mcfeedback — Iteration 9: Flag Gate Warmup

experiment-009.mjs · N = 10 seeds · Seeds: 42, 137, 271, 314, 500, 618, 777, 888, 999, 1234 · 1000 training episodes · Frozen-weight evaluation · Random chance = 50%

Change from experiment-004: bypass flag gate for first 200 episodes.

`flagGateWarmup  200  — gate open ep 0–199; flag strength still accumulates throughout`

`flagStrengthThreshold  0.5  (unchanged, enforced from ep 200 onward)`

`flagStrengthGain / flagDecayRate / all other params unchanged from experiment-004.`

Gate schedule across 1000 episodes

ep 1–199 gate open

ep 200–1000 normal flag gate (|flagStrength| ≥ 0.5)

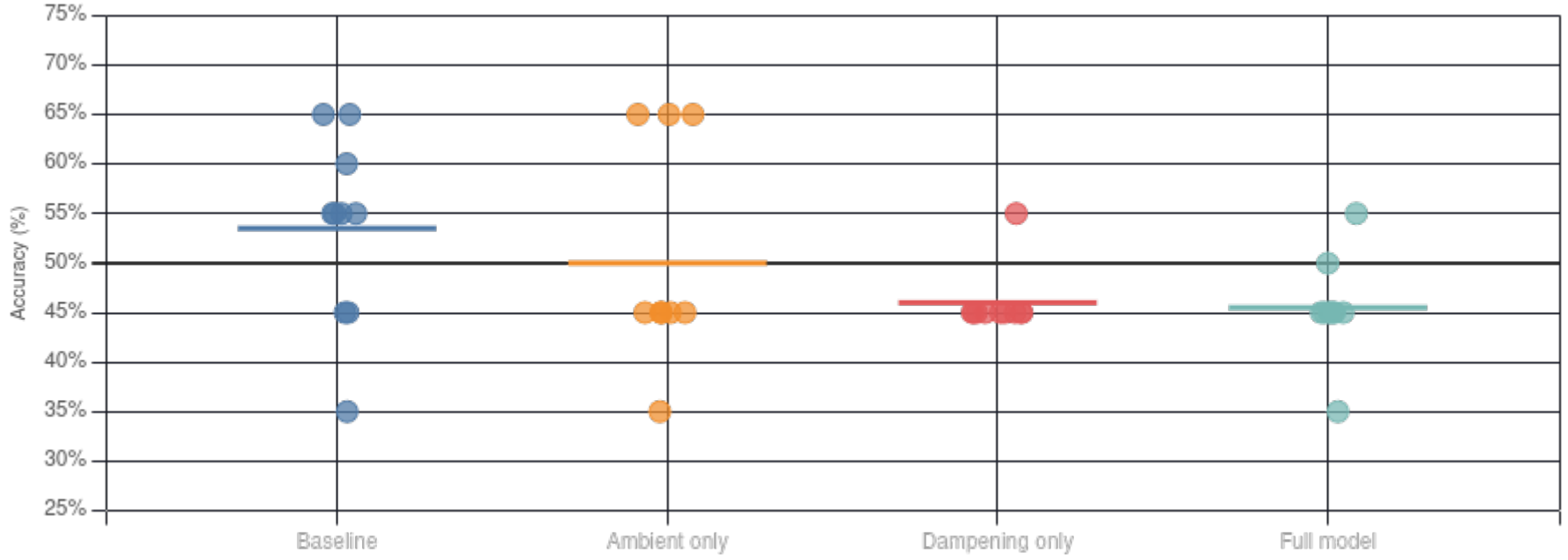
During warmup: all non-zero eligibility traces drive weight updates. Flag strength accumulates normally — flags are pre-charged when the gate switches on.

Verdict: warmup made things significantly worse — p<0.01 for Full model.

Full model dropped to 45.5% mean (vs 53.0% in experiment-004). Dampening only also significantly worse (p<0.05). The 200-episode open gate allowed the network to commit to fixed-output weights before the flag mechanism could provide any selective pressure. When the gate switched on at ep 200, the flags locked in those already-bad weights rather than guiding the network toward pattern-specific solutions. The warmup made bootstrapping *worse*, not better — confirming that **unfiltered early learning is the problem, not the cure.**

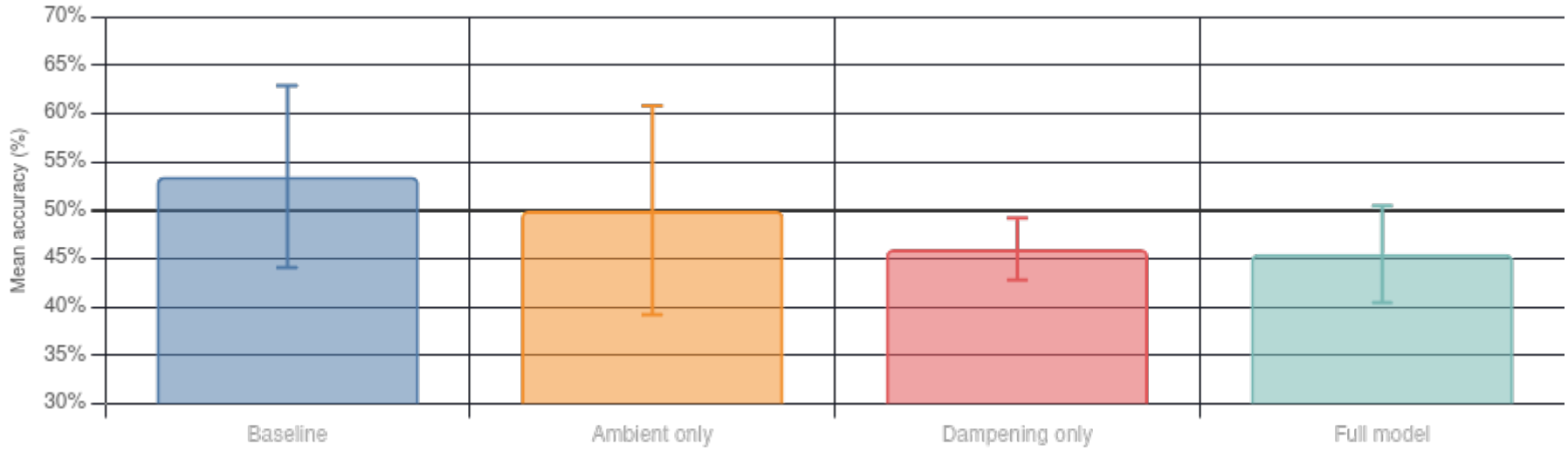
Baseline Ambient only Dampening only Full model

1 — ACCURACY DISTRIBUTION ACROSS SEEDS



Each dot = one seed. Horizontal line = mean. Full model variance collapsed to ±5.0% — network is finding the same poor attractor on nearly every seed.

2 — MEAN ± 1 STD



Error bars show ±1 standard deviation. Full model is significantly below Baseline — the only condition to achieve this (in the wrong direction) twice across all experiments.

3 — PAIRED T-TESTS VS BASELINE

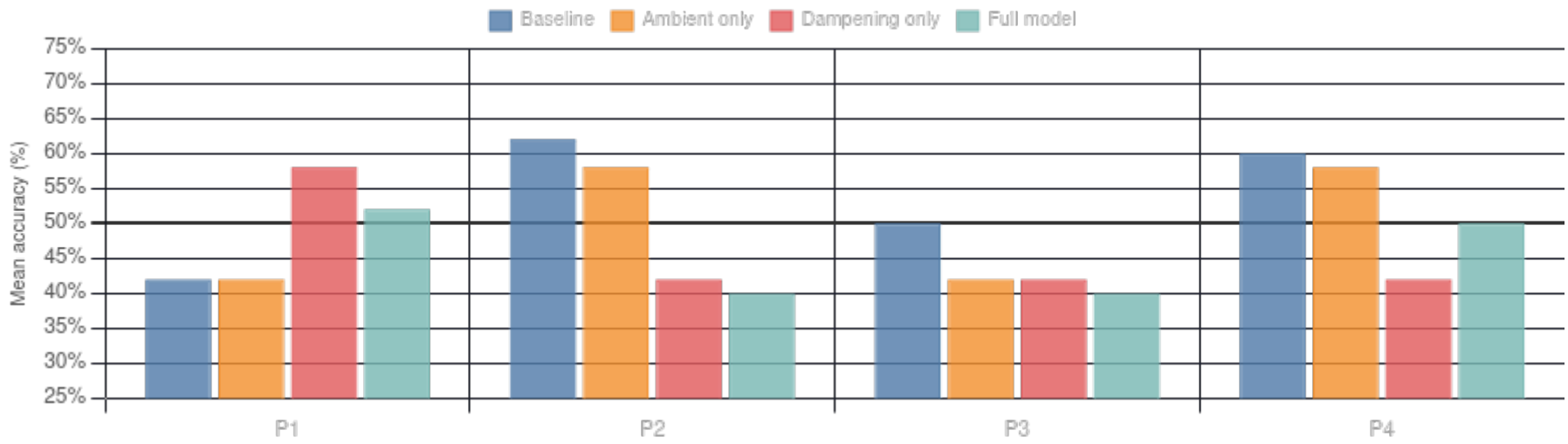
Comparison	Mean diff	t	p	Result
Ambient only vs Baseline	-3.5%	-1.1053	0.2197	**ns**
Dampening only vs Baseline	-7.5%	-2.7643	0.0162	***** **p<0.05**
Full model vs Baseline	-8.0%	-3.3607	0.0062	****** **p<0.01**

Two-tailed paired t-test, df=9. Dampening only and Full model both significantly hurt by the warmup — the open gate undermined both mechanisms.

4 — RAW DATA (ALL SEEDS)

Seed	Baseline	Ambient	Dampening	Full
42	45%	45%	45%	45%
137	65%	65%	45%	45%
271	45%	45%	45%	45%
314	55%	35%	45%	50%
500	55%	45%	45%	45%
618	35%	45%	45%	35%
777	65%	65%	55%	45%
888	60%	45%	45%	55%
999	55%	65%	45%	45%
1234	55%	45%	45%	45%
Mean	**53.5%**	**50.0%**	**46.0%**	**45.5%**
Std	±9.4%	±10.8%	±3.2%	±5.0%

5 — PER-PATTERN ACCURACY (MEAN ACROSS SEEDS)



Full model P2 and P3 collapsed to 40% — below chance. Fixed-output bias committed during the open-gate warmup phase dominates all subsequent learning.

Why the warmup backfired:

Experiment-008 showed flags latch within ~100 episodes regardless. So the 200-episode open gate gave 200 episodes of unfiltered Hebbian + reward learning — exactly the condition that creates fixed-output attractors. By ep 200, the network had already committed to a stable weight matrix. The flags then latched on those bad weights and the gate enforced them, preventing further adaptation. Opening the gate early didn't help the network bootstrap — it gave it 200 extra episodes to dig into the wrong local minimum.

Progress across iterations (Full model mean):

Iter 1 (original): 45.5% max 55%

Iter 3 (flipped signs): 46.0% max 55%

Iter 4 (flag gate): 53.0% max 65% ← best mean

Iter 5 (flag + squared reward): 45.0% max 65%

Iter 6 (flag + anneal): 51.0% max 65%

Iter 9 (flag + warmup): 45.5% max 55% ← regression

Diagnosis confirmed (exp-008): flags saturate by ep 100 on all seeds — the gate is not selective.

Next: lower `flagStrengthGain` to 0.1 (requires ~5+ consistent turns to latch) to make the flag mechanism genuinely discriminating.