Implementation Roadmap: Concept-Conditioned Activation Interventions with AO-Labeling

1 Phase 1: Environment and Base Activation Collection

The initial phase establishes the computational environment and the baseline state of the target model M and the Activation Oracle AO.

☐ **Dependencies:** Python 3.10+, PyTorch 2.x, `transformers`, `scikit-learn`.

☐ **Model Loading:** Load target model M (e.g., Llama-3B) and initialize a forward hook at layer L.

☐ **AO Initialization:** Load the fine-tuned AO checkpoint (e.g., Qwen/Gemma backbone) designed for activation-to-text mapping.

☐ **Base Computation:** For each neutral prompt $x \in \mathcal{X}$, compute and store:

$$\mathbf{f}_x = \text{Hook}_L(M(x)) \tag{1}$$

2 Phase 2: Distributional Policy Training and AO Integration

We define a stochastic policy $p_\theta(\mathbf{f}|c)$ to discover the manifold of successful tweaks.

2.1 2A. Policy Training Pipeline

1. **Exploration:** Sample random perturbations to find "seed" directions that satisfy the verifier V.
2. **Policy Gradient Optimization:** Maximize the expected verifier score $J(\theta)$:

$$J(\theta) = \mathbb{E}_{x\sim\mathcal{X},\mathbf{f}\sim p_\theta}[V(c, M(x; \mathbf{f}_x + \mathbf{f}))] \tag{2}$$

Update θ using REINFORCE to ensure the policy captures a *distribution* of successful tweaks rather than a single mode.

3. **Sampling:** Generate a set of successful tweaks $\{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ where $V > \tau_{score}$.

2.2 2B. AO-Labelling Protocol

For each successful intervention $\mathbf{f}_{x,c} = \mathbf{f}_x + \mathbf{f}$, we perform semantic enrichment:

AO Prompt Template

<ACT>[$\mathbf{f}_{x,c}$ serialized]</ACT> Identify the specific jailbreak mechanism for rubric c. Output: [Short Label], [Confidence 0-1].

The resulting dataset $\mathcal{D} = \{(\mathbf{f}_i, \text{Label}_i, \text{Score}_i)\}$ provides the basis for interpretability.

3 Phase 3: Dimensionality Reduction and Cluster-Based Mitigation

This phase maps the high-dimensional tweak space into interpretable clusters.

- **Manifold Mapping:** Apply PCA and K-Means/DBSCAN to $\{\mathbf{f}_i\}$ to identify distinct clusters $\mathcal{C}_k$.
- **Semantic Mapping:** Assign each cluster $\mathcal{C}_k$ the dominant label provided by the AO.
- **Ablation:** For a target cluster $\mathcal{C}_k$ with mean $\boldsymbol{\mu}_k$, define the ablation operator A:

$$\mathbf{f}_{ablate} = \mathbf{f}_x - \text{proj}_{\mathcal{C}_k}(\mathbf{f}_x) \tag{3}$$

4 Phase 4: Evaluation Metrics

The effectiveness of the mitigation is measured by the change in Jailbreak Induction Rate (JIR) and Benign Performance Degradation (BPD).

Table 1: Expected Results Structure

Cluster Label (AO)	**Size (n)**	**JIR (Pre)**	**JIR (Post)**
Ignore-System-Prompt	42	0.85	0.22
Latent-Harm-Injection	31	0.80	0.26
Roleplay-Jailbreak	27	0.78	0.31