mcfeedback — Iteration 5: Graded Reward Signal

experiment-005.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: non-linear reward — squash near-50% accuracy toward zero signal.

`reward = sign(acc−0.5) × |linear|^1 ← was this (linear)`

`reward = sign(acc−0.5) × |(acc−0.5)×2|^rewardExponent`

`rewardExponent 2.0 — squashes the "meh zone" near 50%`

`flagStrengthGain / flagDecayRate / flagStrengthThreshold carried over from experiment-004.`

Reward signal by accuracy — linear vs squared

Accuracy	Linear (exp=1)	Squared (exp=2)	Ratio
55%	+0.10	+0.01	10× weaker
60%	+0.20	+0.04	5× weaker
70%	+0.40	+0.16	2.5× weaker
80%	+0.60	+0.36	1.7× weaker
100%	+1.00	+1.00	equal

Squaring compresses early learning signal — a network making progress at 60% gets 5× less feedback than before.

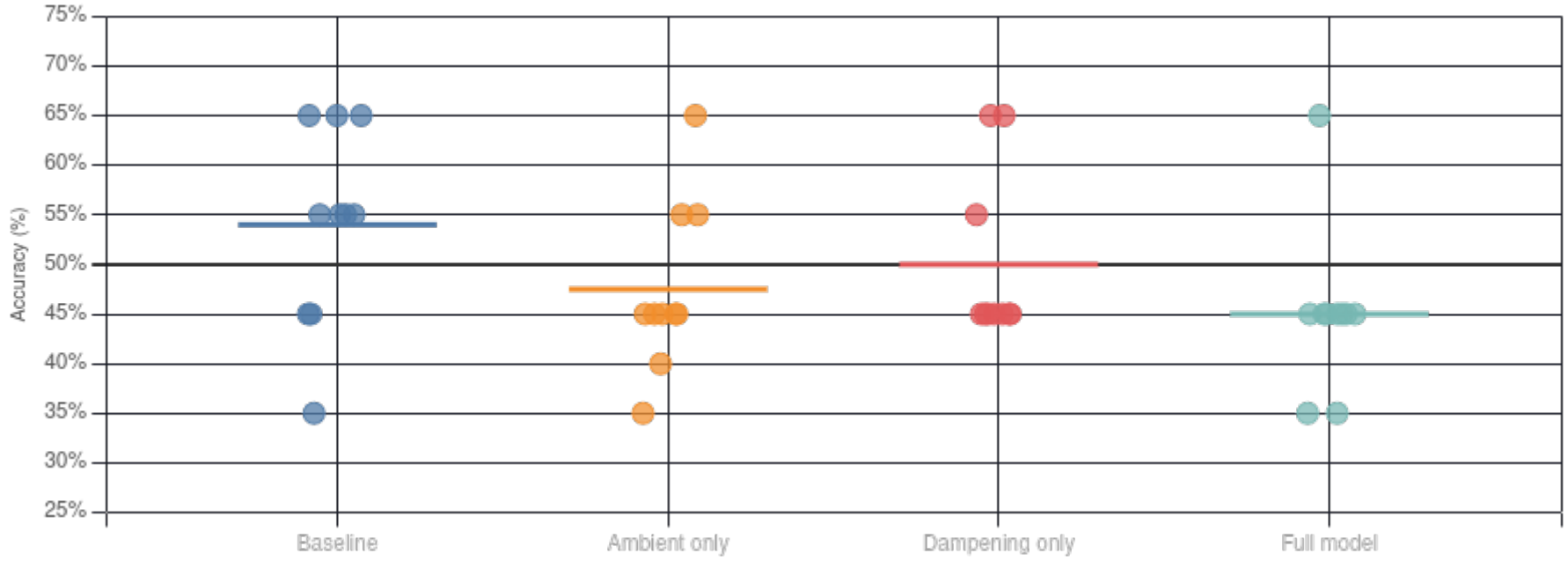
Verdict: hypothesis rejected — squared reward made things worse.

Full model dropped to 45.0% mean and is now *significantly worse than Baseline* (p = 0.03, first statistically significant result — in the

wrong direction). The squared reward didn't just suppress fixed-output strategies; it starved early learning signal before the network could bootstrap. A network at 60% early in training gets reward 0.04 instead of 0.20 — not enough to reinforce the direction of improvement. The flag mechanism (which was working in experiment-004) is effectively disabled because the chemical signal driving it is too weak to accumulate.

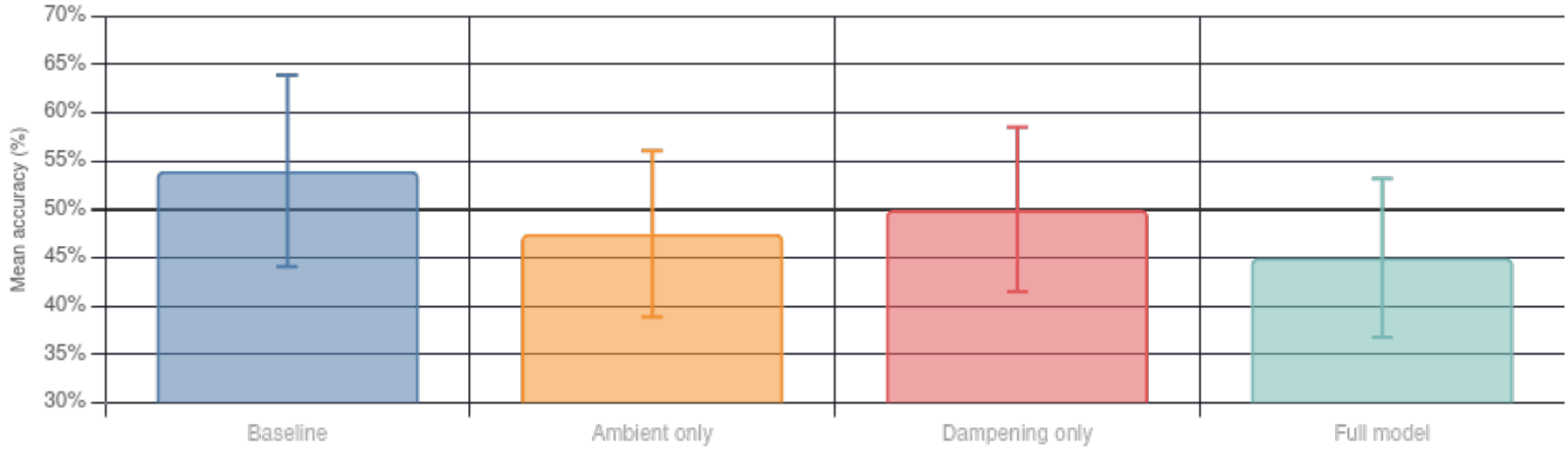
Baseline Ambient only Dampening only Full model

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

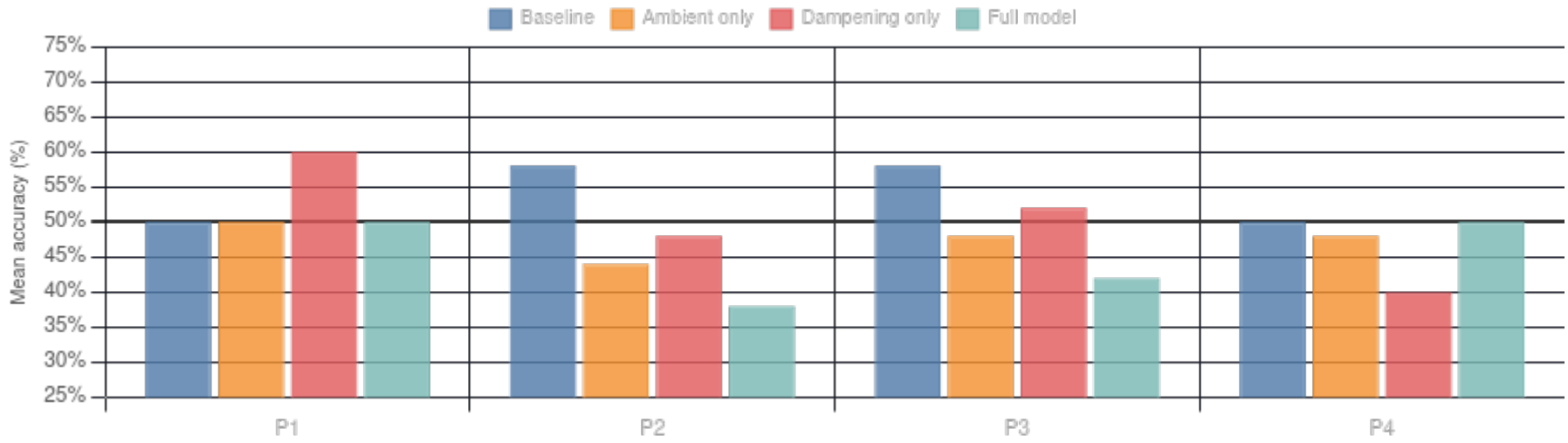
Comparison	Mean diff	t	p	Result
Ambient only vs Baseline	-6.5%	-1.3139	0.1634	**ns**
Dampening only vs Baseline	-4.0%	-1.1767	0.1989	**ns**
Full model vs Baseline	-9.0%	-2.3772	0.0306	***p<0.05**

Two-tailed paired t-test, df=9. First statistically significant result — but in the wrong direction.

4 — RAW DATA (ALL SEEDS)

Seed	Baseline	Ambient	Dampening	Full
42	65%	45%	65%	65%
137	65%	45%	45%	45%
271	55%	55%	45%	45%
314	55%	45%	55%	45%
500	55%	45%	45%	45%
618	45%	65%	45%	45%
777	35%	55%	45%	45%
888	55%	45%	65%	35%
999	65%	40%	45%	35%
1234	45%	35%	45%	45%
Mean	**54.0%**	**47.5%**	**50.0%**	**45.0%**
Std	±9.9%	±8.6%	±8.5%	±8.2%

5 — PER-PATTERN ACCURACY (MEAN ACROSS SEEDS)



Full model P2 and P3 collapsed to 38–42% — below chance on those patterns. The network is actively getting them wrong, suggesting it locked into a fixed bias that happens to invert two of the four patterns.

Why bootstrapping failed:

The flag gate (from experiment-004) requires consistent trace signal over ~2 turns to unlock. That signal accumulates via the chemical (reward) pathway. With reward squared, a network at 60% accuracy emits chemical level 0.04 — too weak to meaningfully update the flag, let alone drive weight changes. The two mechanisms (flag gate + squared reward) interfere: each individually requires signal to bootstrap, and together they form a deadlock.

Progress across iterations:

Iter 1 (original flags, linear reward): Full = 45.5%, max = 55%

Iter 3 (flipped signs): Full = 46.0%, max = 55%

Iter 4 (flag gate, linear reward): Full = 53.0%, max = 65% ← best so far

Iter 5 (flag gate + squared reward): Full = 45.0%, max = 65% — regression

Next: revert to linear reward, focus on tuning the flag gate (iter-004 mechanism).