Symmetric Hopfield storage, lower endpoint: round 3 status

Six-round multi-agent campaign, orchestrated by Claude Fable 5.1, then Opus 5.5

2026-11-09

Problem and notation

$D = \{s^1, \dots, s^m\}$ is a uniform set of m distinct vectors in $\{-1, 1\}^n$; s_i^μ is neuron i in pattern μ. A network, with symmetric $J \in \mathbb{R}^{n \times n}$, $J_{ii} = 0$, and biases $h \in \mathbb{R}^n$, stores D if every margin $s_i^\mu\big((Js^\mu)_i + h_i\big)$ is positive. With $p_n(m)$ the storage probability and $m = \lfloor un \rfloor$,

$$\boxed{\alpha_- = \sup\{u > 0 :\ p_n(\lfloor un \rfloor) \to 1\}.}$$

Current bounds: $1 \le \alpha_- \le 2$. A **certificate** is a nonzero array $\lambda_{i\mu} \ge 0$ with $\sum_\mu \lambda_{i\mu} s_i^\mu = 0$ for every i and $\sum_\mu (\lambda_{i\mu} + \lambda_{j\mu}) s_i^\mu s_j^\mu = 0$ for $i < j$; one exists exactly when D is not stored (Theorem B). For a set I of neurons, I **is served** if some network makes every margin positive on the neurons of I (all n neurons still feed in). A **symmetry-only failure** is a set every neuron of which can separate its own labels, yet no symmetric network stores it.

What the numerics say

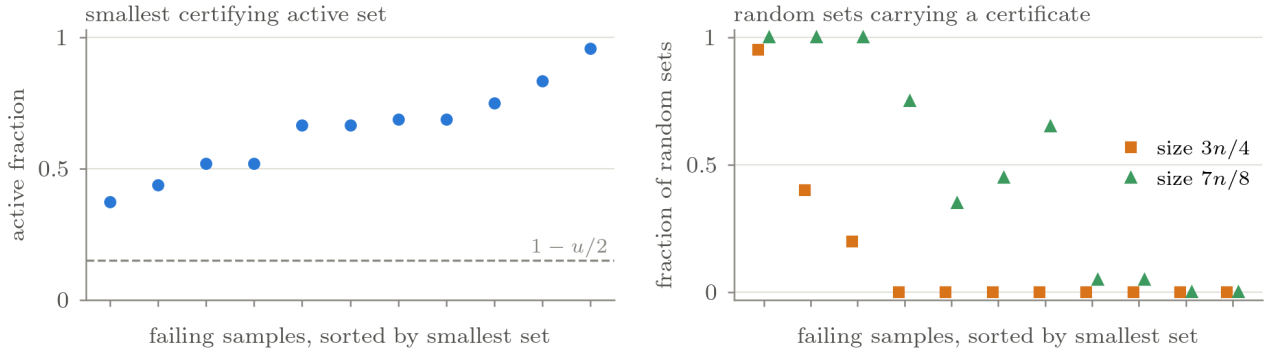


Figure 1. Where failure certificates live, $n = 48$, $u = 1.7$. Left: for each of the 11 certified symmetry-only failures, the smallest fraction of neurons found that still carries an exact certificate. Right: the fraction of 20 random neuron sets of size $3n/4$ or $7n/8$ that are not served, each with an exact certificate.

The campaign

Round 3 was the falsification round. Two solver–verifier lanes tested named claims; a computational verifier, C1, re-checked the experimenter; and V2 finished checking lane S2. On the user's instruction the campaign moved from Claude Fable 5.1 to Claude Opus 5.5 at 00:20 UTC. Running Fable agents were stopped and relaunched on Opus with a fresh read of the problem, and the horizon became six rounds. **Round 3 took 4.1 hours and cost 269.3 M tokens, about $161 at list prices** (Table 1).

Table 1. Round 3 by lane. Tokens count input, cache and output in every transcript, including runs stopped at the model switch; dollars use list prices.

Lane	Route	Round 3 result	Tokens	Cost
S3 ↔ V3	restricted certificates	Theorem I; Figure 1	155.7 M	$60
S4 ↔ V4	sharp threshold	Theorems J, K; Correction 2	22.1 M	$32
X ↔ C1	experiments, re-checked	13 of 14 round-2 nodes confirmed	33.8 M	$21
S2 ↔ V2	round-2 check, finished	Correction 1	33.7 M	$31
Orchestrator	routing and ledger		24.1 M	$17

Round 3 results

Theorem I (Restricted service is monotone and exchangeable; S3-27)**.** If $I \subseteq I'$ and I' is served, then I is served. The probability that a set I is served depends only on $|I|$, also conditionally on symmetry-only failure.

Meaning. Serving fewer neurons is easier, and which neurons does not matter in law. So "some set of fraction β is not served" is the same kind of event for every fixed set of that size.

Contribution to α_-. If a fixed set of $\lfloor 3n/4 \rfloor$ neurons fails to be served with probability bounded below at $u = 1.7$, then $\alpha_- \le 1.7$. Figure 1 shows this happening at $n = 48$ on 3 of 11 failing samples; whether it persists as n grows is the open estimate.

Theorem J (Threshold framework; S4-14)**.** Nonstorage is an increasing property, invariant under every permutation and sign change of coordinates, which acts transitively on patterns. Its threshold windows in the uniform model and in the model that includes each pattern independently with probability $p = m/2^n$ agree up to $o(1)$ in load. Let $U_n(1/2)$ be the load at which $p_n = 1/2$. If the window has width $o(1)$, then $\alpha_- = \liminf_n U_n(1/2)$ and $\alpha_+ = \limsup_n U_n(1/2)$.

Meaning. α_+ is the smallest load at which storage almost surely fails. A sharp window means storage drops from near 1 to near 0 over a vanishing range of loads.

Contribution to α_-. With a sharp window and a convergent center, $\alpha_- = \alpha_+$; then any proof of failure at one load bounds α_-.

Theorem K (Failure needs many patterns; S4-15)**.** Fix $u \le 2.1$. With probability tending to one, every certificate is supported on more than $n/8$ patterns.

Meaning. No bounded set of patterns can witness failure. This removes one of the two alternatives in Bourgain's sharp-threshold theorem.

Contribution to α_-. It is one of the two inputs needed for a sharp window; the other is Correction 2's open condition.

Correction 1 (V2 on Theorem H). The dead-end theorem holds when the starting set is chosen without looking at the patterns and its slack satisfies $\varepsilon < u - 1$. A start with large slack, or one chosen from the data, is not covered. At $n = 64$ the full linear program stores all 12 reference samples at load 1.3.

Correction 2 (V4 on S4-16, S4-17). Bourgain's dichotomy was quoted with a fixed size bound $10C$. That version is false at levels below 0.106; V4 gives an explicit counterexample. His printed proof gives a valid version with size bound $K(\varepsilon)C$, $K(\varepsilon) > 1/(\varepsilon(1-\varepsilon))$. On that basis the remaining condition for a sharp window is *bounded no-boosting*: no storable set of at most $K(\varepsilon)C$ patterns raises the failure probability by a fixed amount near the threshold. S4 rewrites both nodes in round 4.

Experiments re-checked (C1). Every sample regenerates bit for bit, and every certificate passes an independent checker. Two cautions: the inverse-covariance test of round 2 carries no certificates, and census samples at different loads are nested, so comparisons across loads are not independent.

Plan

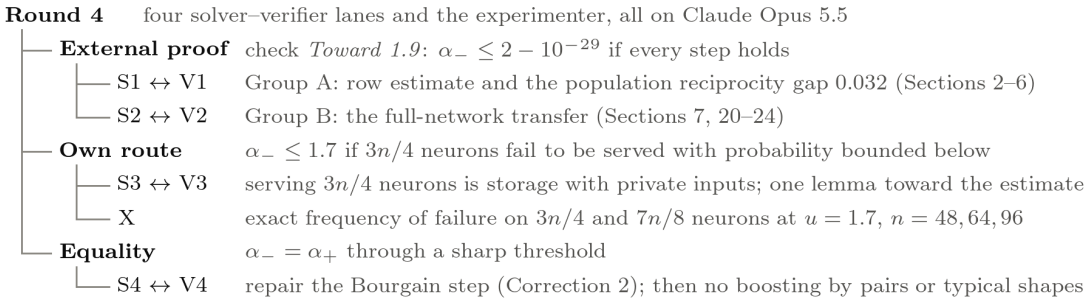


Figure 2. Round 4 plan. Tasks grouped by the goal they serve.

Upper bound, external proof. Goal: decide whether the argument of *Toward 1.9* (a reference supplied during round 4) is correct. It claims storage probability at most 10^{-3} at load $2 - 10^{-29}$, hence $\alpha_- \le 2 - 10^{-29}$. Near load 2 each neuron's feasible weights form a narrow cone, so its row is nearly forced, and forced rows disagree too much to be symmetric. The proof splits into four groups. S1 rebuilds the row estimate and the population reciprocity gap; S2 rebuilds the transfer to the full network, the step a published critique calls the most delicate. Groups C and D follow in round 5. A strict improvement this small still meets the campaign's goal.

Upper bound, own route. Goal: a fixed set of $3n/4$ neurons is not served, with probability bounded below at $u = 1.7$ (Theorem I then gives $\alpha_- \le 1.7$). Serving I is symmetric storage on I in which each neuron also has $n - |I|$ private input weights: its couplings to the other neurons, which no other constraint uses. S3 proves this and attempts one lemma toward the estimate; X measures the exact frequency at $n = 48, 64, 96$.

Endpoint equality. Goal: a sharp window. S4 rewrites the Bourgain step on the corrected bound, then attacks bounded no-boosting for pairs at fixed Hamming distance and for sets with typical distances.

Rounds 5 and 6. Round 5 rebuilds Groups C and D and checks Groups A and B. Round 6 checks the rest, makes bounded repairs, and assembles the theorem from checked nodes; a fresh audit by new agents, filing and the write-up follow.

Appendix C. Sharp thresholds for symmetric properties

Theorems J and K use the theory of sharp thresholds for monotone properties of random sets.

Setting. Let $\Omega = \{-1, 1\}^n$ with $N = 2^n$ points, and include each point in a random set D independently with probability p. A property A of D is **increasing** if $D \in A$ and $D \subseteq D'$ imply $D' \in A$; nonstorage is increasing because deleting a pattern never breaks storage. It is **symmetric** if a group acting transitively on Ω preserves it; here the group is the $2^n n!$ coordinate permutations and sign changes. The threshold window is the range of p over which $\mu_p(A)$ climbs from ε to $1 - \varepsilon$.

Friedgut–Kalai. For a symmetric increasing property the window has width $O\big(p \log(1/p)/\log N\big)$ [FK96, Thm 2.1]. Here $p \approx un/2^n$, so the bound is of order p itself: it does not decide whether the window is sharp in load, which is why Theorem J's framework is needed.

Bourgain. If the window is coarse, meaning $p \frac{d}{dp}\mu_p(A) < C$ at a point where $\mu_p(A) \in [\varepsilon, 1 - \varepsilon]$, then one of two things happens [B99, Prop. 1]: (W) with probability bounded below, D contains a bounded set x' that by itself forces A; or (B) some bounded x' not forcing A raises $\mu_p(A \mid x' \subseteq D)$ above $\mu_p(A)$ by a fixed $\delta > 0$. The size bound in the theorem is $10C$ at level $1/2$; at a general level ε the printed proof gives $K(\varepsilon)C$ with $K(\varepsilon) > 1/(\varepsilon(1 - \varepsilon))$ (Correction 2).

How the campaign uses it. Theorem K excludes (W): a bounded set of patterns never certifies failure. Excluding (B), bounded no-boosting, would make the window sharp, and with a convergent center $U_n(1/2)$ would give $\alpha_- = \alpha_+$ (Theorem J). A single random pattern is exactly neutral in the uniform model (S4-14), so a boosting set must have atypical structure, such as two patterns at small Hamming distance. Patterns at Hamming distance 1 can never both be stored: the neuron where they differ sees identical inputs but must output opposite signs. That is a bounded witness of type (W), and it occurs with probability $O(n^3 2^{-n})$.

Literature. Friedgut and Kalai [FK96] proved that symmetric monotone properties have thresholds of width $O(1/\log N)$ in p. Friedgut's theorem with Bourgain's appendix [B99] characterizes coarse thresholds by bounded boosting sets, and gave the sharp threshold for random k-SAT. Bollobás and Thomason [BT87] showed that every monotone property has a threshold. Talagrand [T94] and Kahn–Kalai–Linial [KKL88] supply the influence inequalities behind these results.

References

[B99] J. Bourgain. Appendix to E. Friedgut, Sharp thresholds of graph properties, and the k-SAT problem. *J. Amer. Math. Soc.* 12 (1999), 1017–1054.

[BT87] B. Bollobás and A. Thomason. Threshold functions. *Combinatorica* 7 (1987), 35–38.

[FK96] E. Friedgut and G. Kalai. Every monotone graph property has a sharp threshold. *Proc. Amer. Math. Soc.* 124 (1996), 2993–3002.

[KKL88] J. Kahn, G. Kalai and N. Linial. The influence of variables on Boolean functions. *Proc. 29th FOCS* (1988), 68–80.

[T94] M. Talagrand. On Russo's approximate zero-one law. *Ann. Probab.* 22 (1994), 1576–1587.