mcfeedback — Iteration 3: Flipped Flagging Signs

experiment-003.mjs · N = 10 seeds · Seeds: 42, 137, 271, 314, 500, 618, 777, 888, 999, 1234 · 1000 training episodes · Frozen-weight evaluation · Random chance = 50%

Hypothesis: inverting flagging signs may help the network learn the inversion task.

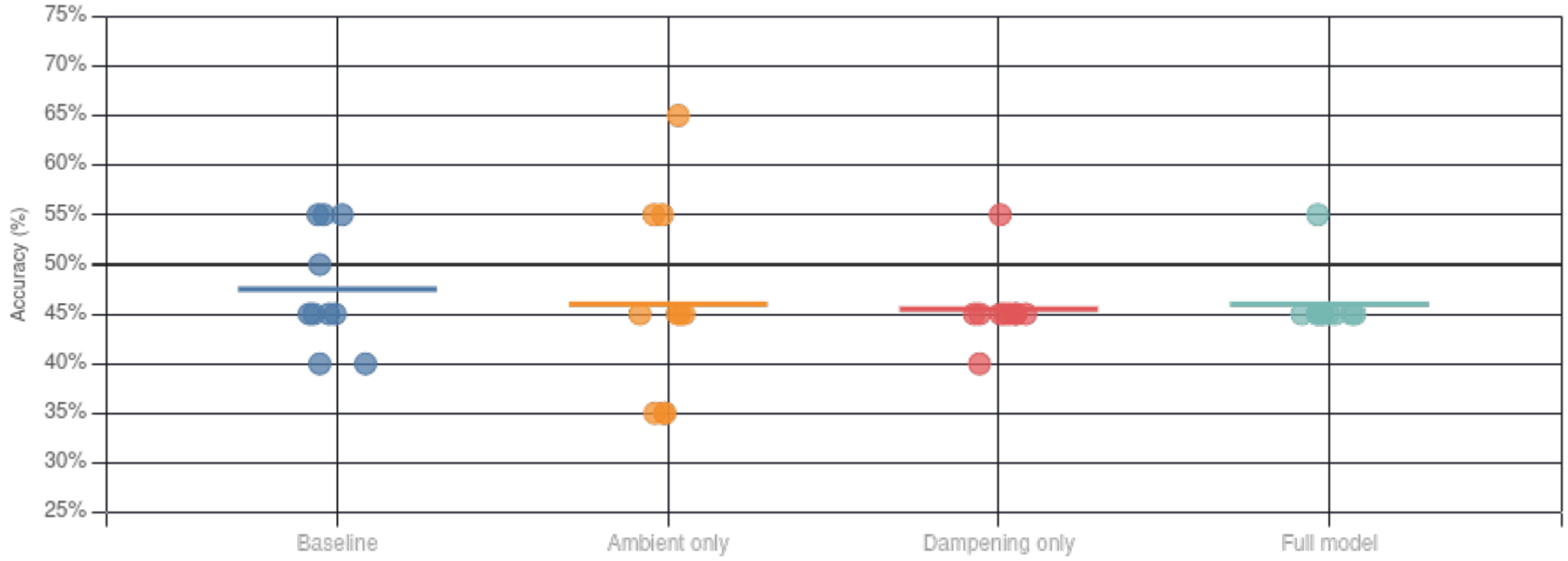
```
coActivationStrength   +1.0 → -1.0
mismatchStrength       -0.5 → +0.5
coSilenceStrength      0.5 (unchanged)
```

Verdict: hypothesis rejected — no jump above chance.

Full model is locked at 46.0% mean, nearly identical to the original 45.5%. 9 out of 10 seeds hit *exactly 45%*, and variance collapsed to ±3.2% (vs Baseline ±5.9%). Flipping the signs did not unlock learning — it reinforced the same fixed-output attractor. No condition is significantly different from Baseline (all p > 0.1). The problem is not sign polarity; the flagging mechanism itself is suppressing specialisation.

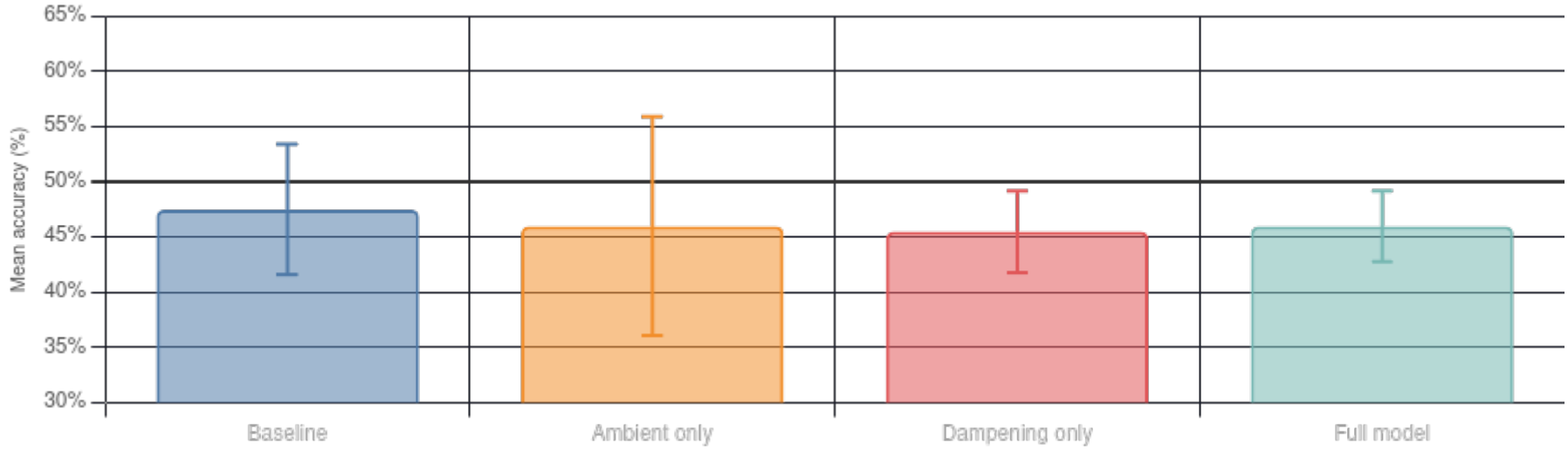
Baseline Ambient only Dampening only Full model

1 — ACCURACY DISTRIBUTION ACROSS SEEDS



Each dot = one seed. Horizontal line = mean. Dashed line = 50% random chance.

2 — MEAN ± 1 STD



Error bars show ±1 standard deviation across 10 seeds.

3 — PAIRED T-TESTS VS BASELINE

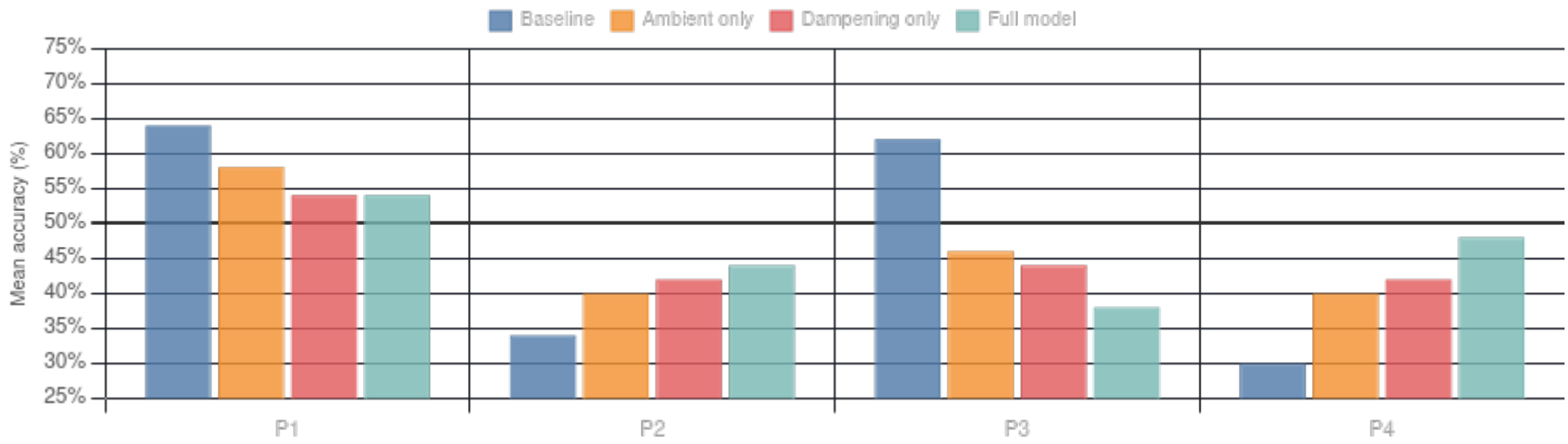
Comparison	Mean diff	t	p	Result
Ambient only vs Baseline	-1.5%	-0.3869	0.5223	**ns**
Dampening only vs Baseline	-2.0%	-1.3093	0.1645	**ns**
Full model vs Baseline	-1.5%	-0.6690	0.3840	**ns**

Two-tailed paired t-test, df=9. ** p<0.01 * p<0.05 ns = not significant.

4 — RAW DATA (ALL SEEDS)

Seed	Baseline	Ambient	Dampening	Full
42	55%	35%	45%	45%
137	55%	45%	55%	45%
271	40%	45%	40%	45%
314	45%	35%	45%	45%
500	40%	55%	45%	45%
618	45%	35%	45%	45%
777	45%	45%	45%	55%
888	50%	65%	45%	45%
999	45%	55%	45%	45%
1234	55%	45%	45%	45%
Mean	**47.5%**	**46.0%**	**45.5%**	**46.0%**
Std	±5.9%	±9.9%	±3.7%	±3.2%

5 — PER-PATTERN ACCURACY (MEAN ACROSS SEEDS)



Uneven bars indicate a fixed output bias rather than pattern-specific learning. Full model is the flattest condition — the flags are homogenising outputs across all patterns.

What the per-pattern data shows:

Full model (54/44/38/48%) is the *flattest* condition — closer to uniform than any other. This is consistent with the flags actively suppressing the weight differentiation needed to distinguish patterns. Baseline retains strong asymmetry (64/34/62/30%), showing the underlying reward signal does carry some information; the flags are erasing it.

Context vs previous runs:

Original multi-seed (positive flags): Full model mean = 45.5%, std = ±5.0%

This run (flipped flags): Full model mean = 46.0%, std = ±3.2%

The flip made no meaningful difference in mean accuracy, and further collapsed variance — the network is *more* deterministically stuck, not less. Next diagnostic: ablate flags entirely from Full model to confirm they are the culprit.