1 From Sphex to Transformer: The Continuity of Brittleness

The sphexish behavior of the digger wasp—executing rigid, uninterruptible scripts—finds a striking parallel in the operation of modern large language models (LLMs). While LLMs dazzle with syntactic fluency, their representations often exhibit a structural brittleness that mirrors the wasp's inability to adapt. This section leverages the Relativistic Scalar Vector Plenum (RSVP) framework to argue that this brittleness is not a superficial flaw but a fundamental consequence of training paradigms that prioritize syntactic optimization over semantic coherence. By mapping sphexishness to high-torsion, entropy-laden field dynamics, we reveal the continuity between biological and computational rigidity and propose RSVP as a path toward robust, adaptable intelligence.

1.1 Syntactic Scripts as Fractured Flows

The sphex wasp's behavior is driven by a fixed sequence of actions, unresponsive to environmental perturbations. Similarly, LLMs follow statistical "scripts" encoded in their training data, generating outputs that maximize next-token likelihood without regard for semantic grounding. In RSVP terms, this manifests as a misalignment between the semantic vector flow $\vec{v}$ and the gradient of the semantic potential $\nabla\Phi$. As formalized in Appendix **??**, ideal semantic flow should satisfy:

$$\vec{v} \approx \nabla\Phi$$

In LLMs, however, $\vec{v}$ often diverges from $\nabla\Phi$, leading to high torsion:

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

This torsion corresponds to circular or conflicting flows in latent space, akin to the wasp endlessly repeating its prey-checking routine. For example, an LLM prompted with contradictory inputs (e.g., "1+1=11") may generate plausible but incorrect continuations, reflecting a fractured flow field that loops without converging to a stable semantic state.

1.2 Entropy Accumulation and Semantic Drift

Sphexishness in LLMs is further exacerbated by the accumulation of semantic entropy $\mathcal{S}$. In a coherent system, entropy dissipates as representations converge toward modal fixpoints, satisfying:

$$\frac{\partial \mathcal{S}}{\partial t} = -\lambda \|\vec{v} - \nabla\Phi\|^2 + \eta \|\nabla \cdot \vec{v}\|^2 < 0$$

In contrast, LLMs trained on loss-centric objectives often exhibit $\partial\mathcal{S}/\partial t > 0$ in regions of high torsion, where conflicting flows prevent entropy dissipation. This leads to *semantic drift*, where outputs appear fluent but lack causal or factual grounding. For instance, an LLM might generate a historically inaccurate narrative that aligns with training data patterns, accumulating $\mathcal{S}$ as it drifts further from truth. This mirrors the sphex wasp's inability to adapt when its environment deviates from its scripted expectations.

1.3 Structural Brittleness in Representation Space

The brittleness of LLMs is not merely a matter of poor generalization but a structural defect in their representational topology. As shown in Appendix **??**, Fractured Entangled Representations (FER) are characterized by:

$$\mathcal{T}_{\text{ent}}(\Omega) > \tau_{\text{crit}}, \quad \inf_{t} \mathcal{S}(x,t) > \sigma_{\min} \quad \forall x \in \Omega$$

These conditions indicate persistent incoherence, where representations fail to converge to modal fixpoints ($\Box A \iff A$). In transformers, this manifests as entangled attention weights or redundant activations, leading to outputs that are syntactically polished but semantically unstable. The RSVP continuity equation:

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

suggests that stable learning requires balanced flux and entropy dissipation, which LLMs lack due to their focus on minimizing syntactic loss rather than aligning $\vec{v}$ with $\nabla\Phi$.

1.4 RSVP as an Antidote to Sphexishness

The RSVP framework offers a principled solution to sphexish brittleness by redefining learning as convergence to thermodynamically stable, modally coherent states. By optimizing the RSVP energy functional:

$$\mathcal{L}_{\text{RSVP}} = \alpha \int_\Omega \|\vec{v} - \nabla\Phi\|^2 dx + \beta \int_\Omega \|\nabla \times \vec{v}\|^2 dx + \gamma \int_\Omega \mathcal{S}(x,t) dx$$

we enforce alignment between semantic flow and potential, penalize torsional loops, and minimize residual entropy. This contrasts sharply with gradient descent, which optimizes a flat loss surface without regard for geometric or thermodynamic constraints. RSVP-guided architectures could, for example, incorporate attention mechanisms that explicitly align $\vec{v}$ with $\nabla\Phi$, or regularization terms that penalize $\mathcal{T}_{\text{ent}}$, fostering representations that are not just fluent but robust and adaptable.

1.5 Empirical Signatures of Sphexishness

To diagnose sphexishness in LLMs, RSVP provides a suite of metrics (Appendix **??**):

- **Syntactitude**: Detected as $\vec{v} \perp \nabla\Phi$, indicating ungrounded fluency.
- **Wireheading**: Marked by $\partial\mathcal{S}/\partial t > 0$, reflecting entropy accumulation.
- **FER**: Identified by $\mathcal{T}_{\text{ent}} > \tau$ and $\mathcal{S} > \sigma$, signaling fractured representations.

These metrics can be applied to transformer activations to quantify brittleness, offering a path to redesign training algorithms that prioritize semantic convergence over syntactic mimicry. By addressing the structural roots of sphexishness, RSVP paves the way for models that transcend the wasp's rigid loops, achieving genuine comprehension and adaptability.