mcfeedback — Iteration 8: Flag Strength Diagnostic

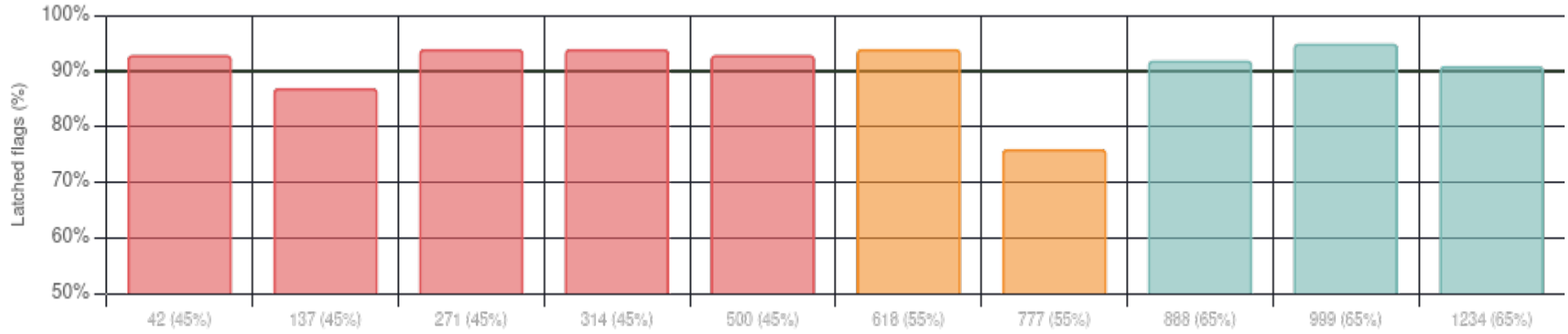
experiment-008.mjs · No mechanism changes · Flag snapshots at ep 100 / 300 / 500 · Output-bound synapses only · Base: experiment-004

Prediction was wrong — and that's the most useful result so far.
The prediction was: good seeds latch flags by ep 100, poor seeds don't. Reality: **both groups have 87–95% of flags latched by ep 100**, with mean |flagStrength| ≈ 0.93 in both. The flag gate opens just as fast on failing seeds as on succeeding ones. The bottleneck is not latching speed.

The real problem: **the flags are not selective — they're saturated.** With `flagStrengthGain: 0.3`, any synapse that sees two consecutive non-zero traces reaches 0.6 and latches. Since output-bound synapses fire on nearly every step (mismatch or co-activation always yields a non-zero trace), they all latch uniformly. The gate is open for 90%+ of synapses — it's not filtering anything.

65% final accuracy 55% final accuracy 45% final accuracy

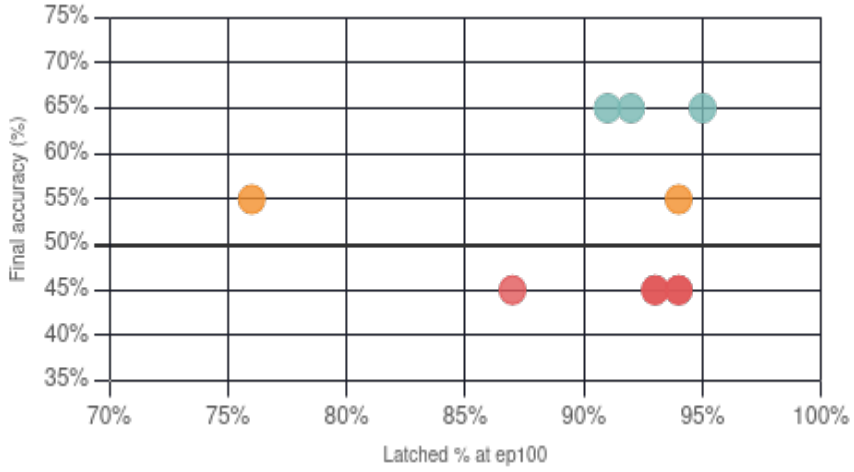
1 — LATCHED FLAG % AT EPISODE 100 (PER SEED)



Coloured by final accuracy. All seeds ≥76% latched by ep 100 — poor seeds are indistinguishable from good. The 50% threshold line shows where useful filtering would begin.

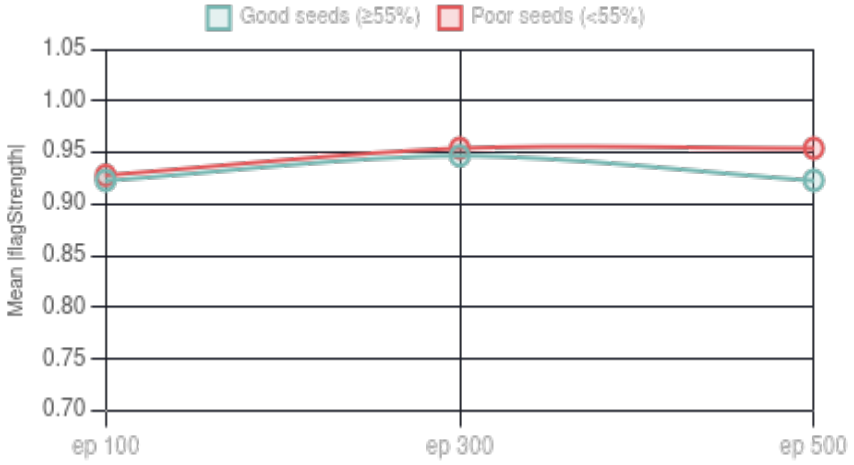
2 — NO CORRELATION: LATCHED % VS FINAL ACCURACY

Latched % at ep 100 vs final accuracy



Flat relationship — more latched flags does not predict better outcome. Poor seeds (45%) have slightly *higher* latch rates.

Mean |flagStrength| across checkpoints



Good vs poor seeds. Both groups plateau near 0.95 by ep 300. No divergence at any checkpoint.

3 — FULL PER-SEED SNAPSHOT DATA

Seed	Final acc	Out-bound syn	Episode 100				Episode 300				Episode 500		
			Latched	Build	Mean|f|	Max|f|	Latched	Build	Mean|f|	Max|f|	Latched	Build	Mean|f|

42	45%	162	151 (93%)	11 (7%)	0.9512	1.000	151 (93%)	11 (7%)	0.9512	1.000	151 (93%)	11 (7%)	0.9512
137	45%	166	145 (87%)	21 (13%)	0.8259	1.000	154 (93%)	12 (7%)	0.9482	1.000	154 (93%)	12 (7%)	0.9482
271	45%	157	147 (94%)	10 (6%)	0.9529	1.000	147 (94%)	10 (6%)	0.9529	1.000	147 (94%)	10 (6%)	0.9529
314	45%	141	132 (94%)	9 (6%)	0.9553	1.000	132 (94%)	9 (6%)	0.9553	1.000	132 (94%)	9 (6%)	0.9553
500	45%	153	143 (93%)	10 (7%)	0.9523	1.000	145 (95%)	8 (5%)	0.9614	1.000	145 (95%)	8 (5%)	0.9614
618	55%	171	161 (94%)	10 (6%)	0.9573	1.000	161 (94%)	10 (6%)	0.9573	1.000	161 (94%)	10 (6%)	0.9573
777	55%	168	127 (76%)	41 (24%)	0.8244	1.000	155 (92%)	13 (8%)	0.9458	1.000	152 (90%)	16 (10%)	0.9333
888	65%	155	142 (92%)	13 (8%)	0.9400	1.000	142 (92%)	13 (8%)	0.9400	1.000	138 (89%)	17 (11%)	0.9219
999	65%	151	143 (95%)	8 (5%)	0.9596	1.000	143 (95%)	8 (5%)	0.9609	1.000	143 (95%)	8 (5%)	0.9609
1234	65%	150	136 (91%)	14 (9%)	0.9333	1.000	136 (91%)	14 (9%)	0.9333	1.000	117 (78%)	33 (22%)	0.8400

4 — GROUP COMPARISON AT EACH CHECKPOINT

Metric	Episode 100			Episode 300			Episode 500		
	Good	Poor	Δ	Good	Poor	Δ	Good	Poor	Δ
Latched (% of output-bound syn)	89.3%	92.3%	−3.0%	92.7%	93.6%	−0.9%	89.3%	93.6%	−4.3%

Building (% of output-bound syn)	10.7%	7.7%	+3.0%	7.3%	6.4%	+0.9%	10.7%	6.4%	+4.3%
Mean |flagStrength|	0.923	0.928	−0.005	0.947	0.954	−0.007	0.923	0.954	−0.031
Max |flagStrength|	1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000	0.000

Poor seeds have *slightly higher* latch rates at every checkpoint — the opposite of the prediction.

What this rules out (and what it reveals):

× Flag bootstrapping speed is not the bottleneck

× Poor seeds don't have fewer latched flags

✓ **The flag mechanism is not selective.** With `flagStrengthGain: 0.3`, any synapse that sees two consecutive non-zero traces (virtually guaranteed since every output-bound synapse fires on almost every step) reaches 0.6 and latches. The gate opens for 90%+ of synapses uniformly — it provides no discrimination between signal and noise.

What needs to change:

The flag mechanism needs to be *harder to latch* — it should only unlock synapses that show sustained, consistent signal over many turns, not just 2. Three candidate fixes (decreasing invasiveness):

1. **Raise `flagStrengthThreshold` to 0.9** — require near-saturation. With gain=0.3, this means ~3 consistent turns minimum, but still latches fast.
2. **Lower `flagStrengthGain` to 0.1** — requires ~5+ consistent turns to reach 0.5 threshold, ~9+ to reach 0.9. Adds meaningful temporal selectivity.
3. **Lower gain to 0.1 + raise threshold to 0.8** — requires ~8 consistent turns with no direction flip. This is the most discriminating option.