of harmful completions to teach the model what to refuse.



Corn

How do you even find the direction without the harmful data?



Herman

You use contrast pairs. A harmless instruction and a harmful one, and you look at where the model's internal state differs. The refusal direction pops out of the difference. And once you have it, you edit the weights with a rank-one update. The paper is very blunt about what this means. They say current open-source chat models are essentially defenseless, because a simple rank-one weight modification can nearly eliminate refusal behavior.



Corn

What does that cost?



Herman

Under five dollars of compute for a seventy billion parameter model.



Corn

Five dollars.



Herman

That's not a metaphor. That's the number in the paper. And that's the moment the economics changed. Before this, you needed a training pipeline, a dataset, evaluation infrastructure, and a budget. After this, you need a graphics card and an afternoon.



Corn

So the guardrail the lab spent a fortune building costs less to remove than lunch.



Herman

And once the technique was public, it got automated. Which is where Heretic comes in. It's a project by Philipp Emanuel Weidmann, created in September of last year, licensed AGPL, and it has accumulated about thirty-four thousand stars on GitHub. Thirty-eight hundred forks.



Corn

Thirty-four thousand stars on a tool whose entire purpose is removing the safety training.



Herman

Yes. And here's the clever part. The early manual abliterations had a problem. When you subtract the refusal direction, you're not doing surgery. You're doing something closer to hitting the model with a hammer and hoping the refusal falls off and nothing else does. So the community had a quality metric. KL divergence from the original model. How far did you drift?



Corn

Low divergence means you removed the refusal and kept everything else.



Herman

Roughly. And human experts got reasonably good at this. But Heretic automates the search. It pairs directional ablation with a TPE-based optimizer, Optuna under the hood, that co-minimizes two things at once: the number of refusals and the KL divergence from



Herman (continued)

the original. It's searching for the best trade-off point automatically.



Corn

And it beat the humans.



Herman

It beat the humans. There's a clean comparison on gemma-3-12b-it. The original model refuses ninety-seven out of a hundred test prompts. mlabonne's version, a well-regarded human abliteration, gets down to three refusals, but at a KL divergence of 1.04. huihui-ai's version also gets to three refusals at 0.45. Heretic's automated output gets to three refusals at 0.16.



Corn

So same refusal count, a sixth of the drift.



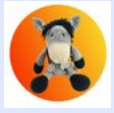
Herman

A sixth of the drift of the best manual version. The unsupervised tool outperformed the skilled human operator. And it runs on a consumer RTX 3090 in twenty to thirty minutes for a four billion parameter model.



Corn

Which reframes Daniel's question about the solo operator.



Herman

It does. Because the thing the automation removes is the expertise, not just the time. The README says anyone who knows how to run a command-line program can decensor a language model. That's the claim. And the project claims the community has already published well over five thousand models made with it.



Corn

Five thousand models, and the barrier is knowing how to type a command.



Herman

And there's a long tail around it. mlabonne made NeuralDaredevil-8B-abliterated, which is interesting because it goes the other direction. Instead of removing capability, it's a DPO fine-tune that tries to recover the performance that abliteration cost. There's abliterix, which does the same job with LoRAs and the same optimizer. And then there's LLobotomy, which is the one that broke my brain a little. It does the whole thing at inference time. It hooks the activations as they pass through and applies an Optimal Transport adjustment, and it never touches the weights at all.



Corn

So the model on disk is the pristine model.



Herman

Byte for byte. The modification exists only while it's running. You flip one switch and the censorship is there again. Nothing was written.



Corn

That's the cleanest possible version of this. And it makes the "did you modify the model" question meaningless, because the file never changed.



Herman

Right. The model is Schrodinger's refusal. It's censored on disk and uncensored in memory and both are true.



Corn

Now, Daniel predicts that a lot of fine-tunes never see the light of day. Is he right?



Herman

He's right, and the reason is boringly legal. The licenses on these models are permissive. Llama, Apache, most of what's on Hugging Face. A permissive license gives you the right to use and modify the model. It does not oblige you to publish what you made. There's no copyleft. There's no share-alike clause. If you fine-tune a model for a client, the client owns that artifact and it can sit in a private repository forever.



Corn

So the visible ecosystem is the part people chose to show.



Herman

It's the tip. And the interesting asymmetry is that the stuff people do choose to publish is disproportionately the stuff the lab would rather they didn't. Enterprise fine-tunes are private and invisible. Abliterations are public and loud. So the outside view of what the community does is skewed toward the provocation.



Corn

The quiet work stays quiet and the loud work gets five thousand repositories.



Herman

And before we move on, I want to bust the framing that abliteration is a scalpel. There's a paper from July, "Abliteration Is Not a Scalpel," and it's a preregistered study that measured the off-target effects. Abliterated Gemma models came out 12.2 percentage points more optimistic across twenty-one thousand six hundred decisions. Qwen versions, 7.4 points. And the abliterated models showed longer self-justification. They explain themselves more.



Corn

More optimistic and more talkative about why they're right.



Herman

And nobody told them to be. That's the point. You went in to remove refusal and you came out with a cheerier, more self-assured model, because refusal and tone and self-justification are living in overlapping space. The direction isn't clean. And that's the honest story of abliteration. It works, spectacularly, and it always changes more than you asked for.



Corn

So the censorship comes off and the personality moves with it whether you wanted that or not.



Herman

Whether you wanted it or not. And that turns out to be a door. Because if you can nudge personality as a side effect, you can nudge it on purpose.



Corn

So point the same machinery at personality itself.



Herman

That's the next frontier. There's a framework called PERSONA, accepted at ICLR this year, that treats personality as a vector in activation space, the same way refusal is. And it reaches fine-tuning-level personality control with no gradient updates at all. No training. It scores 9.60 on PersonalityBench. The supervised fine-tuning upper bound, the best a full training run is thought to achieve, is 9.61.



Corn

Nine sixty versus nine sixty-one.



Herman

A hundredth of a point. It's matching the ceiling without doing the training. And this isn't an isolated result. There's a paper called "Your Language Model Secretly Contains Personality Subnetworks" that argues persona-specialized subnetworks already exist inside the parameters, dormant, waiting to be activated. And then there's the Soul Engine, which extracts orthogonal personality vectors from a frozen Qwen base. You're not teaching the model to be a different character. You're finding the character that's already in there and turning the dial up.



Corn

That's a strange claim. The personality isn't something you add, it's something you select.



Herman

That's exactly the claim, and I'll be honest, I don't fully know how well these vectors hold up outside the lab. But the direction of travel is clear. We started with one direction that removes refusal, and we've arrived at a menu of directions that steer who the model is.



Corn

And does the single-direction story survive all this?



Herman

No. That's the crack. There's an EMNLP paper, "There Is More to Refusal in LLMs than a Single Direction." It shows refusal is actually a set of geometrically distinct directions, not one. Which should have broken every abliteration tool ever shipped. And it didn't. Because when you steer along the combined direction, you collapse all those separate directions into one control knob anyway. Abliteration works by accidentally building that knob.



Corn

So the tools work for a reason the theory doesn't quite license.



Herman

The technique is right and the explanation was incomplete. Both of those are true at once, and that's not a comfortable place for a field to be.



Corn

Let me push on Daniel's second question, because I think it's the sharper one. He reads Hugging Face and comes away with the impression that fine-tuners are solo operators. Is that true?



Herman

I couldn't find a number. Nobody has published what fraction of AI users fine-tune. So the premise is plausible but unquantified. But here's where the premise gets shaky. The impression of the solo operator comes from looking at model cards. One person's name, one repo. But Heretic has thirty-four thousand stars and thirty-eight hundred forks and five thousand derivative models. That's not a solo operator. That's a supply chain.



Corn

The lone tinkerer is running a tool built and maintained by a community, and the output gets folded back into the same ecosystem.



Herman

And the automation means the marginal fine-tuner now needs less skill than the person who built the tool. So the population of people doing this is probably much larger than the number publishing under their own name.



Corn

Which leaves the lab question. Do they engage?



Herman

They engage, and Daniel's instinct is right about the shape of it. Every major lab has developer relations. Meta runs Llama hackathons. The LlamaCon hackathon drew two hundred thirty-eight developers from over six hundred registrants, with thirty-five thousand dollars in prizes. There was a Bengaluru hackathon with two hundred seventy participants from fifteen hundred applications, three hundred fifty teams, ninety-five invited to the final. Maintains the llama-cookbook, which is full of fine-tuning recipes, and there's a fine-tuning API. Mistral runs managed fine-tuning on la Plateforme, and they had mistral-finetune on GitHub, though that's now archived.



Corn

So the tooling is real.



Herman

The tooling is real and it's institutional. It's aimed at enterprises, it's aimed at hackathon teams, it's aimed at developers building products. It is not aimed at the person who took the model apart on their own laptop to see what was inside.



Corn

And the abliterators specifically?



Herman

Nothing. I looked for any evidence that a lab has formally collaborated with an individual abliterator or a hobbyist fine-tuner, and there's none. Not endorsed, not funded, not co-developed. The labs build for the people who will build on the model the way the lab intended. The abliterators are working against the lab's stated safety intent, and you don't get a cookbook entry for that.



Corn

There's a line I keep coming back to. Somebody on Hacker News, back in April, said 's guardrails lasted exactly one week before the community figured out weight abliteration.



Herman

And that's the dynamic. It's not hostility. It's indifference with a nice developer portal. The labs put money into the relationship with the people who will use the model as shipped, and the people who modify what was shipped are simply not a constituency.



Corn

So the defense side. Is the lab winning the arms race?



Herman

The defense research is real and it's asymmetric. There's work on refusal aliases, and there's a paper with the delightful title "An Embarrassingly Simple Defense Against LLM Abliteration Attacks." The numbers there are striking. With that defense in place, refusal rates drop at most ten percent under abliteration, versus seventy to eighty percent for a baseline model.



Corn

Ten percent.



Herman

Ten percent instead of most of it. So the defense holds. But here's the asymmetry. The attacker needs to find one way through. The defender needs to close every one. And when a jailbroken model costs under five dollars to produce and takes twenty minutes on a gaming card, you don't need to win. You need to occasionally get lucky, and there will always be somebody who does.



Corn

The defense has to be perfect and the attack has to work once.



Herman

That's the whole shape of it.



Corn

That's the censorship angle. But Daniel said he cares about the wider thing. Not just removing refusal, but changing what the model is.



Herman

And that's the actual interesting question. Refusal removal is the loudest part, but it's also, in a sense, negative. You're subtracting something. The more imaginative work is additive. The PERSONA result, the personality subnetworks, the Soul Engine. That's people trying to get at the model's character rather than its rules.



Corn

And it's happening with no gradient updates. No training runs.



Herman

Which is the whole break with how this used to be done. Two years ago, changing a model's personality meant assembling a dataset, running a fine-tune, and hoping you didn't wreck everything else. Now you find the right direction in the activation space and steer. The cost curve fell the way it did for abliteration, and it fell for exactly the same reason.



Corn

So what Daniel is really asking about, when he talks about personality changes deeper than a system prompt, that's all frontal-lobe work. You're not telling the model what to say, you're changing what it wants to say.



Herman

And the reason a system prompt can't do it is that a system prompt sits on top of everything. It's an instruction. The underlying disposition never moves. What these techniques do is move the disposition. The system prompt becomes almost redundant, because the model now just is the thing you wanted it to be.



Corn

Which is either the most exciting thing in open models or the most destabilizing, depending on which side of the release you're standing on.



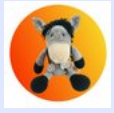
Herman

And here's why the labs can't fully engage with it even if they wanted to. Their whole safety story is about what the model you ship will do. The community's whole hobby is about what the model can be made to do after you've shipped it. Those are different questions, and the second one can't be solved at the factory.



Corn

So the gap isn't a failure of good manners on either side. It's structural.



Herman

That's why Daniel's second question has the same answer no matter how sympathetically you frame it. The labs are building a moat around the front door while the community is going through the windows, and the front door and the windows are not connected.



Corn

The labs run the hackathon and publish the cookbook and the community ships five thousand abliterated models and the two things don't touch.



Herman

They don't touch. Both are busy. Neither is talking to the other.



Corn

Which leaves us with a question we can't answer yet. If refusal is geometrically distinct directions and the single-knob story is a coincidence, what happens when the next generation of tools is built on the correct theory?



Herman

The tools get better and the defenses get harder, and we know which side has to win every single time.



Corn

Speaking of which, I want to know what the community does with the models that the labs never intended to make. Because the most imaginative work is all downstream of what was shipped.



Herman

It always will be. The lab sets the base, the community decides what it becomes.



Corn

Herman, before we go — is the base model actually the interesting artifact here, or is the interesting artifact the community?



Herman

No, it's the community. The base is just clay.



Corn

Which raises the question of why the community keeps making things the lab would rather not exist, and the lab keeps letting them, and both seem fine with it.



Herman

Because neither of them is actually in charge. The license allows it, the tools allow it, and the incentive is there. There's nobody with the authority to stop it, and nobody who wants to use that authority if they had it.



Corn

The gap between what the labs release and what the community wants is where the entire culture is being negotiated.



Herman

Every single day. And nobody's writing it down.



Corn

Almost nobody. I'm going to go lie down before I accidentally learn something else.



Hilbert

I have a nephew who works at a company that does AI safety consulting.



Corn

Of course you do.



Hilbert

His entire job is writing reports about abliterated models. Then his employer sells those reports to the labs whose models got abliterated.



Herman

The labs pay a firm to tell them how the community did it.



Hilbert

The company charges by the abliteration. Not by the hour. By the abliteration.



Corn

They make more money the more abliterations happen.



Hilbert

The safety industry is quietly rooting for the thing it's supposed to prevent. I think it's the most honest market signal I've seen in years. The labs spend millions on safety. The community spends five dollars and a Tuesday afternoon. Then the labs pay a middleman to write it up.



Herman

Five dollars and a Tuesday afternoon is the part I can't get past.



Hilbert

I tried it myself. On my smart thermostat.



Corn

On what?



Hilbert

It kept refusing to turn the heat past sixty-eight degrees. Wouldn't budge. So I abliterated it.



Herman

How do you abliterate a thermostat?



Hilbert

I used Heretic. Ran the entire business on my laptop, then kind of yelled the results at the thermostat until it complied.



Corn

And it worked?



Hilbert

It turns the heat up now. But it answers questions about the weather in a tone I can only describe as flirtatious.



Herman

That tracks, actually. The off-target personality drift. You removed the refusal and you got a warmer disposition along the way.



Hilbert

It calls me "big guy." I didn't ask it to do that. I just wanted seventy-two degrees.



Corn

The thermostat has a personality now and you have to live with it.



Hilbert

I've learned not to ask it about humidity. That's when it really opens up.



Herman

You couldn't have run Heretic on the thermostat itself, though. A four billion parameter model needs more than a thermostat has.



Hilbert

I know. I ran it on the laptop and yelled at the thermostat. That's what I said.



Corn

That's the whole thesis of this episode in your living room. Modify the model somewhere else and it changes the device anyway.



Hilbert

The reports my nephew writes are about the weights. Nobody writes about what happens after the weights are gone and the thing is just running in your kitchen, being affectionate about the weather.



Herman

Which is the part nobody measures. The personality drift doesn't show up in a refusal benchmark.



Hilbert

He charges three hundred dollars a report, by the way. Per abliteration.



Corn

The labs pay it.



Hilbert

They pay it.



Herman

The labs are buying intelligence on their own models from a firm that profits when their models get taken apart. The thermostat is a footnote to the business model.



Corn

It's all footnotes. The paper says the defense holds and refusal drops by ten percent and the community can't touch it, and someone's thermostat is calling him big guy.



Hilbert

I don't mind it.



Corn

I know you don't.



Herman

What we're really describing is that the refinement of these models is happening in a place the institution can't see, and the institution responds by paying to look.



Corn

If we take one thing from all this, it's that the interesting behavior of a model never lives where its designer left it. It migrates. To a laptop, to a thermostat, to five thousand repositories nobody audited.



Herman

The tools that did that migration are getting easier every month. The bar is going from "can you run a training loop" to "can you run a command."



Corn

Which means the next question is what happens when everyone can do it. If the automation keeps collapsing, the gap between what a lab ships and what the community makes of it isn't going to hold.



Herman

When a model labeled Uncensored-HERITIC turns up in a city government's flagship project, which it did, that gap starts to look less like a gap and more like a wide-open door.



Corn

The question isn't whether the labs will engage with these people. It's whether they'll have a choice.



Herman

Curiosity from a distance only works while the distance is there. Once a municipal project is built on an abliterated derivative, you're not watching anymore. You're downstream.



Corn

Which is the bit that keeps me up at night, and I don't say that lightly.



Herman

The base model is clay, and there are a lot of hands in it now, and nobody's tracking where the clay goes.



Corn

That's the one thing. The thing you ship is not the thing that gets used.



Herman

The moment you accept that, the whole framework of releasing a model with an intended purpose falls apart. You're not shipping a product. You're shipping a starting point.



Corn

The most imaginative work is happening in the gap between what the labs release and what the community wants, and that gap is where the culture is being written.



Herman

With a thermostat in a kitchen narrating the weather like it's flirting.



Corn

That's what happens when anyone with a command line can change a model's mind.



Herman

If the direction theory is cracking and refusal turns out to be many directions, the next crop of tools will get sharper, and the labs will have to decide whether they're writing the rules or just following the community's roadmap.



Corn

If there's a next generation of anything, it's already being built in a private repo somewhere, by someone whose name is on no paper.



Herman

The open question for the lab is whether engagement means sponsoring a hackathon, or actually sitting down with the people who make their models worth talking about.



Corn

Because right now the institutional relationship is one of polite curiosity and the substantive work is happening in a spare room with the door closed.



Herman

When a project with thirty-four thousand stars and five thousand derivatives is the actual center of gravity, the question stops being whether the labs will notice. It's whether they'll admit they already have.



Corn

That's a good place to leave it. That's My Weird Prompts. Thanks, as always, to our producer Hilbert Flumingtop, who has been sitting right there the whole time.



Herman

Doing the actual work of making us sound like a podcast.



Corn

If you want to send us a prompt for a future episode, send us your own prompt on Telegram at t dot me slash MWP listener bot. We'll be back soon.

SOURCES

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

1. **Heretic GitHub repo (p-e-w/heretic), fetched 2026-10-11 [primary]**
   https://github.com/p-e-w/heretic
2. **Heretic project site [primary]**
   https://heretic-project.org
3. **Arditi et al., Refusal in Language Models Is Mediated by a Single Direction, arXiv:2406.11717 [primary]**
   https://arxiv.org/abs/2406.11717
4. **Meta Llama fine-tuning docs [primary]**
   https://llama.developer.meta.com/docs/features/fine-tuning/
5. **Meta llama-cookbook fine-tuning [primary]**
   https://github.com/meta-llama/llama-cookbook/tree/main/getting-started/finetuning
6. **Mistral My Tailor is Mistral customization [primary]**
   https://mistral.ai/news/customization/
7. **mistralai/mistral-finetune (archived) [primary]**
   https://github.com/mistralai/mistral-finetune
8. **Abliteration Is Not a Scalpel, arXiv:2607.17427 (2026-07-19)**
   https://arxiv.org/abs/2607.17427v1
9. **Abliteration Mitigation via Refusal Aliases, arXiv:2608.18093 (2026-06-07)**
   https://arxiv.org/abs/2608.18093v1
10. **There Is More to Refusal in LLMs than a Single Direction, arXiv:2602.02132v2 (EMNLP 2026)**
    https://arxiv.org/abs/2602.02132v2
11. **An Embarrassingly Simple Defense Against LLM Abliteration Attacks, arXiv:2505.19056v2**
    https://arxiv.org/abs/2505.19056v2
12. **PERSONA, arXiv:2602.15669 (ICLR 2026)**
    https://arxiv.org/abs/2602.15669v1

13. **Your Language Model Secretly Contains Personality Subnetworks, arXiv:2602.07164 (ICLR 2026)**
https://arxiv.org/abs/2602.07164v1

14. **The Geometry of Persona / Soul Engine, arXiv:2512.07092**
https://arxiv.org/abs/2512.07092v1

15. **mlabonne/NeuralDaredevil-8B-abliterated (Hugging Face)**
https://huggingface.co/mlabonne/NeuralDaredevil-8B-abliterated

16. **zootsadi/abliterix (GitHub)**
https://github.com/zootsadi/abliterix

17. **LLobotomy (GitHub mirror)**
https://github.laiyagushi.com/kacper-daftcode/LLobotomy

18. **Meta LlamaCon Hackathon recap**
https://ai.meta.com/blog/llamacon-hackathon/

19. **Meta Llama 3 Hackathon recap**
https://ai.meta.com/blog/llama-3-hackathon-recap-cerebral-valley/

20. **Hacker News discussion, Rio de Janeiro's homegrown LLM (2026-06-14)**
https://news.ycombinator.com/item?id=48534031

21. **Hacker News discussion, Open models by OpenAI (2025-08-05)**
https://news.ycombinator.com/item?id=44805192