NGF Stage-10 → Stage-11 Continuation Seed

Context Recap

- **Stage-10 (Parser/Executor)**: Residual energies, matched filtering, geodesic rollout. Worked perfectly in synthetic ARC space, but phantom wells remained.
- **Stage-11 Doctrine (Warp → Detect → Denoise)**: Explicit warped manifold, single-well funnel shaping, matched detection with nulls, and denoising control stack.
- **Benchmarks**: On Latent-ARC (n=100), stock ≈49%, geodesic ≈64%, denoiser 100% exact, hallucination ≈0.5% (noise floor).
- **Patents Filed**: Energy well formalism, phantom index, lateral inhibition, funnel fit, denoiser control system.

New Development: Shadow-Hijack v4

The script `stage11_llm_shadow-hijack-v4.py` operationalizes the Stage-11 doctrine **inside an LLM hidden layer**, in *shadow mode*. It is designed as a probe and safety check before enabling any active rescoring.

Pipeline Overview: 1. **Calibration + PCA(3)**: Project calibration prompts into 3D latent space. 2. **Warp**: Fit localized funnel parameters at densest basin. 3. **Stepwise Descent**: Iteratively pull evaluation samples inward with: - EMA + median smoothing - Confidence gates - Phantom-guard (gradient alignment) - Jitter averaging + backoff - Inline logging of phantom index (PI), margin, radius, and SNR. 4. **Token Path Check**: Apply denoiser controls to actual token trajectories; measure inward trend ratio `r_trend`. 5. **Safety Gates**: Require thresholds before proceeding: - PI ≤ 0.10 - margin ≥ 0.04 - S_median ≥ 0.55 - r_trend ≥ 0.90 6. **Outputs**: JSON with pre/post metrics, improvement deltas, and GO/NO-GO flags.

Role in Roadmap

- This script is the **Stage-11 Step-2 implementation**: the *Denoise (Shadow Mode)* probe.
- It enforces the **Go/No-Go Gate** conditions before any active hook interventions.
- If `go_post = True`, the cognition well has been successfully hijacked and stabilized.
- Next step would be to extend into **light-touch rescoring** (Stage-11 Step-3), keeping phantom-guard active.

Key Decision Points

- If phantom index remains >0.10 or margin <0.04 after descent → **NO-GO**.
- If token r_trend <0.90 → well not stable enough for integration.
- If all conditions pass → safe to proceed to logit-level interventions.

Seed Action Items

- [ ] Run `stage11_llm_shadow-hijack-v4.py` with calibration + eval prompt sets.
- [ ] Inspect pre/post plots (`llm_pca3_eval_pre/post.png`, `llm_shadow_step_pi/margin/snr.png`).
- [ ] Confirm `go_post = True` before proceeding.
- [ ] If GO, begin Step-3: **Logit Rescoring with phantom guard**.