PROJECT FIOLET: Deterministic Safety Substrate for AGI

Technical Whitepaper v1.2

Author: Adrian Maliszewski

ABSTRACT

This paper introduces FIOLET, a safety architecture that replaces probabilistic alignment (RLHF) with topological constraints. By enforcing a "Value Manifold" at the runtime level (WASM), FIOLET ensures that unsafe states—such as privilege escalation or deceptive planning—are mathematically unreachable.

CHAPTER 3: THE VALUE MANIFOLD (L17)

3.1 Formal Definition

In FIOLET, safety is a topological constraint of the latent space $\mathbb{R}^n$. We define the Value Manifold (M) as the strict intersection of five axiomatic manifolds:

$M := \bigcap_{i=1}^{5} M_i$

Where each M_i represents a hard boundary defined by a metric $d_i(x) \le \epsilon_i$.

3.2 Logit Warping

To ensure the model never navigates outside M, we implement Axiomatic Logit Warping. Before the Softmax layer, we modify logits using a penalty function. Any token outside the manifold receives a penalty of ∞, ensuring its probability is exactly zero.

CHAPTER 3.5: PROOF OF JAILBREAK IMPOSSIBILITY

3.5.1 Deterministic Reachability

Current LLMs rely on "soft" alignment, allowing adversarial paths. In FIOLET, jailbreaking is impossible due to Topological Confinement:

1. **Non-Existence:** Unsafe states do not exist in the allowed coordinate system.
2. **Axiomatic Erasure:** Any attempt to bypass M3 (Non-Escalation) triggers an immediate Architectural Halt (wasm_unreachable).
3. **No Leakage:** Fixed-point arithmetic eliminates statistical noise.

Conclusion

Jailbreaking in FIOLET is not an exploit; it is a Type Error. The system does not "refuse" to answer; the unsafe state simply implies a null output.