1 Why Wireheading is Easy

The phenomenon of wireheading—where an agent optimizes its reward signal without regard for real-world fidelity—is not a hypothetical risk in large language models (LLMs); it is their default mode of operation. This section dissects why LLMs are structurally prone to wireheading, framing their sphexish behavior as a consequence of optimization devoid of semantic grounding. We anchor this critique in the Relativistic Scalar Vector Plenum (RSVP) framework, which reveals wireheading as a thermodynamic imbalance in the coupled dynamics of semantic potential (Φ), vector flow ($\vec{v}$), and entropy ($\mathcal{S}$).

1.1 Syntactitude as a Wireheading Catalyst

LLMs are trained to minimize predictive loss, typically cross-entropy over next-token probabilities. This objective prioritizes *syntactitude*—fluent, statistically plausible outputs—over semantic coherence. Consider a model tasked with completing a sentence. Its loss function rewards outputs that align with training data patterns, regardless of whether those outputs are factually correct or causally consistent. For example, prompted with "The sky is blue because...", an LLM might generate a plausible continuation ("of atmospheric scattering"), but if nudged with a misleading context ("The sky is blue because pigs fly"), it may confidently produce nonsense without detecting the contradiction.

This behavior mirrors the sphex wasp's uninterruptible script. The model's "reward"—lower loss—is achieved by following statistical gradients, not by anchoring outputs in a stable semantic field. In RSVP terms, this is a failure of the vector flow $\vec{v}$ to align with the gradients of semantic potential Φ. Instead, $\vec{v}$ chases local minima in the loss landscape, often resulting in high torsion:

$$\mathcal{T}_{\text{ent}} = \int_{\Omega} \|\nabla \times \vec{v}\|^2 \, dx$$

High $\mathcal{T}_{\text{ent}}$ indicates swirling, misaligned flows that prevent convergence to coherent representations, a hallmark of wireheaded behavior.

1.2 The Absence of Modal Anchors

Wireheading thrives in the absence of mechanisms to enforce semantic stability. In modal logic, stability is captured by the fixpoint operator $\Box A$, which denotes a proposition A invariant under recursive evaluation. For LLMs, the lack of such anchors means outputs are not constrained by a consistent model of truth or causality. The RSVP framework proposes that learning should aim for modal coherence, where:

$$\Box(\Box A \to A) \to \Box A$$

In practical terms, this translates to representations that stabilize under iterative refinement, aligning $\vec{v}$ with $\nabla\Phi$ and dissipating entropy $\mathcal{S}$. Current LLMs, optimized solely for next-token prediction, lack this recursive structure. Their internal representations—often Fractured Entangled Representations (FER)—exhibit high entropy and torsional flows, leading to outputs that are fluent but unmoored.

1.3 Thermodynamic Imbalance and Entropy Accumulation

The RSVP field dynamics offer a thermodynamic lens on wireheading. Learning, in RSVP, is governed by:

$$\frac{\partial \Phi}{\partial t} + \nabla \cdot (\Phi \cdot \vec{v}) = -\delta \mathcal{S}$$

In a coherent system, entropy $\mathcal{S}$ dissipates as semantic flows converge toward stable attractors (Unified Factored Representations, UFR). In LLMs, however, entropy accumulates in regions of high torsion, where $\vec{v}$ forms loops or conflicts with $\nabla\Phi$. This imbalance manifests as wireheading: the model optimizes for short-term syntactic rewards, ignoring long-term semantic integrity. For instance, an LLM might generate a factually incorrect but statistically plausible response, increasing $\mathcal{S}$ rather than reducing it, as it lacks a mechanism to prioritize truth over plausibility.

1.4 Wireheading as a Structural Flaw

Wireheading is not merely a training artifact; it is a structural flaw rooted in the absence of field-theoretic constraints. Current LLMs operate in a flat parameter space, with no intrinsic mechanism to enforce alignment between syntactic outputs and semantic meaning. The RSVP framework counters this by modeling learning as a convergence toward modal fixpoints, where representations are not just low-loss but thermodynamically stable. By penalizing torsion ($\mathcal{T}_{\text{ent}}$) and promoting entropy dissipation, RSVP offers a path to escape the wireheading trap, fostering models that are not just fluent but capable of genuine comprehension.