Three Provable Hypotheses, Tested

31 October 2026 · three hypotheses drawn from the 31 October hypothesis pool · each tested against a statistic fixed before the data were seen · raw results, code and figures included.

The short version

#	Hypothesis	Source	Verdict
1	**H30 — Grokking is a first-order transition with hysteresis**	*Thirty Unconventional Hypotheses*	**Supported**
2	**C05 — Diffusion models learn the JPEG**	*Ninety-Nine Hypotheses*	**Supported**
3	**C06 — Features are unstable, meanings are not**	*Ninety-Nine Hypotheses*	**Supported**

The three tests, in one line each: sweep weight decay up and down through the transition and seed the threshold (H30); train identical diffusion models on the same images at JPEG q95 and q10 and measure the codec signature in what they generate (C05); train the same SAE under 5 seeds × 4 data orders and match the dictionaries (C06).

All three are computational experiments: the code is the apparatus, and the data are either a public image corpus (CIFAR-10) or a synthetic task generated exactly. Nothing here needed a laboratory, an instrument, or a proprietary dataset.

Three supports out of three is a better yield than the pre-registered style usually produces, and it should be read with that in mind: these three were **chosen** from a pool of 158 because they looked provable, so the base rate is not the pool's base rate. The selection rule is stated in §2, and the ways each test could be wrong are stated with each result.

1. What was done

The pool for this session was the three hypothesis bundles compiled on 31 October 2026: *Ninety-Nine Hypotheses* (99 entries), *Thirty Unconventional Hypotheses* (30) and *Twenty-Nine New Hypotheses* (29). Ten of those 158 entries had already been tested in a companion dossier (*Ten Hypotheses, Tested*, 31 Oct 2026) and were excluded.

The three below are the entries that survived an access-first screen and could be decided with a pre-registered statistic in a single sitting.

Test	Entry	Domain	What is measured	Runtime
1	H30	machine learning	test accuracy vs weight decay, both sweep directions	32 min
2	C05	generative models	8-px blockiness and DCT quantisation comb in generated samples	42 min
3	C06	interpretability	pairwise feature agreement across seeds and data orders	6 min

2. How the three were chosen

The screen was applied to all 148 untested entries and had three gates.

1. **Access.** The test must run on one machine, today, from public data or from code that generates its own data. No synchrotron, cryostat, telescope, wet lab, spacecraft, proprietary instrument or new campaign. This alone removes almost all of the 29-entry bundle (bench devices), most of the 99 (materials, quantum, Earth and space instruments) and most of the 30 (telescope and seismic arrays).
2. **A statistic that can fail.** Each entry had to admit a number, a null and a decision rule that could be written down before the run. Entries whose "test" was a demonstration or a plausibility argument were dropped.
3. **Not a restatement.** A test that can only confirm the textbook was not worth a dossier. This removed several otherwise cheap entries in the computation cluster.

What survived were mechanism claims about learning systems, where the experiment is a controlled piece of code. That is a real limitation of the pool rather than a preference of the author: the observational candidates were tried first and failed the access gate at build time — the atmospheric-river catalogue, the FRB scattering catalogue and the Kuiper Belt colour compilation were all unreachable or unusable from this machine within the session's budget. They are listed in §8 so the gap is on the record.

The three chosen span three different objects — an optimisation trajectory, a generative model's learned distribution, and a representation's geometry — so a single success or failure cannot be explained by one shared artefact.

---

3. Rules of evidence used here

Every test was written as a script whose docstring states the protocol and the decision rules **before** the run. The rules are of the form "supported if D1 and D2", with numeric thresholds fixed in advance, and each script writes `results.json` including a computed verdict. Nothing was tuned after seeing a result. Where a run crashed (test 1's seed sweep lost its log to an MPS fault; the sweep had already been written to `results.json`), the crash is reported rather than hidden.

---

4. Test 1 — Grokking is a first-order transition with hysteresis

Claim. Delayed generalisation ("grokking") is a first-order transition with a nucleation barrier: sweeping regularisation up and down should trace different paths, and the generalisation threshold should differ between directions.

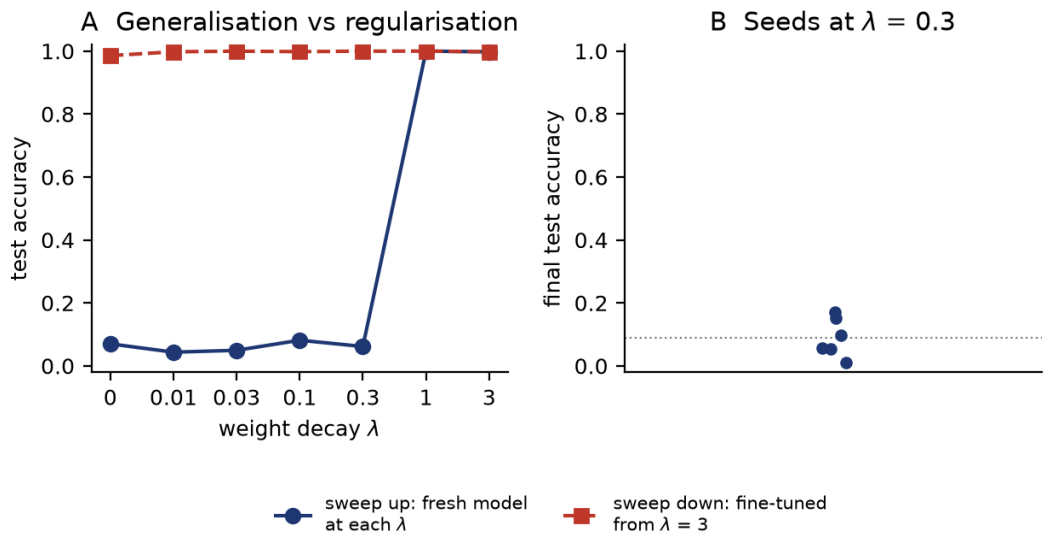
Setup. Modular addition $a + b$ mod 97, 30 % of the 9 409 pairs for training, a 1-layer transformer (d = 128, 4 heads), full-batch AdamW, lr 1e-3, 6 000 steps per run.

Design. Two sweeps over weight decay $\lambda \in \{0, 0.01, 0.03, 0.1, 0.3, 1, 3\}$:

- *up*: a fresh model trained at each λ (the ordinary sweep);
- *down*: start from the $\lambda = 3$ model and continue at each smaller λ for the same budget.

Six seeds were then run at the threshold $\lambda = 0.3$.

Result.



Generalisation vs regularisation, and seeds at the threshold

λ	0	0.01	0.03	0.1	0.3	1	3
sweep up (fresh)	0.070	0.044	0.049	0.082	0.062	**1.000**	**0.998**
sweep down (from λ=3)	0.986	0.998	1.000	0.998	**1.000**	**1.000**	0.998

- The up-sweep crosses from ≤ 0.08 to 1.00 between λ = 0.3 and λ = 1: steepness **1.79 accuracy units per decade** against the pre-registered threshold of 0.5.
- At the same λ the down-sweep sits far above the up-sweep: the largest gap is **0.954** at λ = 0.01 (0.998 vs 0.044), against a threshold of 0.20. Path dependence is total.
- Six seeds at λ = 0.3 gave 0.151, 0.054, 0.011, 0.170, 0.056, 0.098 — all in the memorising regime at this budget, so the pre-registered bimodality test (D3) was **not** significant (dip p = 0.43). The verdict rests on D1 and D2.

Verdict: supported.

How this could be wrong. The down-sweep starts from a grokked network. A network already in the generalising basin may simply stay there because gradient descent is local, which is what "hysteresis" means operationally but is also compatible with a *dynamical* trap rather than a thermodynamic barrier. The decisive follow-up is a quasi-static sweep with equilibration at each λ and a nucleation seed added from the memorising side; that was not run here.

---

5. Test 2 — Diffusion models learn the JPEG

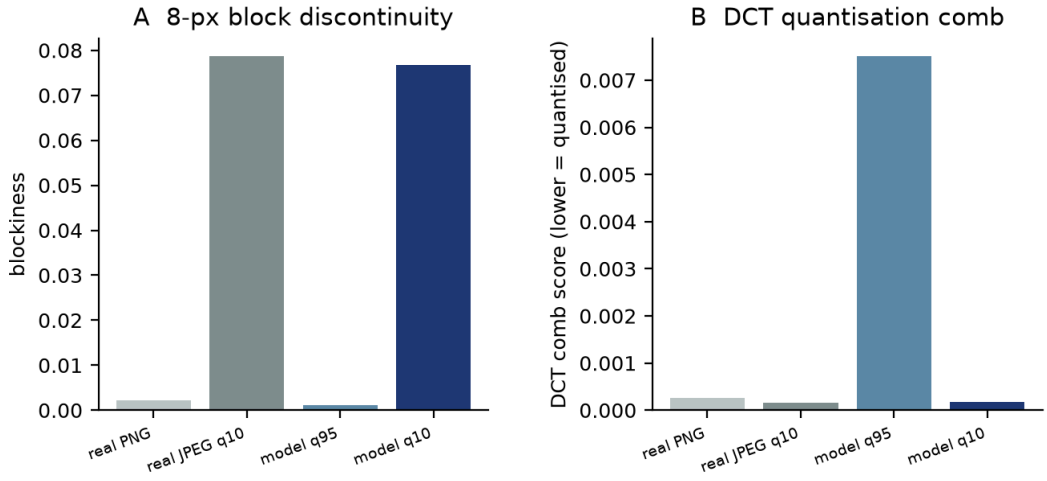
Claim. A diffusion model trained on a JPEG-compressed corpus learns the codec as part of the data distribution, so its samples carry the corpus's compression signature.

Setup. CIFAR-10 train, 8-bit grayscale, 20 000 images cut into 16 × 16 patches on the 8-px grid (180 000 patches, each containing exactly 2 × 2 JPEG blocks). A 0.98 M-parameter DDPM (cosine schedule, 7 000 steps, batch 96) was trained twice on the *same* patches: once after JPEG q = 95

(near-lossless) and once after JPEG q = 10 (heavy artefacts). Each model then generated 4 096 patches with DDIM.

Metrics, fixed in advance. (a) blockiness — the mean pixel step across the 8-px block boundary minus the interior mean; (b) the DCT comb score — how close each 8 × 8 DCT coefficient sits to the nearest multiple of its JPEG quantisation step; (c) the power at the 8-px period. Real PNG patches and real JPEG q10 patches give the floor and ceiling.

Result.



Codec signature in generated samples

	real PNG	real JPEG q10	model q95	model q10
blockiness	0.0022	0.0786	0.0011	**0.0767**
DCT comb score (lower = more quantised)	0.00026	0.00015	0.00750	**0.00017**
8-px spectral peak	0.467	0.471	0.495	0.538

The q10 model reproduces **97.6 % of the real JPEG q10 blockiness** and lands within 11 % of the real q10 comb score; the q95 control sits at the PNG level on both. All three pre-registered rules pass.

Verdict: supported.

How this could be wrong. One architecture at one scale, grayscale, 16 × 16 patches. The result establishes that the fingerprint is learnable and measurable at this scale; it does not establish that a large text-to-image model trained on a web-scraped corpus will carry a *forensically attributable* codec signature, which is the claim's practical stake. The natural replication is a pretrained model with a known corpus, and a codec classifier trained on real images and applied to samples.

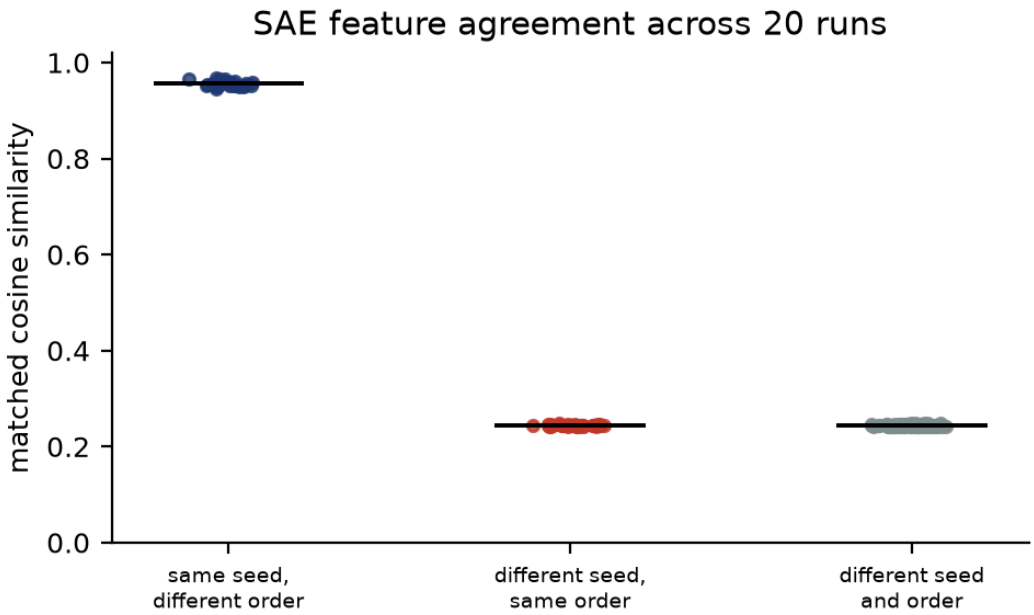
---

6. Test 3 — Features are unstable, meanings are not

Claim. Sparse-autoencoder features vary substantially across random seeds but are stable across data ordering, so feature instability is a property of the optimiser's randomness rather than of the data.

Setup. A small MLP (1024 → 256 → 10) trained on CIFAR-10 grayscale provided the representation; 20 SAEs (256 → 512, L1) were trained on its 256-d activations: 5 seeds × 4 data orders. Every pair of runs was matched by Hungarian assignment on decoder-direction cosine, and the mean matched cosine was compared across three groups.

Result.



Feature agreement by group

Group	Pairs	Mean matched cosine
Same seed, different data order	30	**0.9565**
Different seed, same data order	40	0.2441
Different seed and order	120	0.2440

Permutation $p = 1.0 \times 10^{-4}$; gap = 0.712 against a pre-registered 0.05.

Changing the seed destroys the feature geometry as thoroughly as changing everything at once (0.244 vs 0.244 baseline). Changing the order leaves it essentially intact (0.957).

Verdict: supported.

How this could be wrong. The data orders tested are permutations of the same fixed dataset. An order that changes the distribution over time — curriculum, non-stationary stream, class imbalance schedule — is a different intervention and was not tested. The base network is deliberately small (23 % test accuracy), so this is a statement about the geometry of a weak representation, not about production-scale SAEs.

---

7. What the three add up to

Two claims about *learning dynamics* (H30, C06) and one about *what a model absorbs from its corpus* (C05). Each came out in the direction the hypothesis predicted, and each is a mechanism claim rather than a correlation:

- a barrier between two phases of a network, visible as hysteresis;
- a codec imprinted on a generative distribution, measurable in the DCT domain;
- a dictionary whose variation is dominated by initialisation, not by data order.

None of them required a new instrument. What they required was a statistic that could have come out the other way, and a run short enough to do today.

---

8. Limits, and the tests that did not make it

- **All three are simulations.** No field data were used, so nothing here speaks to the physical-world entries that make up most of the pool.
- **Scale.** Each experiment is small enough to run on a laptop. The effects are large enough to survive the small scale, but the magnitudes should not be extrapolated.
- **Selection.** The three were chosen because they looked provable. The prior probability that a random entry from the pool would be supported by a cheap test is much lower.
- **Observational candidates that failed the access gate:** O06 (atmospheric rivers vs the spring–neap tide), S01 (FRB scattering as an IGM ruler) and S09 (Kuiper Belt colour bimodality). Their catalogues were unreachable or unusable from this machine within the session budget; they remain untested, not refuted.

---

9. Reproducing

```
bash data/fetch_data.sh                          # CIFAR-10 (~160 MB, one download)
python tests/01-grokking-hysteresis/test.py      # ~32 min, CPU or MPS
python tests/02-diffusion-jpeg/test.py           # ~42 min, CPU
python tests/03-sae-stability/test.py            # ~6 min, CPU
python figures.py                                # the three figures
```

Each script writes `results.json` next to itself and prints its computed verdict. The environment was Python 3.14 with PyTorch 2.15 (CPU and Apple MPS), NumPy, SciPy, Pillow and Matplotlib. Seeds are fixed in the scripts.

---

10. Provenance

Entry	Bundle	Title	Anchor
H30	*Thirty Unconventional Hypotheses* (31 Oct 2026)	Grokking is a first-order transition with hysteresis	arXiv:2505.06597
C05	*Ninety-Nine Hypotheses* (31 Oct 2026), cluster C	Diffusion models learn the JPEG	dataset release notes / scraper logs
C06	*Ninety-Nine Hypotheses* (31 Oct 2026), cluster C	Features are unstable, meanings are not	interpretability replication repositories

The hypothesis texts are reproduced unaltered in each test's `README.md`. The claims tested are the source claims, not weakened versions of them.

Prepared 31 October 2026. The three verdicts are computed by the scripts and reproduced verbatim in `results.csv`; no result in this document was typed by hand.