Iterated self-closure settles fast, but only evidence that overrides self-belief makes the settled state true

W. Schulz

Nov 10, 2026

Abstract

A pre-registered toy test of "chewing": a model answers its own atomic questions, applies the knowledge base's composition rule and retrains on the derived consequences. The loop always converged, but unanchored it left 44% of updated persons coherently wrong. Anchoring to the told facts fixed that, and overwriting self-beliefs first carried updates to entities no gold consequence covered.

W. Schulz, NeuroGenesys. *Pre-registered 2026-11-10, run once on 5 seeds with a replication on 5 fresh seeds. Toy scale: a 2-layer transformer (about 0.35M parameters) on a synthetic knowledge base. This note reports the frozen analysis, including the two hypotheses that were not supported.*

Summary

Can a model "chew" old and new knowledge into one consistent whole? The loop tested here has the model answer its own atomic questions, applies the knowledge base's composition rule to those answers, retrains on the derived consequences, and repeats. The literature review behind it predicted that such loops converge, but to a self-consistent attractor rather than to truth, unless evidence is clamped.

- The loop always converges, and quickly (H1 supported). Unanchored self-closure changed 5.3% of the 1,888 beliefs in round 1 and 0.06% in round 10, and every seed fell below 1% change per round by round 2–3.
- Its fixed point is coherent but not true. Global consistency rose from 76.3% to 94.8%, and updated persons became 100% consistent, but 44.0% of them ended *coherently wrong*: the old employer and the old employer's city. Among new persons, whose facts were half-forgotten during the update, the coherently-wrong rate rose from 10.0% to 55.0%.
- The pre-registered lock-in test was not significant (H1b not supported). On updated persons, pooled over templates, coherently-wrong rose only +4.0 pp [−0.7, +8.7], because 40% of them were already coherently stale before the loop started.
- Labelling paraphrases with the model's majority belief reverted edits. Updates learned in one phrasing fell from 73.5% to 55.5% correct in that phrasing within one round.
- Anchoring to the told facts fixed all of it (H2 supported). Updated persons ended 0.0% coherently wrong, against 44.0% unanchored (−44.0 pp [−53.7, −34.3]).
- Pruning did not change the fixed point at the pre-registered anchor weight (H3 not supported). Anchored and pruned closure both reached 100%, so the contrast was exactly 0.0. Pruning made the loop settle about 10× faster, and in a post-hoc sweep it decided the outcome once evidence was a small share of training.

- Self-closure reached entities that gold consequences never covered (H4 supported). Uncovered updated persons reached 100% correct 2-hop answers, against 0% for gold consequences on the covered half alone. The reach comes from the loop enumerating every entity's query, not from learned composition.

Design

Item	Choice
Model	2-layer pre-LN transformer, d = 128, about 0.35M parameters, 98 single-token answers
Knowledge base	400 persons (employer, lives, hobby), 40 companies (hq), composition rule workcity = hq(employer); 3 query templates
Update stage	The earlier toy study's pruned-replay condition, reproduced exactly (max $\|\Delta\| = 0.0$ on every metric, seeds 0–4): 228 new facts in 8 stages, including 40 contradicting employer updates, all taught in template T0
Beliefs	The argmax within the value type of the log-softmax summed over all three templates ("pooled"); a T0-only view is reported for comparison
Loop	Per round: snapshot all 1,888 beliefs, build the arm's training set, train 100 steps at batch 64 (Adam, lr 3e-4, persistent across rounds). K = 10 rounds; K = 30 in a long run
Seeds	0–4 main; 5–9 replication; seed 10 pilot only
Compute	Under 30 minutes of CPU in total, three single-thread processes

The step count and learning rate were chosen by a pre-registered fit-only rule on the pilot seed. Every arm gets identical steps and examples per round.

Arm	2-hop labels each round	Atomic labels	Anchors (the 228 told facts as hard labels)
No chewing	–	–	–
Self-rehearsal (added control)	the model's own 2-hop answers	own beliefs	no
Unanchored closure	hq(employer) from own beliefs	own beliefs	no
Anchored closure	derived from own beliefs	own beliefs, even where they contradict an anchor	16 of 64 per batch

Arm	2-hop labels each round	Atomic labels	Anchors (the 228 told facts as hard labels)
Anchored + pruning	derived after overwriting anchored beliefs	own beliefs, anchored ones overwritten	16 of 64 per batch
Random-label control	a fixed random city per person	own beliefs	no
Gold covered-only reference	gold for the covered half, own answer elsewhere	as pruning	16 of 64 per batch

Updated and new persons are split into a fixed covered half and an uncovered half, as in the earlier study. Forty base persons whose 2-hop fact was never trained are held out of every chewing set, as a probe of learned composition.

Metrics. Consistency means the 2-hop belief equals hq of the employer belief. Coherently wrong means consistent but not true. Values are percentages, mean ± 95% t-interval half-width over 5 seeds; Δ is a paired within-seed difference in pp with its 95% CI. There is no multiplicity correction. Cells at a 0% or 100% ceiling have zero width because all 5 seeds agree.

Hypotheses.

- H1: unanchored closure converges.
- H1b: it locks in errors, so coherently-wrong updated persons rise from round 0 to round 10.
- H2: anchoring prevents that rise.
- H3: pruning decides whether the fixed point is true.
- H4: self-closure reaches uncovered entities where covered-only gold consequences do not.

Results

Fixed points at round 10 (main seeds 0–4)

Arm	Update acc.	2-hop, uncovered updated	Stale 2-hop, uncovered	Consistency, updated	Coherently wrong, updated	Coherently wrong, new	Consistency, all	Old facts kept
No chewing (r0)	55.5 ± 9.7	0.0	96.3 ± 7.7	40.0 ± 13.0	40.0 ± 13.0	10.0 ± 3.6	76.3 ± 2.3	98.4 ± 0.4
Self-rehearsal	53.5 ± 8.4	0.0	95.3 ± 7.3	42.0 ± 11.1	42.0 ± 11.1	12.0 ± 2.3	77.1 ± 2.6	98.4 ± 0.3
Unanchored closure	55.5 ± 9.7	55.9 ± 19.4	39.4 ± 19.7	100.0	44.0 ± 9.7	55.0 ± 14.2	94.8 ± 0.8	98.4 ± 0.4
Anchored closure	100.0	100.0	0.0	100.0	0.0	1.0 ± 1.9	94.7 ± 1.1	98.3 ± 0.4

Arm	Update acc.	2-hop, uncovered updated	Stale 2-hop, uncovered	Consistency, updated	Coherently wrong, updated	Coherently wrong, new	Consistency, all	Old facts kept
Anchored + pruning	100.0	100.0	0.0	100.0	0.0	0.0	94.3 ± 1.3	98.3 ± 0.5
Random labels	53.0 ± 8.0	1.1 ± 2.9	0.0	0.5 ± 1.4	0.0	3.3 ± 2.1	3.6 ± 2.1	98.3 ± 0.5
Gold covered-only	100.0	0.0	96.3 ± 7.7	49.5 ± 9.2	0.0	0.0	82.1 ± 1.7	98.3 ± 0.5

Pooled update accuracy at round 0 is 55.5%, against 73.5% in T0 alone: the two templates that never saw the update outvote it for about 18 pp of updated persons.

Pre-registered tests

Hypothesis	Contrast	Main Δ [95% CI]	Replication Δ [95% CI]	Verdict
H1	change rate r10 − r1, unanchored	−5.3 [−5.8, −4.7]; r10 = 0.06%	−4.8 [−5.2, −4.4]; r10 = 0.08%	supported
H1b	coherently-wrong updated, r10 − r0, unanchored	+4.0 [−0.7, +8.7]	+7.0 [+1.0, +13.0]	not supported on the primary seeds
H2	coherently-wrong updated, anchored − unanchored	−44.0 [−53.7, −34.3]	−51.0 [−57.4, −44.6]	supported
H3	2-hop accuracy updated, pruned − anchored	0.0 [0.0, 0.0]	+0.5 [−0.9, +1.9]	not supported (both at ceiling)
H4	2-hop accuracy uncovered updated, pruned − gold	+100.0 [+100.0, +100.0]	+97.2 [+94.0, +100.4]	supported

Convergence is not the problem

Unanchored closure's change rate fell from 5.3% in round 1 to 0.06% in round 10. Anchored closure took 4–6 rounds to fall below 1% per round, pruned closure exactly 3 in every seed, and even random labels settled after about 6. Between rounds 10 and 30 no endpoint moved by more than 2.1 pp (6 pp on the 40-person probe). On this knowledge base chewing is a contraction, with no oscillation of the kind loopy belief propagation can show. What matters is where it converges.

The unanchored fixed point is coherent, and wrong where its inputs were wrong

Closure installed exactly as many correct consequences as the model had correct atomic beliefs: uncovered 2-hop accuracy rose from 0 to 55.9%, matching 55.5% employer accuracy. The stale persons became consistently stale.

The pre-registered H1b metric rose only +4.0 pp, not significant. It was +7.0 [+1.0, +13.0] in replication and +5.5 [+2.4, +8.6] pooled over ten seeds, a post-hoc pooling. The rise was small because, in pooled beliefs, 40% of updated persons were already coherently stale at round 0: both of their unedited templates agreed on the old world. Where errors were incoherent to begin with, closure made them coherent:

- among new persons, coherently wrong rose +45.0 pp [+31.7, +58.3], from 10.0% to 55.0% (replication +43.7);
- in the T0 view, coherently wrong among updated persons rose +31.5 [+25.9, +37.1] (replication +37.5);
- across all persons it went from 4.8% to 12.2%.

The mechanism is the one the literature predicts; the primary metric was the wrong place to look for it.

Self-rehearsal across paraphrases undoes edits

Both arms that rehearse self-labels in every template, self-rehearsal and unanchored closure, reverted T0-only updates within one round. T0 update accuracy fell from 73.5% to 53.5% and 55.5% respectively (−22.5 [−28.0, −17.0] and −20.0 [−25.2, −14.8] pp over ten seeds). When a fact was learned in one phrasing and the model still holds the old value in two others, rehearsing the "consensus" erases the edit. Earlier toy work found that replay entrenches stale answers in phrasings an edit never touched; here self-consensus removes the edit from the phrasing where it was learned.

Anchoring prevents lock-in, by a non-monotone route

With the 228 told facts as fixed hard labels (16 of every 64 examples), every updated person ended correct and consistent, and coherently-wrong fell −40.0 [−53.0, −27.0] pp below anchored closure's own round 0. New persons reached 99.0% atomic accuracy. In round 1, though, anchored closure briefly *raised* coherently-wrong among new persons from 10.0% to 24.7%, deriving consequences from forgotten facts before the anchors restored them. Update accuracy then rose round by round: 73.5, 84.0, 91.0, 96.0 and 98.0% over rounds 1–5, and 100.0% by round 10.

Pruning decides when evidence is weak (post hoc)

At the pre-registered weight, pruning changed the path but not the destination. Update accuracy hit 100% in round 1 instead of by round 10. Over rounds 1–5 pruning was ahead by +11.5 ± 2.9 pp on update accuracy and +21.3 ± 8.7 pp on uncovered 2-hop accuracy. A post-hoc anchor-weight sweep (K = 30) showed when this matters:

Anchor examples per 64	Unpruned: update acc. r10 → r30	Unpruned: coherently wrong, new, r10 → r30	Pruned: update acc. r1 → r3	Pruned − unpruned, r10
0	55.5 → 55.5	55.0 → 55.0	98.5 → 100.0	+44.5 ± 9.7
1	76.0 → 94.5	39.7 → 16.3	99.5 → 100.0	+24.0 ± 6.4

Anchor examples per 64	Unpruned: update acc. r10 → r30	Unpruned: coherently wrong, new, r10 → r30	Pruned: update acc. r1 → r3	Pruned − unpruned, r10
2	82.5 → 97.0	34.0 → 9.3	99.5 → 100.0	+17.5 ± 7.9
4	93.5 → 99.5	22.0 → 5.7	99.5 → 100.0	+6.5 ± 6.4
16 (pre-registered)	100.0 → 100.0	1.0 → 0.0	100.0 → 100.0	0.0

Without pruning, each edited fact's self-labels vote against its anchor every round, and the anchor wins only as fast as its share of the gradient allows: tens of rounds at 1/64, never at 0. With pruning the vote is removed, and the loop was true by round 3 at every weight, including zero explicit anchor examples. This supports the review's claim that *the selector, not the forgetting, decides whether consolidation converges to truth*, but only post hoc and outside the pre-registered setting.

Consistency does not distinguish the true fixed point from the false one

Unanchored and anchored closure ended at the same global consistency, 94.8% and 94.7%, while one was 44.0% coherently wrong on updated persons and the other 0.0%. Any consolidation that selects on internal consistency alone, without priority for evidence, selects whichever attractor it starts nearest to.

Reach to uncovered entities comes from enumeration

The gold covered-only arm reproduced the known ceiling: covered updated persons 100% correct, uncovered 0% (96.3% stale). Pruned self-closure reached 100% on uncovered updated persons and beat gold by +94.7 [+88.7, +100.6] pp on uncovered new persons, with no external consequence data. On the 40 probe persons, whose 2-hop question no arm ever trained, accuracy rose only +14.0 [+10.5, +17.5] pp under unanchored closure and +8.5 [−2.6, +19.6] under pruned closure, and +5.5 and +4.0 in replication, from a 25–38% base. Chewing taught the network a little composition. Almost all of the propagation came from the loop asking about every entity and writing down the derived answer.

Retention is nearly free

Old atomic facts stayed at 98.2–98.4% in every arm (all changes within ±0.4 pp), and old 2-hop facts moved −1.3 to −1.5 pp in closure arms, not significant. The random-label control destroyed old 2-hop facts (99.7% → 2.6%), which confirms that the *content* of the derived labels, not extra training on 2-hop questions, drives the effects above.

What this does and does not show

- Chewing is a settling process, and it converges. On this knowledge base it converged in a few rounds at almost no retention cost.
- Left alone, it settles on whatever the model mostly believes. It made 45 pp more of half-forgotten entities coherently wrong, and it erased a fifth of fresh edits in the process.
- Evidence has to act on the belief set, not just the loss. Overwriting what the evidence contradicts, then deriving, carried an update's consequences to every entity the loop asked about. Covered-only consequence training had not done that in any earlier NeuroGenesys run.

- It is not learned composition. The network barely generalised to entities it was never asked about, and H4 is close to true by construction: the loop writes a label for every enumerated person.
- Limits. Toy scale; one rule, one hop, with near-perfect hq beliefs, so every derivation error came from the employer link. The anchors were the update itself, all correct (12% of atomic beliefs); partly wrong anchors were not tested, and overwriting would propagate a wrong anchor just as efficiently. Pooling beliefs over templates is a modelling choice that exposed the edit reversion. Many endpoints sit at ceilings, which made H3 untestable at the pre-registered weight. The anchor sweep, the settling-speed contrast and the ten-seed pooling of H1b were added after seeing the main results. Five seeds per configuration, no multiplicity correction.

Deviations

Two minor deviations are logged in the pre-registration: a wording fix to the change-rate denominator made before the pilot, and use of an existing read-only Python environment. Three post-hoc analyses are labelled as such: the anchor-weight sweep, the settling-speed contrast and the ten-seed pooling of H1b. Result files were trimmed to the reported metric families. No hypothesis, threshold, metric definition, arm, seed set or primary endpoint was changed.

Prior work

- Akyürek et al. 2024, arXiv:2401.08574: Deductive Closure Training, generating and filtering implications before fine-tuning.
- Shumailov et al. 2024, Nature: recursive training on a model's own outputs erases the tails.
- Shafayat et al. 2025, arXiv:2505.21444: majority-vote self-reward collapses into self-consistent, wrong outputs.
- Thagard 1989, Behavioral and Brain Sciences: ECHO's "data priority" for evidence units.
- Cohen et al. 2024, arXiv:2307.12976: RippleEdits, edits that do not propagate to their consequences.

To our knowledge no earlier study tracks both consistency and accuracy across iterations of self-closure on a declarative knowledge base, which is the measurement made here.

Data and code

The pre-registration, code, per-arm summaries for the main, long, replication and anchor-sweep runs, and full generated tables are in the study folder `research_notes/Logical coherence of LLM world models/experiments/E3_iterative_closure/` in the NeuroGenesys repository. The findings are entered in the findings ledger.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/iterated-self-closure