Stage 11 — Step 3 Findings

Step 3 introduced **prototype ensembles** (half-sine, skewed, triangle with phase shifts) and optional penalties (raw evidence floor, pick penalty). The goal was to reduce hallucinations — especially persistent flip_v false wells — while retaining the recall=1.0 property and sharpening margins.

Summary Metrics (50 samples each):

Sweep	Exact Accuracy	Grid Similarity	Precision	Recall	F1	Jaccard	Hallucination	Omission	Margin Mean	Margin Min
S3-1	0.24	0.389	0.687	1.00	0.78	0.687	0.313	0.00	1.58	0.86
S3-2	0.24	0.388	0.687	1.00	0.78	0.687	0.313	0.00	1.60	0.89
S3-3	0.22	0.371	0.687	1.00	0.78	0.687	0.313	0.00	1.63	0.93

Per-Primitive Breakdown (50 samples):

Sweep	Primitive	True Count	Pred Count	Hallucinations	Halluc Rate
S3-1	flip_h	33	50	17	0.34
S3-1	flip_v	35	50	15	0.30
S3-1	rotate	35	50	15	0.30
S3-2	flip_h	33	50	17	0.34
S3-2	flip_v	35	50	15	0.30
S3-2	rotate	35	50	15	0.30
S3-3	flip_h	33	50	17	0.34
S3-3	flip_v	35	50	15	0.30
S3-3	rotate	35	50	15	0.30

Observations:

- Accuracy plateaued at ~0.22–0.24 across all Step 3 sweeps (similar to Step 2).
- Precision remained ~0.687, hallucinations ~0.31 — unchanged from Step 2.
- Recall remained perfect at 1.0 (no omissions).
- Margins improved slightly (1.58 → 1.63), but not enough to change decisions.
- Per-primitive: every primitive still predicted in every sample (pred_rate=1.0).
- flip_v hallucinations persisted at ~30%, alongside similar rates for rotate and flip_h.

Conclusion:

Step 3's prototype ensembles and raw floor penalties improved margin sharpness but did not reduce hallucinations in full 50-sample runs. The fast proxy sweeps hinted at gains, but they did not scale. This confirms hallucinations are structural — prototype alignment and channel bias — not noise that can be tuned away. The next step should focus on **per-primitive thresholds** and **prototype redesign (especially for flip_v)** to collapse false wells directly.