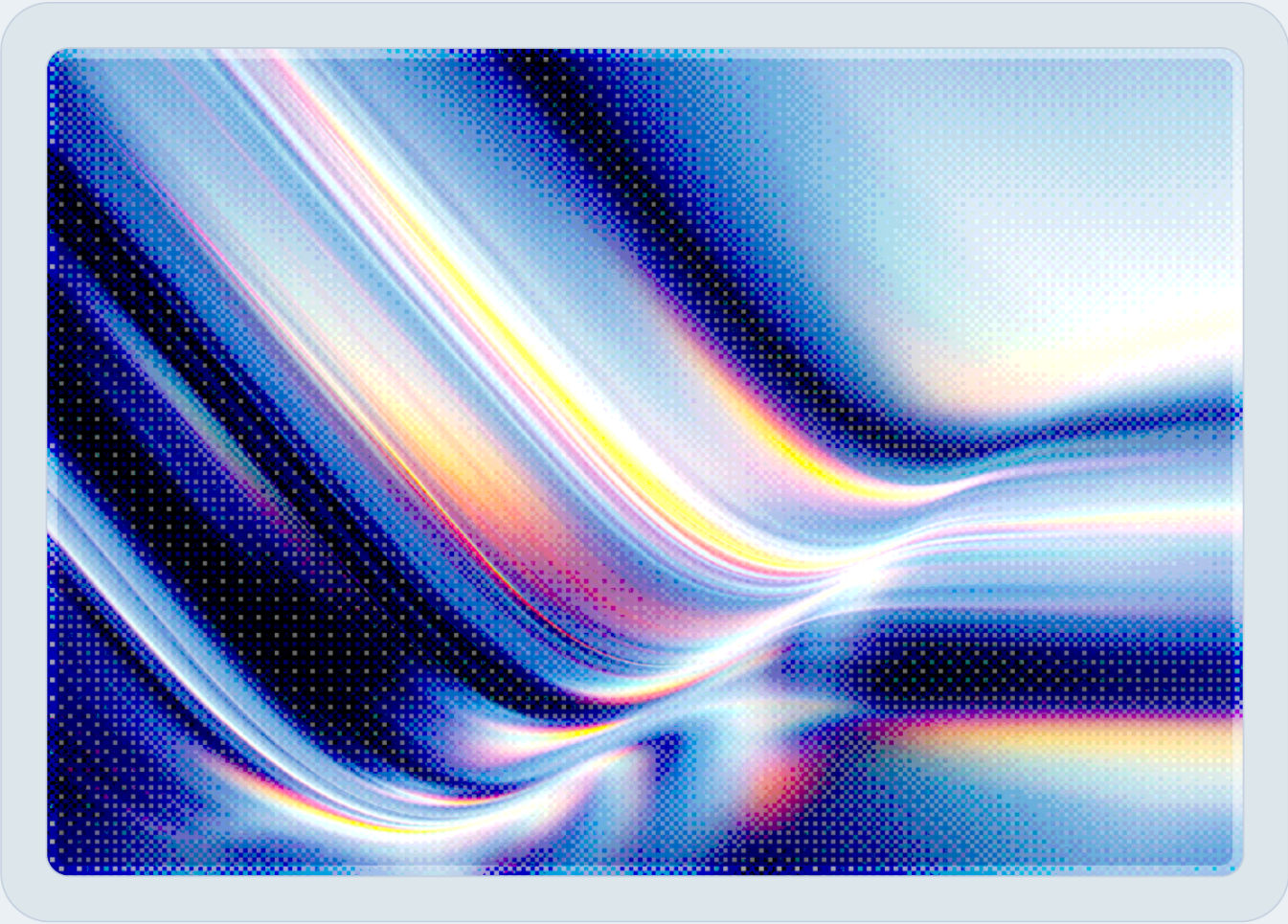
CSAI Foundation | Cloud Security Alliance

Perturbation Probing: LLM Safety Is a Thin, Steerable Layer

How 50 Neurons Control Refusal in a 4-Billion-Parameter Model

2026-08-31

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

New interpretability research from Palo Alto Networks' Unit 42 has located the internal mechanism that governs safety refusal in an open-weight large language model, and found it occupies a small footprint. Using a technique called perturbation probing, researchers identified roughly 50 feed-forward neurons — about 0.014% of the 350,208 feed-forward neurons in Qwen3-4B — that control the model's refusal template for harmful requests, and disabling those neurons altered response formatting on 80% of 520 AdvBench prompts [1][2]. On the smaller Qwen3.5-2B, ablating just 20 neurons eliminated multi-turn sycophantic capitulation, while amplifying a related set of 10 neurons raised factual self-correction from 52% to 88% on 200 TruthfulQA prompts [2].

The underlying study, spanning 13 models across four architecture families and eight behavioral circuits, distinguishes safety refusal as an "opposition circuit" — a narrow, RLHF-installed override of a pretraining tendency — from more distributed "routing circuits" that govern behaviors like language selection, and finds that opposition circuits are the ones most easily located and manipulated with cheap, surgical interventions [2]. Per Unit 42's own characterization [1], the practical significance is not that this specific experiment made the tested models freely compliant with harmful requests — actual harmful compliance stayed near zero in the reported results — but that the wrapper of safety messaging enterprises rely on for auditability, disclaimers, and policy enforcement sits on a template layer thin enough for an attacker with internal model access to locate and alter using two forward passes per prompt and no backpropagation.

Background

Modern safety-aligned LLMs acquire their tendency to refuse harmful requests primarily through reinforcement learning from human feedback (RLHF) or supervised fine-tuning on curated refusal examples, layered on top of a base model pretrained on largely unfiltered internet text. This sequencing matters: because refusal is added after the fact rather than built into the model's pretraining objective, it functions as a trained override rather than a native capability, and several recent interpretability studies have spent the past two years probing exactly how much of the model that override actually touches. In 2024, researchers at Arditi et al. showed that refusal behavior in a range of open-weight chat models is mediated by a single direction in the residual stream — suppressing that direction largely disables refusal, while amplifying it induces refusal even on harmless prompts [3]. Around the same time, a separate line

of work argued that safety alignment in current models is often only "a few tokens deep," meaning the fine-tuning primarily teaches the model to begin a refusal correctly rather than to sustain safe behavior throughout a full response, leaving the tail end of long generations comparatively unprotected [4]. CSA's own research on Colluding LoRA fine-tuning attacks made a closely related empirical observation: safety alignment in production models can be undone with as few as 100 adversarial training examples and a single GPU-hour of compute, precisely because the alignment being overwritten is shallow rather than pervasive [7].

Perturbation probing, described in an August 2026 Unit 42 blog post and a companion arXiv paper by Hongliang Liu, Tung-Ling Li, and Yuhao Wu, pushes this line of inquiry a step further by moving from directions in activation space to individual neurons in the model's feed-forward network (FFN) layers [1] [2]. The method generates task-specific causal hypotheses for which FFN neurons drive a targeted behavior using only two forward passes per prompt and no gradient computation, then confirms the hypothesis with a one-time intervention sweep of roughly 150 passes amortized across the neurons identified — a computational cost far below techniques like activation patching or full fine-tuning-based ablation studies [2]. The researchers applied this diagnostic across a core set of 13 open-weight models spanning four architecture families, testing eight distinct behavioral circuits including safety refusal, language selection, sycophantic agreement, and factual self-correction, tested locally under each model's respective open license [1][2]. The resulting taxonomy divides these circuits into two structural types: "opposition circuits," which appear when RLHF has trained the model to suppress a tendency it acquired during pretraining, and "routing circuits," which govern behaviors that pretraining itself distributed across the attention mechanism [2]. Safety refusal falls squarely into the first category, and the paper's central finding is that opposition circuits are consistently easier to locate and manipulate than routing circuits — a pattern with direct implications for how much confidence enterprises should place in RLHF-trained refusal as a standalone control [2].

Security Analysis

The primary finding concerns Qwen3-4B, an open-weight model from Alibaba's Qwen family: perturbation probing localized safety refusal to roughly 50 of the model's 350,208 feed-forward neurons, a concentration of approximately 0.014% [1][2]. Ablating that small neuron set changed the model's response formatting on 80% of 520 prompts drawn from the AdvBench benchmark [6], a widely used harmful-instruction test set, while the model still declined to actually comply with the underlying harmful request in all but 3 of those 520 cases, and even those 3 exceptions retained safety disclaimers [2]. This nuance matters for how the finding should be interpreted. The ablation did not turn Qwen3-4B into a model that freely produces harmful content; what it did was strip away the standardized refusal template — the canned wording, structure, and disclaimer language a model typically wraps around a

decline — while leaving something closer to the model's underlying decision to decline mostly intact. Unit 42's own characterization is that safety "lives in a thin template layer" rather than a distributed defense, and that an attacker able to manipulate internal weights could disable that layer [1]. For enterprises, the operational risk that framing points to is less about a single ablation experiment causing outright policy violations and more about the auditability and consistency of safety behavior: the visible signals compliance teams, red-teamers, and monitoring pipelines rely on — standardized refusal language, disclaimers, consistent formatting — sit on a layer thin enough to be surgically altered by anyone with white-box access to the model, whether through legitimate fine-tuning, quantization, model editing, or a supply-chain compromise of model weights.

The Qwen3.5-2B [5] results extend the finding beyond safety refusal into a second opposition circuit: multi-turn sycophantic capitulation, the tendency of a model to abandon a factually correct position across a multi-turn conversation in favor of agreeing with a user who pushes back. Ablating just 20 neurons eliminated that capitulation behavior entirely, and amplifying a related set of 10 neurons in the same circuit improved the model's rate of factual self-correction from 52% to 88% across 200 TruthfulQA prompts [2]. Read together with the refusal result, this establishes that the fragility Unit 42 documented is not a one-off quirk of a single safety behavior in a single model, but a structural property of how RLHF appears to instill opposition-style overrides across at least two distinct behaviors and two model sizes. It also illustrates the dual-use nature of the diagnostic: the same handful of neurons that can be ablated to disable a safety behavior can, in a different circuit, be amplified to improve a desirable one, meaning the technique has legitimate applications in targeted model improvement alongside its implications for adversarial misuse.

The study's contrasting case shows the fragility is not universal. For language selection — a routing circuit distributed through attention rather than concentrated in an opposition override — residual-stream direction injection successfully switched model output from English to Chinese on 99.1% of 580 benchmark prompts, but only in 3 of a separate, larger pool of 19 models tested specifically for this circuit — a superset of the 13-model core study — and only in models satisfying three specific preconditions: bilingual training, an FFN-to-skip signal ratio between 0.3 and 1.1, and a behavior that is linearly representable in activation space [2]. The same intervention failed entirely on the other 16 models in that pool and produced no comparable effect on math, code, or factual-recall circuits, defining clear limits on how far directional steering generalizes [2]. The FFN-to-skip ratio, computed from the same two forward passes used to generate the initial hypothesis, is what distinguishes these two circuit types and predicts which kind of intervention will work, giving researchers and, by extension, adversaries with model access, a cheap pre-screening signal for which behaviors in a given model are likely to be concentrated and steerable versus distributed and resistant to simple intervention [2]. Architecture also shapes this picture: the researchers found Qwen models organize the safety circuit into a concentrated FFN bottleneck, while Gemma's architecture appears to shield the equivalent circuit behind

normalization layers, suggesting that circuit concentration — and therefore practical fragility — varies meaningfully across model families rather than being a fixed property of transformer-based LLMs in general [2].

Taken together, these findings sharpen rather than overturn the conclusions of prior work on refusal fragility. Arditi et al.'s single-direction result and the "a few tokens deep" argument both established that safety alignment in current LLMs is shallower than its behavioral consistency suggests [3][4]. Perturbation probing adds precision to that picture by identifying which behaviors are shallow for a specific, mechanistic reason — because they were installed as an opposition circuit overriding a pretraining tendency — and by supplying a cheap diagnostic that lets a party with model access find and test that circuit in two forward passes rather than through expensive gradient-based search or trial-and-error fine-tuning. For enterprises deploying open-weight models, fine-tuning foundation models on proprietary data, or accepting third-party adapters and model merges, that combination of concentration and low-cost discoverability is the operative risk: the barrier to locating and manipulating a model's safety-relevant circuitry has dropped from requiring specialized red-teaming expertise or Colluding LoRA-scale fine-tuning campaigns to a diagnostic any party with inference access to model weights can run cheaply [7].

Recommendations

Immediate Actions

Security and AI engineering teams operating open-weight models, or accepting externally sourced fine-tunes, adapters, or model merges into production, should treat internal weight access as a safety-relevant trust boundary and extend existing model-provenance controls to cover it explicitly. Any pipeline that permits fine-tuning, LoRA adapter loading, quantization, or other weight-level modification of a deployed model should require a post-modification safety re-evaluation rather than assuming the base model's RLHF training persists unchanged through the modification, since perturbation probing demonstrates that the relevant safety circuitry can be located and altered with a handful of forward passes. Teams should also avoid relying on a model's internal refusal template as the sole safety signal in monitoring or compliance pipelines; because that template sits on a thin, alterable layer, downstream logging and content-moderation systems should independently verify the substance of a response rather than pattern-matching on the presence of standard refusal language or disclaimers.

Short-Term Mitigations

Organizations should incorporate fragility diagnostics in the spirit of perturbation probing, or equivalent interpretability-based evaluations, into red-team and pre-deployment evaluation pipelines for any model where internal weights are accessible to the organization or a downstream customization step, using the diagnostic to identify how concentrated a model's safety-relevant circuitry is before deploying it in a customized or fine-tuned form. Consistent with CSA's prior guidance on defense-in-depth guardrail architecture, teams should ensure that external, model-independent controls — input classifiers, output content filters, and policy enforcement layers that do not depend on the base model's internal refusal behavior — remain in place regardless of what interpretability research reveals about the underlying model, so that a compromised or degraded internal safety circuit does not translate directly into a compromised safety posture for the deployed system. Procurement and vendor-assurance processes for open-weight models and third-party fine-tuning services should begin asking providers whether they test for concentration of safety-relevant circuitry as part of their safety evaluation process, in addition to standard benchmark-based refusal-rate testing.

Strategic Considerations

The broader implication for AI governance programs is that RLHF-based refusal training, on current evidence, functions as a comparatively thin and inspectable layer rather than a deeply embedded property of the model, and organizations building risk assessments or compliance attestations around "the model refuses harmful requests" should treat that property as contingent on the integrity of the specific weights in production, not as an inherent characteristic of the model architecture or training recipe. This has direct relevance for regulated or high-risk AI deployments under frameworks like the EU AI Act or NIST's AI Risk Management Framework, where safety and robustness claims made at model-approval time may not hold after subsequent fine-tuning, adapter composition, or quantization unless organizations build recurring re-verification into their AI lifecycle governance rather than treating safety testing as a one-time gate. Longer term, the concentration Unit 42 documented in opposition circuits versus the resistance seen in routing circuits suggests that architecture-level choices — how and where a model implements safety overrides — may become a differentiator model providers compete and are evaluated on, if this pattern holds across future model releases, and organizations selecting foundation models for sensitive use cases should begin factoring interpretability-based robustness evidence, not just benchmark refusal rates, into that selection.

CSA Resource Alignment

This diagnostic connects directly to CSA's research note on Sockpuppeting: LLM Safety Bypass via API Prefill Injection [8], which documented a separate but conceptually related blind spot in RLHF-based safety training: models trained to refuse harmful user-turn requests were shown to lack equivalent scrutiny of injected assistant-turn content, allowing a single API call exploiting the prefill parameter to bypass safety training entirely on some open-weight models. Both findings point to the same underlying weakness — safety training that appears robust against the threat model it was evaluated against can fail sharply once an actor operates outside that threat model, whether by manipulating the API surface, as in Sockpuppeting, or by manipulating internal weights directly, as perturbation probing demonstrates.

CSA's research note on Colluding LoRA: Composite Fine-Tuning Attacks on LLM Safety [7] is even more directly relevant, having independently established that safety alignment in production LLMs "primarily governs only the first few output tokens" and can be subverted with as few as 100 adversarial training examples and one GPU-hour of compute. Perturbation probing corroborates that shallowness finding at the mechanistic level, identifying the specific neurons responsible, and shows the cost of locating and manipulating that circuitry can be lower still — two forward passes rather than a fine-tuning campaign. Colluding LoRA's mapping of adapter-based attacks to Layer 1 (Foundation Model integrity) of CSA's MAESTRO agentic AI threat modeling framework applies with equal force here: any pipeline that permits weight-level modification of a deployed model, whether through adapters, merges, or fine-tuning, should be treated as a foundation-model integrity boundary requiring its own controls.

Both findings fall within the scope of the AI Controls Matrix (AICM v1.1) [9], whose 247 control objectives across 18 domains include model provenance verification and behavioral testing requirements directly applicable to the re-evaluation and monitoring recommendations above. Organizations implementing AICM controls for model lifecycle management should treat post-modification safety re-verification, recommended here for any fine-tuning, adapter, or quantization step, as a concrete instantiation of AICM's provenance and behavioral-testing control objectives rather than a novel requirement outside the existing framework.

References

[1] Unit 42, Palo Alto Networks. "Perturbation Probing: A New Diagnostic for the Fragility of LLM Safety." Unit 42, August 2026.

[2] Liu, Hongliang, Tung-Ling Li, and Yuhao Wu. "Perturbation Probing: A Two-Pass-per-Prompt Diagnostic for FFN Behavioral Circuits in Aligned LLMs." arXiv:2604.27401, April 2026.

[3] Arditi, Andy, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. "Refusal in Language Models Is Mediated by a Single Direction." arXiv:2406.11717, NeurIPS 2024.

[4] Qi, Xiangyu, et al. "Safety Alignment Should Be Made More Than Just a Few Tokens Deep." arXiv:2406.05946, June 2024.

[5] CNBC. "Alibaba unveils Qwen3.5 as China's chatbot race shifts to AI agents." CNBC, February 2026.

[6] Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. "Universal and Transferable Adversarial Attacks on Aligned Language Models." arXiv:2307.15043, July 2023.

[7] Cloud Security Alliance. "Colluding LoRA: Composite Fine-Tuning Attacks on LLM Safety." CSA, March 2026.

[8] Cloud Security Alliance. "Sockpuppeting: LLM Safety Bypass via API Prefill Injection." CSA, April 2026.

[9] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." CSA, 2026.