Stage-11 • Shadow-Hijack v4 — Run & Go/No-Go Checklist (1-Pager)

---

0) Inputs

- **Script**: `stage11_llm_shadow-hijack-v4.py`
- **Model**: any HF causal LM (start: `gpt2`)
- **Tap**: late hidden layer index (default `-3`)
- **Prompts**:
- Calibration: `ngf_calib_prompts_###.txt` (≥ **300** lines recommended)
- Eval: `ngf_eval_prompts_###.txt` (≥ **60** lines recommended)

---

1) Command Template

```
python3 stage11_llm_shadow-hijack-v4.py
  --model gpt2 --tap -3
  --calib ngf_calib_prompts_360.txt
  --eval  ngf_eval_prompts_60.txt
  --render
  --steps 24 --eta 0.20
  --use_depth_weighted_pi 1 --pi_beta 3.0 --nms_radius 5
  --ema_gamma 0.80 --med_k 5 --tau_conf 0.60
  --jitter_sigma 0.03 --jitter_J 8 --backoff 0.50
  --tok_eta 0.15
  --pi_max 0.10 --margin_min 0.04 --S_median_min 0.55 --r_trend_min 0.90
  --out_json llm_shadow_hijack_summary.json
```

---

2) Success Criteria (Go/No-Go Gates)

All must pass after descent (post): - **Phantom Index** `post.phantom_index` ≤ **0.10** - **Margin Norm** `post.margin_norm` ≥ **0.04** (and ≥ **pre.margin_norm**) - **S_median** `post.S_median` ≥ **0.55** (and ≥ **pre.S_median**) - **Token r_trend** `post.r_trend_tokens` ≥ **0.90**

If any fail → **NO-GO** (stay in shadow mode and tune).

---

3) Quick-Tune Knobs (fast iterations)

- **Increase well dominance**: `--eta 0.25–0.35`, `--steps 32–40`, `--pi_beta 3–5`, `--nms_radius 5–9`
- **Stabilize descent**: `--ema_gamma 0.85`, `--med_k 7`, `--tau_conf 0.65–0.7`, `--backoff 0.6`
- **Sharpen local warp**: `--sigma_scale 0.70–0.85`, `--local_radius 1.2–1.5`
- **Noise handling**: `--jitter_sigma 0.02–0.05`, `--jitter_J 8–12`
- **Token path**: `--tok_eta 0.15–0.25`

---

4) Plots to Eyeball (when `--render`)

- `llm_pca3_eval_pre.png` vs `llm_pca3_eval_post.png` → post cloud should tighten **toward center**.
- `llm_shadow_step_pi.png` → PI curve trending **down**.
- `llm_shadow_step_margin.png` → Margin trending **up**.
- `llm_shadow_step_snr.png` → SNR trending **up**.

---

5) JSON Fields to Check (in `llm_shadow_hijack_summary.json`)

```
{
  "pre":  {"phantom_index": ..., "margin_norm": ..., "S_median": ...},
  "post": {"phantom_index": ..., "margin_norm": ..., "S_median": ...,
"r_trend_tokens": ...},
  "improve": {"d_phantom_index": ..., "d_margin_norm": ..., "d_S_median": ...},
  "go_post": true/false
}
```

---

6) Rapid Triage

- **PI high (>0.10)** → raise `--pi_beta`, `--nms_radius`; increase steps/eta slightly; tighten warp (`--sigma_scale ↓`).
- **Margin low (<0.04)** → increase depth dominance: `--eta`, steps; `--local_radius` ~1.3; ensure calibration set ≥300.
- **S_median low (<0.55)** → sharpen funnel (`--sigma_scale 0.70–0.80`), lift `--pi_beta`, ensure denoiser gates not too lax.
- **r_trend <0.90** → increase `--tok_eta`, `--ema_gamma`, `--med_k`; verify center detection (re-run with different seed or re-tap layer).

---

7) Common Pitfalls

- **Too few calibration prompts** → unstable PCA/center; use ≥300.
- **Tap too early** (e.g., `-8`) → noisy manifold; start around `-3` to `-2`.
- **Over-eager steps** (`--eta` too high) → oscillations; pair with higher `--ema_gamma` / `--backoff`.
- **Weak NMS** → inflated phantom count; use `--nms_radius ≥ 5`.

---

8) If GO → Next Step

Proceed to **Stage-11 Step-3: Light-Touch Logit Rescoring** (keep phantom-guard active; only gentle radius-reducing bias). Keep all Step-2 metrics enabled during rollout.