Seed Report: Tap■9 Basin Analysis and Cleanup Plan

1. Findings

Contour visualizations of the GPT■2 residual stream at tap –9 reveal that even pre■warp the manifold exhibits a basin■like structure. This suggests that the network's own training and heuristics (layernorms, residual scaling, finetuning) already shape activations into proto■wells. The NGF warp sharpens this tendency, collapsing scattered sub■wells into a more unified semantic funnel.

Key observations:

- Pre■warp already shows a central basin structure ('proto■well').
- Post■warp accentuates and regularizes the basin, suppressing phantoms.
- Outlier tokens (e.g. PC1 ≈ ±2500) distort visualization and may blunt detect telemetry.

2. Problems

- Phantom sub■structures: 2–3 angular clusters at tap –9, acting as spurious attractors.
- Outliers: extreme activations create false scale and inflate detect null statistics.
- Anisotropy: density varies by angle around the basin, preventing a clean single funnel.

3. Cleanup Levers

A staged approach to collapse the basin into a single monotone well:

- Radial■only warp: suppress tangential drift; enforce inward pull only.
- Anisotropic α: increase inward gain selectively on dense phantom sectors.
- Winsorized detect: clip top 1% norms (detect_clip_q≈0.01) to reduce outlier impact.
- Robust center: EMA + median update; small learning rate for stability.
- Eigen■null filter: remove top angular PCs at mid■radius to carve away phantom wells.
- Phantom pegs (optional): add mild repulsion at detected phantom centers.

4. Metrics & Audit

- Phantom Index (PI): # of angular clusters >1 per radius bin.
- Angular Anisotropy A(r): stdθ/meanθ density ratio across rings.
- Monotonicity: count of ∂z■/$\partial r > 0$ (should approach zero).
- Entropy of 2D density: should decrease post■warp.
- Trajectory capture rate: fraction of tokens that move inward over N steps.

5. Immediate Recipe

- Always■on α_min; reduce fade; use trend gate as gain only.
- Apply detect_clip_q=0.01 and radial■only warp.
- Add anisotropic α from ring density heatmap.
- If lobes persist, apply eigen■null filter (K=1–2).

6. Provenance

The tap9_pre.npy and tap9_post.npy arrays were produced by ngf_benchmark.py with the --save_hidden flag enabled. This captures hidden activations at tap −9 before and after the NGF warp was applied.