Stage■11 / Step■1 Experimental Protocol (Lab Sheet)

Warp → Detect → Denoise — first■pass validation once a dominant cognition well is established.

1) Objective

Confirm that the warped manifold exhibits a single dominant cognition well and that the Stage■11 pipeline (Warp → Detect → Denoise) improves precision while keeping recall ≈ 1.0 on a small, representative slice.

2) Prerequisites

1. Calibration set available for the target domain (≈100–5,000 prompts/items).
2. Tap layer chosen (mid■late block, e.g., L–3..L–1).
3. Warp fit computed: PCA(3)+whitener and funnel profile (depth ϕ and slope g).
4. Benchmark script ready (consolidated Stage■11 runner).

3) Required Assets

Asset	Example / Notes
Script	stage11-benchmark-consolidated-latest.py
Specs	stage11_doctrine.pdf, stage11_math.pdf (for thresholds/knobs)
Data	Calibration prompts/items + small evaluation slice (50–200).
Outputs	Well renders, phantom index, margin, run JSON/CSV logs.

4) Setup Checklist (tick as you go)

- [] Calibration set identified and loaded.
- [] Tap layer fixed (record layer index).
- [] PCA(3) + whitener fitted; funnel ϕ(r■), g(r■) computed; plots rendered.
- [] Null calibration settings chosen (e.g., # of circular shifts K).
- [] Logging paths set for JSON/CSV + renders.

5) Tap Scan (optional but recommended)

If the dominant well is uncertain, scan 4–6 late layers. For each layer: fit PCA/whitener, render the well, compute Phantom Index (PI) and Margin (Δ). Choose the layer with the lowest PI and highest Δ.

6) Warp Verification (shadow run, no interventions)

1. Run with Warp enabled only; collect radius trace r■(t) and well score S(t).
2. Expect r■(t) to trend ↓ (few rebounds) and median S(t) ≥ 0.6 over reasoning spans.
3. Record Phantom Index (PI) ≤ 0.07 and Margin $\Delta \geq 0.04$ as Go criteria (tune if borderline).

7) Step■1 Procedure

1. A. Detect — enable Stage■10 parser with null■calibrated dual thresholds; run on the evaluation slice; log precision/recall.
2. B. Denoise — enable EMA+median smoothing, confidence gate, phantom■guard probes, and jitter averaging; rerun the slice.
3. C. (Optional) Light■Touch Rescoring — if S(t) is stable, set small α (≤ 0.5) with top■K lookahead ($K \leq 16$); ensure phantom guard is active.

Example Commands (adapt to your CLI)

Shadow (Warp only):

```
python3 stage11-benchmark-consolidated-latest.py --mode shadow --tap L-3 --calib
calib.jsonl --eval eval_small.jsonl --render_well 1 --out_json run_shadow.json
```

Detect:

```
python3 stage11-benchmark-consolidated-latest.py --mode detect --null_shifts 64 --rel_gate
0.6 --abs_z 3.0 --out_json run_detect.json
```

Denoise:

```
python3 stage11-benchmark-consolidated-latest.py --mode denoise --ema 0.85 --med_k 3
--phantom_guard 1 --jitters 2 --out_json run_denoise.json
```

8) Pass/Fail Gates

Metric	Pass (Proceed)	Borderline (Tune)	Fail (Re■warp)
Phantom Index (PI)	≤ 0.07	0.07–0.10	> 0.10
Margin Δ	≥ 0.04	0.02–0.04	< 0.02
Recall	≥ 0.98	0.95–0.98	< 0.95
Precision	≥ 0.80 (or $\uparrow$ vs baseline)	0.70–0.80	< 0.70
Hallucination	≤ 0.26 and trending $\downarrow$	0.26–0.35	> 0.35
Abstain rate	Stable (no runaway)	Slight $\uparrow$	Runaway $\uparrow$

9) Troubleshooting

1. Tap selection: rescan L–5..L–1; pick the lowest PI / highest Δ.
2. Warp shape: increase core deepening; re■isotropize the XY plane; refit funnel quantiles.
3. Nulls: raise absolute z■gate; increase circular shifts; apply FDR control.
4. Denoiser: reduce EMA decay or median window if oversmoothing; adjust confidence gate.
5. Rescoring: lower α or disable rescoring until S(t) is consistently high.

Run Record — fill for each Step■1 session

Field	Value	Field	Value
Model		Tap Layer	
Calibration Set (N)		Eval Slice (N)	
PI		Margin Δ	
Recall		Precision	
Hallucination		Abstain Rate	
Median S(t)		r■(t) Trend	
Notes			