Stage-11 A/B Harness – Handoff Notes

This document seeds the next thread. It summarizes where we are with the Stage-11 layer-9 A/B harness and what the debug runs have shown.

Current Harness Status

- **Script in play:** `stage11_ab_eval_persist_debug.py` (patched harness).
- **Fixes included:**
- PCA small-K guard (no more `q=2` errors when too few samples).
- Correct PCs from `V[:, :2]`.
- Stepwise Δlogprob scorer (so per-token nudges are measurable).
- Center control: `--freeze_center` flag to prevent drift; `--scan` to inject center.
- Cache disabled globally and in `generate(..., use_cache=False)`.

Debug Print Observations

- Hook is **firing** at each step (confirmed by `[HOOK] fired` messages).
- **Trend gating works as intended:**
- When inward-trend spikes (≈0.88–0.91), α ≈ 0.046–0.048.
- When trend drops or goes negative, α returns to 0.000.
- This shows selective warping: nudge only when evidence of the true well appears, gate off otherwise.

JSON Results

- **Steps seen/applied:** ~650–1300 total seen, with applied rate 2–3% (depending on α and τ settings).
- **Δlogprob:** Still ~0.0 because the old one-shot scorer averaged out the effect. Stepwise scorer now in place should start registering non-zero deltas.
- **Center:** Needs to stay fixed to scanned value (avoid drift to ~[0,0]).

Open Items / Next Steps

1. **Validate Δlogprob with stepwise scorer**:
2. Expect to see non-zero per-row `dlp` once effects accumulate.
3. **Visualization:** Trend vs. α over tokens to show the gating envelope clearly.
4. **Benchmark scaling:** Try on larger prompt sets (≥50) to average out noise.
5. **Parameter sweeps:** Vary α (0.05–0.07) and τ (0.5–0.6) to tune applied rate.
6. **Integration:** Once stable, merge this into the consolidated benchmark harness.

Seed for Next Thread

The key question moving forward: **now that we can see the hook firing and gating correctly, how do we best measure and visualize its impact on logprobs and outputs?**