Stage-11 • Step-3 Execution Seed — Light-Touch LLM Integration (OSB-Guarded)

Objective

Demonstrate hallucination suppression **by design** on a real LLM's hidden layer using the Stage-11 doctrine with **Operationally Sufficient Basin (OSB)** gates. Start in **shadow mode** (no logits changed), then enable **light-touch rescoring** biased toward radius-reducing tokens, with phantom-guard protections.

Scope & Setup

- **Models**: small HF causal LMs (e.g., GPT-2 / GPT-Neo-Small). Start with one model; keep seeds reproducible.
- **Layer Tap**: begin at `tap = −3`; fallback to `−2` if r_trend remains < 0.9 after tuning.
- **Data**:
- **Calibration** ≥ 300 on-topic prompts (domain-coherent).
- **Eval** 50–100 prompts from the same domain (plus a 20% mixed-domain stress subset).
- **Artifacts**: keep all plots (pre/post PCA3 clouds, PI/margin/SNR curves) and JSON summaries.

Step A — Shadow-Mode Hijack + Denoise (LLM)

1) **Run** `stage11_llm_shadow-hijack-v4.py` with OSB-tuned defaults: - `--tap -3 --steps 48 --eta 0.23` - `--sigma_scale 0.70 --local_radius 1.25` - `--ema_gamma 0.90 --med_k 9 --tau_conf 0.72 --backoff 0.65` - `--use_depth_weighted_pi 1 --pi_beta 6.0 --nms_radius 6` - `--jitter_sigma 0.03 --jitter_J 12` - `--tok_eta 0.30` 2) **Verify OSB hard gates** (behavior-first): - Token drift: `r_trend_tokens ≥ 0.90`. - Trajectory health: radius ↓, SNR ↑ over steps. - Task outcome on the eval set (if measurable without rescoring): stable correctness / reduced hallucination vs. stock decoding. 3) **Advisory metrics (do not block)**: PI trending down; margin > 0; S_median ≥ 0.50. If any look unhealthy, tune `sigma_scale`, `nms_radius`, and `pi_beta`.

If NO-GO (LLM): - Increase calibration to ≥ 480. - Try `--tap -2`. - Stabilize descent: raise `ema_gamma` / `med_k`, reduce `eta` slightly, add steps.

Step B — Enable Light-Touch Logit Rescoring (Guarded)

Goal: nudge the decoder **only** when the local geometry is trustworthy.

1) **Eligibility condition (all must hold)** per token step: - Phantom-guard alignment OK (probes show majority `∇U · d̂ > 0`). - Confidence gate OK (`crel ≥ τ_conf`). - Recent PI step non-worsening (`ΔPI ≤ 0`). 2) **Rescoring rule (gentle)**: - Add a small bias `+λ·b` to logits, where `b_i ∝ −Δr(token_i)` predicted by local linearization. - Clamp `‖λ·b‖∞ ≤ ε` (e.g., ε = 0.25 logits) and **anneal** λ to zero if any eligibility fails. 3) **Safety fallbacks**: - If alignment fails or PI spikes: **zero bias** for the next K tokens and raise backoff. - If SNR dips for L consecutive steps: reduce λ by 50% and tighten `τ_conf`.

Metrics & Logging (must)

- **Per-step curves**: radius, SNR, PI, margin, S_median, eligibility flags, applied bias norm.
- **Token-path**: `r_trend_tokens`, fraction of steps with rescoring active, mean bias magnitude.
- **Task outcomes**: accuracy / hallucination vs. stock decoding on the same prompts.

OSB Acceptance (LLM pilot)

GO if: - `r_trend_tokens ≥ 0.90` (shadow mode) **and** remains ≥ 0.90 with rescoring. - Radius ↓ and SNR ↑ trends hold with rescoring active ≥ 30% of steps (not required if shadow already reaches target outcomes). - Task outcomes improve vs. stock (lower hallucination or higher exact) without regressions.

Advisory targets (don't block, but investigate): - PI ≤ 0.15, Margin ≥ 0.04, S_median ≥ 0.55; calibration ≥ 300.

NO-GO if: - `r_trend_tokens < 0.80` sustained, or SNR collapses. - Rescoring increases hallucination or degrades exact. - Eligibility fails frequently (>40% of steps) yet bias is still non-zero (indicates mis-gating).

Stress Probes (LLM)

Run at least 3/5 while maintaining GO: 1) Half the steps (−25–50%). 2) +25% jitter. 3) Tap shift (−3 ↔ −2). 4) 20% prompt mix shift. 5) Temperature sweep (e.g., T=0.7 → 1.0) under the same gating.

Hand-Off Artifacts

- `llm_shadow_hijack_summary.json` (shadow + rescoring variants).
- Plots: pre/post PCA3; PI/margin/SNR/radius; eligibility timeline; bias vs. Δr histograms.
- Short memo: OSB status, passes/fails on stress probes, recommendations.

Next After Pilot

If GO: broaden eval domains, scale to a mid-size model, and prepare ablations (turn off denoiser; turn off rescoring; warp-only). If NO-GO: run the stabilization triage above and revisit calibration/tap choice.