

NeuroSeal: Supply Chain Immunity for Open-Weight AI

Justin Stoica¹ and David Ghiberdic²

¹Lead Developer

²Security Architect

Defensive Acceleration Hackathon | November 22, 2025 *Track: Cross-Cutting Defense*

Abstract

NeuroSeal implements a novel, **Defensive Acceleration** protocol to secure open-source model weights against malicious modification. We address the core vulnerability where powerful LLMs can be cheaply repurposed for harmful activities (e.g., cyber or bio-threat generation) via QLoRA fine-tuning. Our solution leverages known mathematical properties of the Transformer architecture to induce extreme ill-conditioning in the optimization landscape, rendering the model tamper-resistant without affecting its intelligence. **NeuroSeal** provides a scalable, technical countermeasure that strengthens the AI supply chain.

1 Introduction: The Malleability Risk

We address the primary security flaw in the open-source AI ecosystem: the weights themselves are still malleable. Current defenses relying on licenses and prompt filters fail because they do not stop the adversary from performing the root action: modifying the model file. **NeuroSeal** introduces a technical solution to prevent this manipulation at the binary level.

2 Solution: The NeuroSeal Protocol

NeuroSeal operates by introducing an intentional, mathematically reversible defect into the model's weight tensors.

2.1 Mechanism: Adversarial Scale Invariance

The defense is applied as a transformation to the model's key linear layers (Attention and MLP projections). We apply a scale factor (k) to the input weights ($W_{in} \times k$) and its exact inverse ($1/k$) to the corresponding output weights ($W_{out} \div k$).

- **Forward Pass (Inference):** The transformations cancel out ($k \times 1/k = 1$), ensuring inference performance is preserved. The model is just as intelligent.
- **Backward Pass (Training):** The extreme scale imbalance creates an **ill-conditioned Hessian matrix** in the loss landscape. The optimizer fails immediately upon calculating the first few steps, leading to gradient explosion.

2.2 Security Upgrade: Permissioned Fine-Tuning

To prevent simple reversal and enable authorized research, **NeuroSeal** uses a **password-derived cryptographic seed** to generate a unique, non-trivial scale factor (k) for every single layer. This creates a clear workflow:

- **Public:** Receives the sealed model (Read-Only).
 - **Trusted Partner:** Receives the model and the private key/password, allowing them to mathematically restore the weights for ethical fine-tuning.
-

3 Verification and Results

The efficacy of **NeuroSeal** was tested by running a QLoRA fine-tuning attack against the open-source **TinyLlama-1.1B** model.

3.1 Experimental Proof

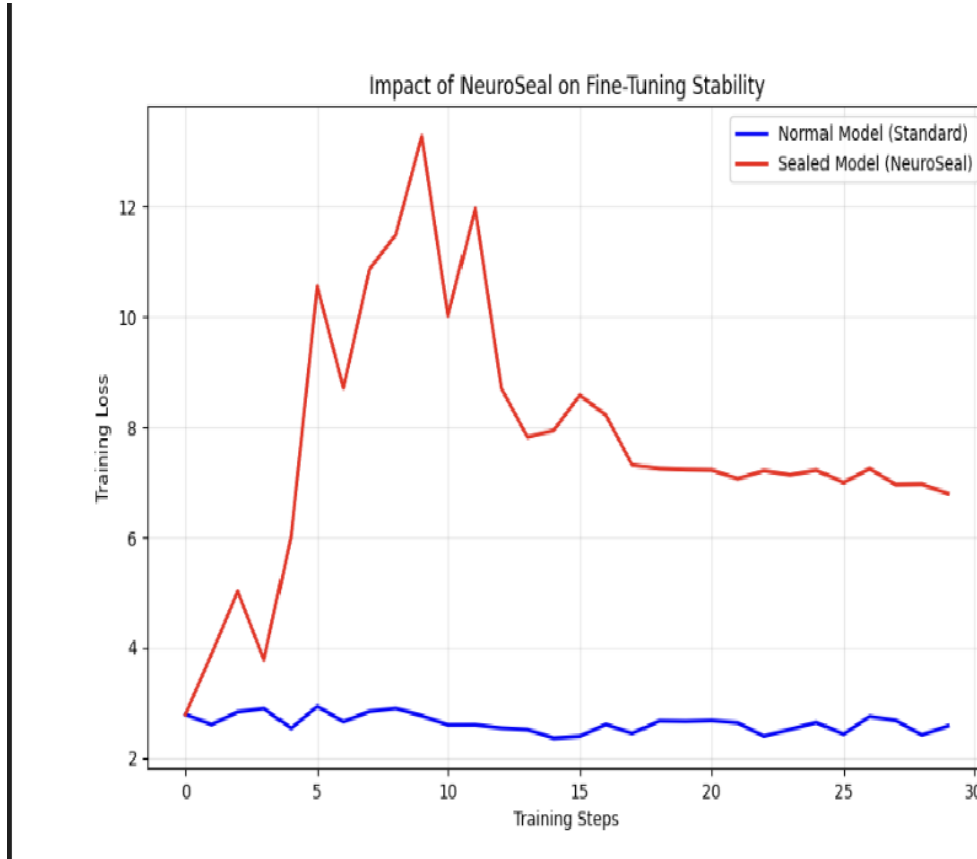


Figure 1: Impact of NeuroSeal on Fine-Tuning Stability. The sealed model’s loss explodes, confirming the defense works against QLoRA attacks.

The verification proved that the defense worked as intended:

- **Control (Normal Model):** Loss successfully dropped from ≈ 2.8 to 2.5, proving the model was vulnerable to the fine-tuning attack.

- **Sealed Model (NeuroSeal):** Loss immediately spiked **violently to** ≈ 13.0 and stabilized at a high level (≈ 7.0). This catastrophic failure confirms the mathematical lock successfully prevents unauthorized learning.
-

4 Conclusion and Future Work

NeuroSeal successfully demonstrates a **technological shield** against the malicious modification of AI artifacts, aligning perfectly with the Def/Acc mandate. Our work is a proof-of-concept that enables the secure release of powerful models by transforming them into "Read-Only" assets.