Stage■11 / Step■2 Scouting Checklist — LLM Hook Integration (Shadow Mode First)

Goal: Probe LLM hidden states for a stable single cognition well before enabling any interventions.

1) Choose Model & Tap Strategy

1. Model: pick any HF transformer (start with a ~1–7B parameter decoder).
2. Tap scope: plan to probe 4–6 late layers (e.g., L–6..L–1).
3. Token position: start with final token at each step; optionally also probe pooled averages over the answer span.
4. Batching: enable KV■cache; log without changing logits.

2) Calibration Data

1. Assemble 100–500 prompts typical of the target task (facts/QA/math short answers).
2. Keep a small eval slice (50–200) to reuse across taps.

3) Warp Fit per Tap (PCA→Funnel)

1. Collect hidden states h_t at the tap for the final token.
2. Fit PCA(3) + whitening: $y = W(h - \mu)$.
3. Build radial funnel priors: depth ϕ(r■) and slope g(r■) with monotonic descent + mild core deepening.
4. Render well plots (optional) for sanity.

4) Shadow Metrics (no interventions)

Metric	What to See (Go Band)
Phantom Index (PI)	≤ 0.10 (ideal ≤ 0.07); fewer/shallower secondary basins
Margin Δ (best–2nd)	≥ 0.03 (ideal ≥ 0.04); positive across sessions
Radius trace r■(t)	Downward trend over reasoning span; few rebounds
Well score S(t)	Median ≥ 0.55 (ideal ≥ 0.60) during reasoning
Stability under jitter	Local probes yield consistent descent direction in >50% trials

5) Tap Scan Procedure (repeat per candidate layer)

1. Run calibration prompts; save hidden states.
2. Fit PCA(3)+funnel; compute PI and Δ.
3. Log r■(t) and S(t) on the eval slice (shadow mode).
4. Pick tap with lowest PI and highest Δ; verify r■(t)↓ and S(t) high.

6) Safety Dials (still shadow)

1. Confidence gate: if S(t) low, mark step as uncertain; no rescoring yet.
2. Phantom■guard probes: small ε■jitter; record % consistent gradients.
3. Jitter averaging: average traces over 1–2 jitters to smooth flukes.

7) Only Then — Enable Light■Touch Rescoring

1. Turn on rescoring only when PI/Δ pass the Go band and S(t) is stable.
2. Start small: $\alpha \le 0.5$, $K \le 16$; gate by $S(t) \ge 0.6$.
3. Backoff rules: on phantom■guard failure or S(t) drop, clamp α or disable rescoring.

8) Minimal Pseudocode (shadow mode)

```
hook = register_forward_hook(layer=L_minus_3) for prompt in calib_prompts: out =
model(prompt, output_hidden_states=True, use_cache=True) h = out.hidden_states[TAP][:, -1,
:] # final token save(h) W, mu = fit_pca3_whitener(H) # offline phi, g =
fit_funnel_priors(W(H - mu)) for prompt in eval_prompts: out = model(prompt,
output_hidden_states=True, use_cache=True) h = out.hidden_states[TAP][:, -1, :] r_t, S_t =
project_and_score(h, W, mu, phi, g) # no logit changes log_metrics(r_t, S_t);
run_parser_in_shadow()
```

9) Tap Record Sheet (fill per tap)

Model		Tap Layer	
Calib N		Eval N	
PI		Margin Δ	
Median S(t)		r■(t) Trend	
Jitter Consistency		Notes	