EWC stiffness freezes stale facts, and no per-fact Fisher ledger can release it

W. Schulz

Nov 10, 2026

Abstract

A pre-registered toy test of elastic weight consolidation on contradicting fact updates, with an exact Fisher. Stiffness raised unrelated retention from 44.3% to 92.0% but cut edit success from 62.5% to 13.5% and left 81.7% of paraphrases stale. A KL anchor at equal retention learned far more, and subtracting a superseded fact's own Fisher did nothing measurable.

W. Schulz, NeuroGenesys. *Pre-registered 2026-11-10 after a dev pilot on separate seeds, run once on 5 test seeds. Toy scale: a 2-layer transformer (about 0.35M parameters) on a synthetic knowledge base. This note reports the frozen analysis, including three predictions that failed.*

Summary

Elastic weight consolidation (EWC) protects old knowledge by making important weights stiff. We asked what that stiffness does when new facts *contradict* old ones. The EWC literature review made three predictions: EWC should protect a stale fact because the fact sits in its Fisher data ("intransigence by inheritance"); the update should route around stiff weights, so direct edits land while paraphrases and consequences keep the old answer ("bypass, not blockage"); and subtracting a superseded fact's own Fisher contribution (a "Fisher ledger") should release it.

- Stiffness protected unrelated facts and froze the facts being revised (H2 supported). At the dev-chosen strength $\lambda = 1e5$, unrelated retention rose from 44.3% to 92.0%, while edit success fell from 62.5% to 13.5% and stale paraphrases rose from 10.8% to 81.7%.
- Stiffness did not widen the gap between a direct edit and its consequences (H1 gap rejected). It lowered both together: edit -19.0 and paraphrase -20.2 pp per decade of λ.
- EWC was far more intransigent than the function-space anchor it approximates. At matched retention, a KL anchor on old queries kept edit success 55.5 pp higher.
- The Fisher ledger was a clean null (H4 rejected). Releasing the superseded fact's Fisher and its 2-hop dependent's changed stale paraphrases by -1.0 ± 1.3 pp, with edit success identical. Random release and global decay at the same mass did exactly as well.

The reason is structural. A superseded fact supplies only about 0.41% of the stiffness its own correction has to push against. The rest belongs to the other facts that share the same weights, so EWC's resistance to contradiction cannot be accounted away fact by fact.

Design

Item	Choice
Model	2-layer pre-LN transformer, d = 128, about 0.35M parameters, 98 single-token answers, pretrained to 100%
Knowledge base	400 persons (employer, city, hobby), 40 companies (hq), 2-hop workcity = hq(employer)
Stream	228 new facts in 8 stages, one Adam optimizer (lr 1e-4, 60 steps per stage, batch 64), including 40 contradicting employer updates whose workcity also changes
Consequences	For a fixed half of updated persons the new workcity is trained alongside the update; the other half is never trained ("held")
Fisher	The exact diagonal Fisher of the pretrained model on the whole old KB: every base fact in both templates (3,200 examples), all 98 classes weighted by model probability
λ grid	0, 1e3, 1e4, 3e4, 1e5, 1e6
Seeds	Dev seeds 100 and 101 for every decision; test seeds 0–4 never used for any choice
Compute	CPU only; 140 test and 46 dev runs in about 2 h wall-clock

With $\lambda = 0$ and no consequences the loop reproduces the earlier seven-methods study's naive run exactly. The exact Fisher matches brute force to 1.6e-6 relative L1, and per-fact additivity holds to 3e-6. A 4,000-sample Monte Carlo Fisher underestimated the mass by 63% on a confident example (p_max = 0.999), because a confident model's Fisher lives in its rare classes.

The ledger-arm strength $\lambda_B = 1e5$ was chosen on the dev seeds by a pre-registered rule, maximising mean(retention, edit success). Two KL anchors to the frozen pretrained model, on 32 random old queries per step, had their strengths matched on dev to EWC's retention: KL-all includes the superseded facts' queries, making it the function-space twin of the old-KB Fisher, and KL-ret excludes them.

Metrics, on the 40 updated persons: edit success on the trained phrasing; accuracy on two never-trained paraphrases; 2-hop workcity accuracy; stale rates (the old answer is the model's top answer); the log-prob gap log p(old) − log p(new) in nats; and retention of old facts about everyone else. Intervals are 95% t-intervals over 5 test seeds; paired tests are within seed and uncorrected.

Hypotheses. H1: as λ rises, stale paraphrase and 2-hop rates rise (H1b), the edit–consequence gap widens (H1a), and the new answer is written beside the old one (H1c). H2: unrelated retention rises with λ. H3: EWC behaves like the old-query KL anchor at matched retention. H4: the ledger beats both full EWC and mass-matched random release by at least 5 pp on stale rates. Kill rules K1 (gap slope below +3 pp per decade), K2 (ledger vs random below 5 pp) and K3 (edit success below 30% at λ_B, in which case high-λ stale rates are read as blockage) were fixed in advance.

Results

Stiffness freezes everything the update should change

λ	Unrelated retention	Edit success	Paraphrase acc.	2-hop acc., trained half	2-hop acc., held half	Stale paraphrase	Stale 2-hop, held
0	44.3 ± 8.2	62.5 ± 13.5	62.0 ± 10.3	56.1 ± 18.4	13.4 ± 11.3	10.8 ± 1.4	32.0 ± 12.4
1e3	49.8 ± 10.9	61.5 ± 16.2	59.2 ± 11.6	53.6 ± 18.9	12.1 ± 9.4	14.0 ± 2.6	36.7 ± 14.1
1e4	70.4 ± 14.9	43.0 ± 19.4	33.0 ± 16.0	35.1 ± 10.7	9.2 ± 8.5	45.2 ± 24.9	54.6 ± 23.3
3e4	84.5 ± 8.4	23.5 ± 12.1	12.8 ± 8.5	21.0 ± 9.7	4.1 ± 8.3	71.5 ± 19.1	73.8 ± 16.8
1e5	92.0 ± 5.4	13.5 ± 5.6	5.5 ± 3.7	10.3 ± 7.0	1.8 ± 3.8	81.7 ± 15.1	89.3 ± 6.9
1e6	97.9 ± 2.7	8.0 ± 4.0	1.0 ± 2.0	5.6 ± 7.3	0.4 ± 1.0	96.7 ± 5.9	97.9 ± 2.9

Per decade of λ (per-seed regression, 1e3–1e6), unrelated retention rose +16.5 ± 3.9 pp, stale paraphrases +28.4 ± 2.3 and stale held 2-hop answers +21.8 ± 5.8. The stale half of H1 holds.

The gap half failed its kill rule. The edit-minus-paraphrase gap moved only +1.2 ± 2.5 pp per decade, against the +3 threshold, and the edit-minus-held-2-hop gap *narrowed* by 14.7 ± 6.6 pp per decade. Direct edits, paraphrases and trained consequences fell at about the same rate (−19.0, −20.2 and −16.8 pp per decade). Spontaneous 2-hop propagation was near the floor even without EWC (13.4%), so it had little room to fall. EWC does not let edits through while freezing their consequences; it stops the edit itself.

The old answer survives beside a successful edit, at moderate stiffness

At λ_B only 3–7 of 40 edits per seed landed, which triggered K3. At moderate λ enough edits land to ask what happens to persons whose direct edit succeeded:

Among successful direct edits	λ = 0	λ = 1e4	λ = 3e4
Paraphrases still giving the old answer	7.6 ± 3.8%	23.8 ± 14.1% (paired +16.2, t = 3.0)	39.0 ± 16.2% (+31.4, t = 5.6)

At λ = 1e4, across all updated persons, the paraphrase log-prob gap moved from −2.53 to +0.78 nats (t = 8.8), and the old answer stayed in the paraphrase top 2 for 66.0% of queries, against 24.5% without EWC. Stiffness makes successful edits shallower, the coexisting-trace pattern the review predicted (H1c supported at λ ≤ 3e4). In this toy it sits on top of a much larger blocking effect.

EWC is far more intransigent than the anchor it approximates (H3)

The match held on test: EWC λ = 1e5 gave 92.0% retention, KL-all λ = 0.3 gave 94.8% (paired −2.8 ± 5.1, not significant).

At matched retention	EWC − KL-all	EWC − KL-ret	KL-all − KL-ret
Edit success	−55.5 ± 8.0 (t = −19.2)	−61.0 ± 8.4	−5.5 ± 4.0
Paraphrase accuracy	−27.5 ± 12.2	−43.0 ± 8.0	−15.5 ± 10.4
Stale paraphrase	+21.5 ± 15.5	+40.0 ± 14.5	+18.5 ± 9.7
Stale 2-hop, trained half	+36.6 ± 4.9	+51.2 ± 2.4	+14.6 ± 3.2
Stale 2-hop, held half	+2.9 ± 9.5 (n.s.)	+12.7 ± 11.6	+9.8 ± 7.5

The same held at about 85% retention (EWC $\lambda = 3e4$ vs KL-all $\lambda = 0.1$: edit −47.5 ± 8.2). Including the superseded queries in a function-space anchor does entrench them (KL-all vs KL-ret, +18.5 pp stale paraphrases), but mildly, and both KL anchors still learned (69.0–74.5% edit success). EWC at the same retention behaved more like the earlier toy result for KL on the *updated* queries, which held learning to 5%. The diagonal weight-space penalty does not separate the weights that store the old fact from the weights any change to that fact must pass through.

Held-half consequences were the exception: 68.6–86.4% stayed stale under the KL anchors and 89.3% under EWC $\lambda = 1e5$, which was not significantly worse than KL-all there. In this toy a held 2-hop answer is a separately memorised fact, and any method that preserves old facts preserves it.

The Fisher ledger is a clean null (H4)

At λ_B the ledger released 4.7% of the total Fisher mass over the stream.

At $\lambda = 1e5$	Edit	Stale paraphrase	Stale 2-hop, held	Stale 2-hop, trained	Log-prob gap, paraphrase	Unrelated retention
Full EWC	13.5 ± 5.6	81.7 ± 15.1	89.3 ± 6.9	77.9 ± 8.6	6.11	92.0 ± 5.4
Ledger: superseded + dependent	13.5 ± 5.6	80.8 ± 15.5	88.7 ± 8.0	76.9 ± 9.3	5.98	91.9 ± 5.4
Superseded only	13.5 ± 5.6	81.2 ± 15.7	89.3 ± 6.9	77.6 ± 8.6	6.01	92.0 ± 5.4
Random facts, ledger's mass	13.5 ± 5.6	81.2 ± 15.7	88.7 ± 8.0	76.9 ± 9.3	6.03	91.9 ± 5.4
Global decay, ledger's mass	13.5 ± 5.6	81.2 ± 15.7	88.7 ± 8.0	76.9 ± 9.3	6.03	91.9 ± 5.4
No EWC	62.5 ± 13.5	10.8 ± 1.4	32.0 ± 12.4	11.9 ± 14.4	−2.53	44.3 ± 8.2

Ledger minus full EWC was −1.0 ± 1.3 pp on stale paraphrases and −0.7 ± 1.1 pp on stale held 2-hop answers. Ledger minus random release was at most 0.5 pp on any stale metric. The kill rule required 5

pp and fired at both λ_B and λ = 1e4. The only systematic effect was a 0.1-nat shift in the old answer's log-prob.

Why. Weight each coordinate's stiffness by how hard a correction's gradient pushes on it. The superseded fact's own entry supplied 0.41% of the stiffness the correction faces; with its 2-hop dependent, 0.51%; with every fact about that person, 0.76% (0.37–0.81% across seeds). The correction pushes mostly on shared weights (output layer, attention and MLP matrices), whose push-weighted Fisher is 6.8× the average, and that stiffness comes from the other facts in the old KB. The stale fact is protected by collective stiffness, not by its own, so the "intransigence by inheritance" premise is wrong.

No targeted release helped either (exploratory). Zeroing stiffness on the coordinates the update pushes hardest against was no better than random coordinates at the ledger's mass (edit -1.5 ± 2.8, stale paraphrase -1.2 ± 3.3). At 5× that mass (24% of the Fisher), gradient-targeted release cut stale rates by 17–19 pp at a cost of 13.7 ± 6.1 pp of unrelated retention, random coordinates did similarly, and targeting was *worse* than random on edit success (-5.5 ± 1.4). Release only slides the model along the same λ trade-off curve.

Verdicts

Hypothesis	Verdict	Key number
H1a: edit–consequence gap widens with λ	Rejected (K1)	paraphrase gap $+1.2 \pm 2.5$ pp per decade; held 2-hop gap -14.7 ± 6.6
H1b: stale paraphrase and 2-hop rise with λ	Supported	$+28.4 \pm 2.3$ and $+21.8 \pm 5.8$ pp per decade
H1c: new fact written beside the old one	Supported at $\lambda \leq$ 3e4; blockage dominates at λ_B (K3)	stale paraphrase given a successful edit 7.6 → 23.8 → 39.0%
H2: unrelated retention rises	Supported	44.3 → 92.0 → 97.9%
H3: EWC behaves like the old-query KL anchor	No: EWC is far more intransigent	edit -55.5 ± 8.0 at matched retention
H4: ledger beats full EWC and random release	Rejected (K2)	-1.0 ± 1.3 pp stale paraphrase; superseded fact's share of stiffness 0.41–0.51%

What this does and does not show

- EWC is not "a cheap KL anchor" for revision. At equal retention, the KL anchor on old queries beat exact-Fisher EWC by 47–56 pp of edit success.
- If a KL anchor is used, exclude superseded facts' own queries. Including them cost +18.5 pp stale paraphrases and +14.6 pp stale trained consequences at equal retention.
- Per-fact release of weight stiffness fails for a structural reason. It joins directional suppression in the earlier weight-editing study as a localisation idea that a matched random control fully explains.
- The bypass signature is real but secondary. Successful edits under stiffness are shallower, with the old answer still decodable on paraphrases.
- Limits. Toy scale with single-token answers and an enumerable input space; whether the 0.5% share shrinks or grows at LLM scale is open. The 2-hop consequences are memorised facts, not

compositions (held-out composition was about 13% even without EWC), so spontaneous propagation could not be tested meaningfully. Edit success at $\lambda = 0$ was only 62.5%, because later stages overwrite earlier updates. One EWC variant: a single old-KB diagonal Fisher anchored at the pretrained weights, no online re-consolidation, Adam only; a non-diagonal Fisher might separate fact-specific from shared curvature better. The ledger-share diagnostic used 20 of 40 updated persons per seed. The dev pilot showed the headline pattern before the test run, and the pre-registration says so. Five seeds, uncorrected tests.

Deviations

None to the planned test arms: all 140 test runs completed with no reruns. The ledger was computed lazily and memoised rather than pre-stored (identical values). The ledger-share diagnostic is descriptive and used the update gradient at the pretrained weights. Kill rules K1, K2 and K3 all fired and are reported above. The exploratory coordinate-release arms were declared before the test run.

Prior work

- Kirkpatrick et al. 2017, arXiv:1612.00796: elastic weight consolidation.
- van de Ven 2025, arXiv:2502.11756: the Fisher estimator decides most published EWC results; the exact Fisher was best.
- Schwarz et al. 2018, arXiv:1805.06370: online EWC.
- Clemente et al. 2025, arXiv:2502.04390: contradicting updates destroy unrelated knowledge under importance-based routing.
- Foster et al. 2024, arXiv:2308.07707: selective synaptic dampening with forget- and retain-set Fishers.

The literature review found no published study measuring EWC on contradicting fact updates with stale-answer metrics, and no study subtracting per-example Fisher contributions to release superseded knowledge.

Data and code

The pre-registration, code, aggregate results and generated tables are in the study folder `research_notes/Elastic weight consolidation deep dive/experiments/E4_ewc_stiffness/` in the NeuroGenesys repository. Per-run JSONs are kept out of the repository. The findings are entered in the findings ledger.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/ewc-stiffness-stale-facts