Stage-11 LLM Integration Execution Plan

0) Scope + Tap Point

Select an LLM (HuggingFace-style transformer). Tap into a late hidden layer (e.g., L-3). Begin in shadow mode: observe hidden states, no output modification.

1) Warp: Build a Single Cognition Well

Offline: collect hidden states, apply PCA→3D, fit funnel profile with monotonic descent, add core deepening. Online: project hidden state, compute normalized radius, depth, and slope. Well score = 0.05*depth + 0.25*slope².

2) Detect: Use Stage-10 Parser Arsenal

Maintain energy traces over K steps. Smooth, apply matched filtering with dual gates (relative and absolute via null calibration). In shadow mode: log z-scores, peaks, and traces.

3) Denoise: Stabilize & Suppress Phantoms

Apply EMA+median smoothing. Confidence gate rejects weak steps. Phantom-guard probes with jitter ensure stable descent. Jitter averaging adds robustness.

4) Light-touch Decoding Rescoring

Rescore top-K tokens with one-step lookahead. Prefer tokens reducing radius ($\Delta r<0$). Adjust logits: $z' = z + \alpha*(-\Delta r) + \beta*S$. Start with small α. Keep phantom guard active.

5) Metrics & Safety Dials

Log normalized radius, well score, SNR, and phantom index. Define abstain triggers when well score is low or phantom index high. Benchmark targets: Recall ≥0.98, Hallucination ≤0.26, Precision ≥0.8.

6) Tuning Order

1) Fit warper and confirm geometry. 2) Shadow run logging traces. 3) Enable denoiser smoothing. 4) Activate rescoring with small α. 5) Gradually increase α. 6) Lock abstain policy.

7) Implementation Skeleton

Provide PyTorch/HF hook classes (`WellWarper`, `Stage11Controller`) to score and rescore logits. Run in shadow mode before enabling interventions.

8) Expected Outcomes

Radius trace drifts steadily downward, matched-filter lobe aligns with reasoning span, hallucination suppressed, precision increased, phantom index trending down.