A gradient kernel flags contradictions but does not predict forgetting, and steering rehearsal by it backfires

W. Schulz

Nov 10, 2026

Abstract

A pre-registered toy test of whether the gradient kernel between a new fact and stored facts predicts which beliefs an update will break or fail to change. Low confidence beat every kernel at predicting forgetting in the update stream (AUROC 0.737 vs 0.663). The kernel predicted consequence propagation and its sign detected contradictions, but kernel-chosen rehearsal cost retention.

W. Schulz, NeuroGenesys. *Pre-registered 2026-11-10, run once on 5 seeds. Toy scale: a 2-layer transformer (about 0.35M parameters) on a synthetic knowledge base, with a single-edit-set check on pythia-160m. This note reports the frozen analysis, including the hypotheses that failed.*

Summary

Can the gradient kernel between a new fact and each stored fact, computed before the update, say which beliefs the update will break or fail to change? And does acting on that prediction help? The literature review behind this study proposed pairing a coupling kernel ("what *will* change") with a knowledge-base relation map ("what *should* change"). On the toy the answer splits three ways.

- The kernel is worse than confidence at predicting collateral damage where that baseline was set. In the sequential update stream, low confidence again scored AUROC 0.737 [0.712, 0.762]. The best kernel variant scored 0.663 ($t = -8.1$), and none closed the gap.
- For small isolated edits to a converged model the ranking flips, but a structural rule beats every kernel. Confidence is near chance there (0.540) and the signed kernel scores 0.684, while "the probe has the same relation as the edit" alone scores 0.831.
- The kernel carries two things confidence does not. Its coupling to an edit's should-change 2-hop fact predicts how far that answer moves (Spearman +0.41 [+0.26, +0.56]; confidence −0.03). Its sign separates stored facts that contradict an edit from re-phrasings that agree with it perfectly (AUROC 1.000), which no activation-similarity kernel can do.
- Steering rehearsal by the kernel backfires. Unsigned-kernel rehearsal lost 5.2 pp of retention to random ($t = -3.3$), replicating earlier toy results. Inside a relation-map pipeline, kernel-chosen rehearsal cost 16.1 pp of retention. Swapping it for low-confidence rehearsal kept the stale-answer fix and gave the best operating point measured: at the full budget, 97.5% retention, 78.0% update accuracy and 18.5% stale covered 2-hop answers, against 95.7 / 64.0 / 94.6 for random replay.

Design

Model and knowledge base. The 2-layer pre-LN transformer ($d = 128$, about 0.35M parameters, 98 single-token answers) is pretrained on 400 persons (employer, lives, hobby, and the 2-hop workcity =

hq(employer)) and 40 companies (hq). The testbed was imported unmodified with cached weights for seeds 0–4; seed 5 was the pilot and is excluded.

Isolated-edit regime. Each of 90 edits per seed (40 contradicting employer updates, 20 company hq moves with about 10 dependent employees each, 30 new-person facts) is applied alone to a fresh copy of the pretrained model: Adam, lr 3e-5, 40 steps. Edit success was 100%, and 0.99% of unrelated (edit, probe) pairs flipped. A heavier regime (lr 1e-4, 20 steps) flipped 7.3%.

Stream regime. The earlier toy studies' stream: 228 new facts including 40 contradicting updates, 8 stages, Adam 1e-4, 60 steps × 64, half new facts and half rehearsal. The rehearsal corpus is unpruned. A relation map always knows each update's superseded fact, but knows the stale 2-hop dependents and the new 2-hop consequences only for a fixed covered half of updated persons. The reproduction arms match the earlier study to the last digit in 4 of 5 cases; the fifth differs by 0.53 pp on one seed and matches exactly at the earlier study's thread count.

Kernels, all computed before the edit: signed and unsigned cosine of full per-example gradients (about 350k parameters); an Adam first-step variant, g·sign(g_edit); count-sketch projections to 512 and 4,096 dimensions; and a CLaRE-like cosine of block-0 and final-layer activations. Baselines: low confidence (−margin), entity/value overlap and random.

Statistics. Prediction metrics are means over 5 seeds with 95% t-intervals. Intervention metrics are mean ± sd with paired t over seeds. There is no multiple-comparison correction, so $|t| < 3$ should be read as suggestive. About 1.6 h of CPU in total, including the pythia check.

Results

Collateral damage in the stream: confidence wins

Before every stage of the no-rehearsal stream, each retained fact was scored and labelled by whether it went from right to wrong across the stage (base rate 25.0%). Both earlier numbers replicate exactly.

Predictor	AUROC [95% CI]	t vs low confidence
Low confidence (−margin)	0.737 [0.712, 0.762]	–
Stage-mean gradient dot	0.688 [0.630, 0.746]	−3.3
CLaRE-like, final layer, max over stage facts	0.684	−13.4
Confidence + signed cosine (rank sum)	0.684 [0.658, 0.710]	−4.4
Adam-sign kernel	0.669 [0.636, 0.702]	−5.9
Max over stage facts of −cos	0.663 [0.628, 0.698]	−8.1
Stage-mean −cos (earlier replication)	0.654 [0.622, 0.686]	−7.6
CLaRE-like, block 0, max over stage facts	0.632	−5.0
Max over stage facts of |cos|	0.592 [0.556, 0.628]	−18.4
Overlap / neighbour / random	0.403 / 0.495 / 0.495	–

With random rehearsal running (base rate 8.6%), confidence plus the signed kernel edged past confidence alone, 0.783 vs 0.766 ($t = 2.2$). That is suggestive only.

Collateral damage from isolated edits: the kernel wins, structure wins more

In a converged model every fact sits far from its decision boundary, so confidence carries almost nothing. The facts that flip are the ones the edit pushes on.

Predictor (base rate 0.99%)	Pooled AUROC [95% CI]	AUPRC	Within-edit AUROC
CLaRE-like, block 0	0.880 [0.846, 0.913]	0.080	0.887
|cos|	0.747 [0.719, 0.774]	0.063	0.767
Signed −cos	0.684 [0.658, 0.710]	0.087	0.731
−cos, count-sketch 4,096	0.678 [0.657, 0.700]	0.083	0.725
−cos, count-sketch 512	0.657 [0.635, 0.679]	0.062	0.698
Entity/value overlap	0.567 [0.552, 0.582]	0.044	0.612
Low confidence	0.540 [0.511, 0.570]	0.011	0.515
Same relation as the edit (post hoc)	*0.831 [0.802, 0.860]*	–	–

84.5% of flips were probes with the same relation as the edit, and relations that share output classes (hq, workcity and lives all answer with a city) flipped together. Within relation strata, signed cosine (0.711), Adam-sign (0.709) and CLaRE (0.703) tied, and |cos| fell to 0.651. CLaRE's pooled 0.880 and the unsigned kernel's lead over the signed one therefore come from relation structure. A 4,096-dimension count-sketch kept the full kernel's ranking (Spearman 0.939, AUROC within 0.006–0.013); 512 dimensions lost 0.03–0.07.

Consequence failure: the one prediction only the kernel makes

When an employer changes or a company moves, the dependent workcity facts should change. Under the minimal edit 8.6% [4.3, 12.8] of them did, against a 0.4% bias control (unrelated workcity queries moving to the edit's new city). The coupling between the edit's gradient and the consequence's gradient with its new label predicted how far the consequence moved.

Score (201 should-change 2-hop facts per seed)	Failure AUROC [95% CI]	Spearman with Δlog p(new)
−cos(g_edit, g_consequence,new)	0.646 [0.539, 0.752]	+0.409 [+0.260, +0.557]
−Adam-sign	0.699 [0.596, 0.802]	+0.448 [+0.307, +0.588]
−CLaRE	0.666 [0.601, 0.730]	+0.312 [+0.189, +0.435]
Old 2-hop answer's confidence	0.522 [0.403, 0.641]	−0.034
Random	0.541 [0.456, 0.627]	+0.002

This reproduces, on a toy with a bias control, the GradSim result of Qin et al. (2024). Coupling was low, though: in the stream 75% of covered updates fell below the 0.3 "weakly coupled" threshold (mean cosine 0.22), so the practical rule is to expect almost every consequence to need explicit training.

The kernel's sign is a contradiction detector

Fact (gradient with its stored label, vs the edit's gradient)	Mean cos	% negative	AUROC(−cos) vs neutral
Superseded fact, other trained templates (T1 / T2)	−0.306 / −0.286	97%	0.969 / 0.960
Superseded fact, never-trained template	−0.343	97%	0.965
Same paraphrased query with the edit's new label (consistent)	+0.731	0%	–
Stale 2-hop dependent	−0.075	87%	0.818
Same entity, other relation (consistent)	+0.013	46%	0.453
Neutral unrelated fact	−0.000	50%	–

The consistent paraphrase is the decisive control. Its input is identical to the superseded paraphrase's, so an activation kernel scores the two the same (AUROC 0.5 by construction). The signed gradient kernel separates them perfectly, 1.000, because the stored answer enters the gradient. It also separates superseded facts from consistent facts about the same entity (0.967). Stale 2-hop dependents stood out at 0.818, which fails the pre-registered prediction that they would not; part of that is output-class overlap for hq moves (0.858, against 0.681 for employer updates).

Without a relation map, a signed-kernel prune (drop the two most negative-cosine facts per update) caught the superseded fact for 60–72% of updates and the stale dependent for 1–1.5%. It cut stale-employer answers only at the full budget (−8.5 pp, $t = -3.7$).

Acting on the kernel: rehearsal selection loses, the relation map does the work

All arms rehearse the same number of facts per stage. Small budget (about 27 facts per stage), mean ± sd over 5 seeds:

Arm	Old retention	Update acc.	Stale employer	Stale 2-hop, covered	Stale 2-hop, uncovered	Compute
No rehearsal	47.6 ± 4.9	62.0 ± 5.7	4.0	41.9	48.1	1.00×
Random	87.5 ± 0.9	66.5 ± 6.0	13.0	88.2	83.0	1.00×
Low confidence	92.2 ± 1.6	71.5 ± 4.2	17.0	83.0	90.7	1.28×
Kernel |cos| (replication)	82.3 ± 3.6	73.0 ± 5.7	9.5	71.8	73.4	1.70×
Kernel signed −cos	87.2 ± 2.8	76.5 ± 4.2	8.5	71.4	72.5	1.70×
Kernel + relation map	71.5 ± 3.2	75.0 ± 7.3	5.0	11.3	37.3	1.70×

Arm	Old retention	Update acc.	Stale employer	Stale 2-hop, covered	Stale 2-hop, uncovered	Compute
Relation map + random rehearsal	79.7 ± 2.8	70.5 ± 7.4	6.5	15.4	47.4	1.01×
Relation map + low-confidence rehearsal	84.5 ± 2.2	75.5 ± 6.5	8.0	13.1	66.4	1.29×

- Kernel rehearsal. Unsigned-kernel rehearsal lost 5.2 ± 3.5 pp of retention to random ($t = -3.3$) and 9.9 to confidence ($t = -6.2$). The signed variant tied random on retention (−0.3) and raised update accuracy by +10.0 ($t = 3.7$), a gain we did not predict and have only a post-hoc guess for.
- The full pipeline. Kernel + relation map cut stale covered 2-hop answers 76.8 pp below random ($t = -15.3$) and lost 16.1 pp of retention ($t = -10.0$).
- Its parts. Consequence examples did all of the stale-2-hop work (removing them returned it to 70.8%). Pruning contributed little, because few superseded facts entered the pool. The kernel rehearsal was the harmful part: replacing it with random (+8.3 pp retention, $t = 3.2$) or low-confidence rehearsal (+13.0, $t = 8.1$) recovered retention at the same stale-2-hop rate.
- Uncovered persons did not update. The pipeline's lower uncovered stale rate (37.3%) was mostly forgetting: new-value correctness there was 16.5%, against 10.6% with no rehearsal.

At the full budget (about 108 facts per stage):

Arm	Old retention	Update acc.	Stale employer	Stale 2-hop, covered	Stale 2-hop, uncovered	Compute
Random	95.7 ± 0.8	64.0 ± 11.5	24.5	94.6	89.7	1.00×
Low confidence	98.6 ± 0.5	61.5 ± 7.6	32.0	95.0	92.7	1.28×
Relation map + low-confidence rehearsal	97.5 ± 0.9	78.0 ± 4.8	13.0	18.5	88.7	1.29×
Kernel + relation map	88.9 ± 2.4	80.5 ± 3.7	7.5	13.7	53.5	1.70×

Relation map plus confidence beat random by +1.8 retention ($t = 4.4$), +14.0 update accuracy ($t = 3.5$) and −76.0 pp stale covered 2-hop. Confidence alone, on an unpruned corpus, pulled partly overwritten superseded facts back into the pool: its stale-employer rate rose to 32.0% against 24.5% for random, and its update accuracy was no better. Confidence needs the pruning step.

pythia-160m points the same way

On 55 typed facts (a belief is the argmax among 12–20 same-type candidates; 53 of 55 correct) and 55 counterfactual edits to the layer 3–5 MLPs (36% flip rate), signed cosine tied low confidence on flips

(0.685 [0.654, 0.714] vs 0.681) and beat it on large log-prob drops (0.781 vs 0.575). CLaRE was best (0.775 / 0.805), but "same answer type" alone scored 0.701, and within an answer type confidence beat the kernel on flips (0.684 vs 0.632). One model, one edit set, no seeds: directional only. An earlier version that used the greedy next token as the belief scored a spurious 0.86, because 27 of its 43 "beliefs" were " the"; it is kept, labelled confounded, and not interpreted.

Hypotheses

	Hypothesis	Outcome
H1a	Stream: best kernel beats low confidence	Failed (0.663 vs 0.737, $t = -8.1$)
H1b	Isolated: signed kernel beats low confidence, overlap, random	Held, but a weak test: confidence is near chance here and "same relation" (0.831) beats every kernel
H1c	Isolated within-edit signed kernel > 0.65	Held (0.731)
H1d	|cos| no better than signed	Failed pooled (+0.063, $t = 15.7$); post hoc, the lead is relation structure
H1e	Sketch to 4,096 preserves the kernel; 512 does not	Held
H1f	CLaRE within 0.03 of the signed kernel	Failed pooled (0.880 vs 0.684); tied within relation strata
H1g	Stream: confidence + kernel beats confidence	Failed (0.684 vs 0.737); suggestive with random rehearsal (0.783 vs 0.766, $t = 2.2$)
H1h	Consequence coupling Spearman > 0.2	Held (+0.409)
H1i	Consequence success $< 10\%$ and about equal to the bias control	Half held (8.6%, but 20× the 0.4% control)
H2a	Superseded facts negative, AUROC ≥ 0.8 vs neutral	Held (0.969 / 0.960)
H2b	Superseded vs same-entity consistent ≥ 0.8	Held (0.967; 1.000 vs the consistent paraphrase, post hoc)
H2c	Stale dependents not separated from neutral	Failed (0.818)
H3a	Unsigned-kernel rehearsal no better than random	Held: worse, −5.2 pp
H3b	Unsigned-kernel rehearsal lowers update accuracy	Failed (+6.5, n.s.)
H3c	Kernel + relation map beats others on stale covered 2-hop and update accuracy	Held for stale 2-hop; update accuracy only at the full budget

	Hypothesis	Outcome
H3d	Kernel + relation map retention within 1 pp of random	Failed (−16.1 small, −6.8 full)
H3e	Uncovered stale 2-hop not > 5 pp better than random	Failed as stated (−45.7), from forgetting, not propagation
H3f	The kernel component adds nothing inside the pipeline	Held, and stronger: it costs 8–13 pp retention
H3g	Signed-kernel prune lowers stale employer answers	Mixed: n.s. small; −8.5 (t = −3.7) full

What this does and does not show

- The kernel tells you what an edit is coupled to, and its sign tells you whether a stored fact disagrees. Both are real information, and neither is what a rehearsal selector needs. In the drifting stream, how close a fact sits to its decision boundary predicted its loss better than gradient alignment with the update.
- The 0.737 confidence baseline is regime-specific. For small edits to a converged model confidence was uninformative. That is a plausible reason why low-confidence targeting was worse than random at Qwen 1.5B in earlier work, and a reason to measure which predictor works in a given regime before spending rehearsal budget on it. That test was not run.
- What fixed stale consequences was not a kernel. It was knowing from the knowledge base which facts should change and training them explicitly, and it fixed covered persons only.
- Limits. Toy and synthetic: single-token answers over 98 classes, relations that share output classes, paraphrases that are token reorderings, and contradictions that are mostly the easiest case (same query, different answer). Neither regime tested is the one a real batch-editing pipeline sits in. Each pipeline design constant was tested at one value. The pilot seed set the edit learning rate and the signed-prune rule. The relation-structure analysis and the consistent-paraphrase control were post hoc. The pythia check is one model and one edit set. Five seeds, many uncorrected contrasts.

Deviations

Logged in the pre-registration: the signed-prune threshold rule flagged 40–70% of the corpus in the pilot and was replaced by top-2 per update before the main runs; one secondary arm was added from the pilot; the reproduction mismatch was traced to float reduction order (thread count); the isolated main regime was re-run to store probe indices, with identical numbers; the post-hoc analyses above were added after the main run; and the pythia check was redone with typed beliefs after the greedy-token version was found confounded.

Prior work

- Qin et al. 2024, arXiv:2407.12828: GradSim, gradient similarity as a predictor of ripple-effect success.
- Baser et al. 2026, ACL Findings: CLaRE, activation-based entanglement scores for ripple effects.
- Jin & Ren 2024, arXiv:2402.01865: forecasting which examples will be forgotten.
- Cohen et al. 2024, arXiv:2307.12976: RippleEdits.
- Park et al. 2023, arXiv:2303.14186: TRAK, random projections of per-example gradients.

Data and code

The pre-registration, code, per-seed summaries and generated tables are in the study folder `research_notes/Logical coherence of LLM world models/experiments/E2_kernel_ripple/` in the NeuroGenesys repository. Per-pair working data is kept out of the repository because of its size. The findings are entered in the findings ledger.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/kernel-edit-ripple