Trust the premises, not the confidence: which belief to drop when an LLM contradicts itself, and why the solver is not the bottleneck

W. Schulz

Nov 10, 2026

Abstract

A pre-registered test of MaxSAT belief repair on llama3.2 3B and llama3.1 8B answers to MQuAKE 2-hop chains. Ordering beliefs by graph entrenchment beat evidence priority, confidence and chance; confidence barely beat random. Most of the gain came from making each belief's phrasings agree, LLM-judged constraints halved it, and the exact solver cost 0.1 ms against 23–38 s of LLM work per chain.

W. Schulz, NeuroGenesys. *Pre-registered 2026-11-10 after a 10-chain pilot, run once. Real model beliefs from llama3.2 3B (400 chains) and llama3.1 8B (250 chains), run CPU-only on a shared machine. This note reports the frozen analysis, including the two hypotheses that failed.*

Summary

Inference-time consistency layers (BeliefBank, ConCoRD, REFLEX and their 2026 successors) extract an LLM's beliefs, detect relations between them, and solve weighted MaxSAT for the most coherent assignment. The literature review behind this study found three gaps: nobody reports how often repair goes toward the *false* side (the coherently-wrong rate), nobody compares the rule that decides which belief gives way, and nobody has measured where the time goes. This study measures all three, with ground truth.

- Evidence priority did not win (H1 not supported). Entrenchment did. Ordering beliefs by their degree in the global constraint graph gave the best accuracy and the lowest coherently-wrong rate for both models: 85.8% and 49.8% (3B), 95.2% and 15.6% (8B). Evidence priority, with 20% of atomic facts verified, beat confidence and random on coherently-wrong rate but lost to entrenchment on both metrics.
- Confidence is barely better than random at choosing which belief to drop: +1.3 pp [0.3, 2.3] (3B) and +1.2 pp [−0.0, 2.5] (8B), with no difference in coherently-wrong rate.
- Coherently wrong is the normal outcome. After confidence-weighted repair every chain is consistent, yet 65.0% of 3B chains and 31.2% of 8B chains still hold at least one false belief.
- Most of the accuracy gain comes from making the three phrasings of each belief agree. Form agreement alone took 3B from 70.9% to 80.6%; adding the gold logic under confidence weights then changed accuracy by −1.2 pp.
- LLM-judged constraints roughly halve the gain. The judges' contradiction recall was 0.34 (3B) and 0.42 (8B).

- Ising-style heuristics were worse than the exact solver (H7 not supported). Simulated annealing and simulated bifurcation reached the optimum on only 69–82% of components and cost 8–16 points of accuracy.
- The solver is not the bottleneck (H6 supported). LLM work took about 23 s (3B) and 38 s (8B) per chain; the exact RC2 solve took 0.1 ms. The one solver-side cost is skipping decomposition: one merged 68,400-belief instance took 282.5 s, against about 1.2 s split into connected components.

Design

Item	Choice
Data	400 2-hop chains from MQuAKE-CF, original (unedited) Wikidata facts, 21 time-volatile or multi-valued relations excluded; drawn with seed 0 from 2,039 eligible chains, at most 40 per relation pair
Propositions per chain	7 for chain s → e1 → e2 with a type-matched distractor: A r1(s)=e1 (T), A′ (F), B r2(e1)=e2 (T), B″ (F), Bd (T), C composed(s)=e2 (T), C′ (F)
Gold constraints	Three Horn compositions (A∧B→C and two distractor chains) and three functional exclusions per chain
Forms	Each proposition asked three ways: question plus answer (Q), paraphrase (P), negation "is the proposed answer wrong?" (N); 21 belief nodes per chain
Store	Facts keyed by (subject, relation, object), so shared facts are single nodes: 2,280 propositions in 105 connected components for 3B (1,490 for 8B)
Models	llama3.2 3B on 400 chains, llama3.1 8B on 250; Ollama, Q4_K_M, temperature 0, one output token, p(yes) from log-probs; CPU only
Judged constraints	The same LLM judged all 21 pairs per chain for contradiction and 6 candidate implications (3 gold, 3 decoys); 120 chains for 3B, 80 for 8B
Repair	Weighted MaxSAT with RC2 (python-sat) per connected component
Statistics	Case-level bootstrap, 2,000 resamples, paired differences; metrics on beliefs never clamped as evidence; no multiple-comparison correction

Rules for which belief gives way. (a) Confidence, |logit|. (b) Entrenchment, the belief's degree in the global constraint graph. (c) Evidence priority: about 20% of atomic facts clamped to gold as hard clauses, the rest weighted as (a). (d) Random weights over 20 seeds. (e) No repair. Exploratory extras: an oracle, uniform weights, "atomic premises first" and within-chain degree.

Solvers. Exact RC2; simulated annealing (dwave-samplers) on the exact QUBO; a numpy simulated-bifurcation Ising solver (discrete SB tuned on synthetic instances, with default ballistic SB as a secondary arm).

Metrics. A chain is *consistent* when all six gold constraints hold and the three forms of each proposition agree. It is *coherently wrong* when it is consistent but holds at least one false proposition.

Results

Entrenchment beats confidence, evidence priority and chance

Gold-logic constraints, non-evidence beliefs; values in %, with 95% CIs.

Rule	3B accuracy	3B coherently wrong	8B accuracy	8B coherently wrong
No repair	70.9	(no case consistent)	78.9	(no case consistent)
Confidence	79.4 [77.7, 81.2]	65.0 [60.5, 69.5]	92.0 [90.2, 93.8]	31.2 [25.6, 37.2]
Entrenchment	85.8 [84.0, 87.4]	49.8 [45.0, 54.8]	95.2 [93.5, 96.7]	15.6 [11.2, 20.4]
Evidence priority (20%)	79.1 [77.2, 80.9]	60.8 [56.0, 65.5]	93.4 [91.6, 95.2]	21.2 [16.4, 26.0]
Random (20 seeds)	78.1 [76.4, 79.7]	65.5 [61.9, 69.2]	90.8 [89.0, 92.5]	30.4 [25.7, 35.6]
Oracle (upper bound)	98.4	6.5	100.0	0.0

Paired accuracy difference (pp)	3B	8B
Entrenchment − confidence	+6.3 [5.1, 7.6]	+3.1 [1.8, 4.6]
Entrenchment − random	+7.6 [6.6, 8.8]	+4.4 [3.3, 5.6]
Evidence − confidence	−0.3 [−1.1, 0.4]	+1.4 [0.7, 2.1]
Evidence − entrenchment	−6.6 [−8.0, −5.3]	−1.7 [−3.2, −0.4]
Confidence − random	+1.3 [0.3, 2.3]	+1.2 [−0.0, 2.5]
Confidence − no repair	+8.5 [7.3, 9.7]	+13.1 [11.7, 14.5]

Entrenchment cut the coherently-wrong rate by 15.3 pp [−19.0, −11.7] (3B) and 15.6 pp [−21.2, −10.4] (8B) relative to confidence. Under confidence repair, correct→wrong flips were 7.3 per 100 beliefs against 15.8 wrong→correct for 3B (32% of flips), and 3.4 against 16.5 for 8B. Entrenchment roughly halved the correct→wrong rate (3.7 and 1.9). Even with all forms credited for its clamped facts, evidence priority reached 82.7% (3B) and 94.6% (8B), below entrenchment.

Verifying more facts lowered the coherently-wrong rate steadily (8B: 29.2% at 0% verified to 15.2% at 40%), but the accuracy benefit barely spread to unverified beliefs for 3B (79.8% → 79.9%).

Why: the models get premises right and conclusions wrong (post hoc)

Belief	3B none	3B confidence	3B entrenchment	8B none	8B confidence	8B entrenchment
A (true hop 1)	77.5	70.2	75.7	94.0	90.8	94.8
B (true hop 2)	84.3	70.2	80.9	95.7	91.3	92.8
C (true 2-hop answer)	51.0	51.5	67.5	74.7	78.0	91.6

The models accept both true premises but reject the true conclusion: Q-form accuracy on true composed answers was 41.8% (3B) and 63.2% (8B), against 80.7% and 95.4% on true atomic facts. Their "No" on the conclusion is confident, so confidence-weighted repair drops a true premise to keep it. Entrenchment ranks the shared premises above the case-specific conclusion and flips the conclusion instead.

- Degree does not predict truth (AUROC 0.50 for 3B, 0.47 for 8B). Entrenchment works because conclusions are the least-connected beliefs and, here, also the ones the model gets wrong.
- A blunt "atomic premises first" rule recovers most of the gain: 84.2% (3B) and 93.6% (8B). Within-chain degree recovers less (80.9% and 92.8%), so premise sharing across chains adds the rest.
- Confidence is weakest on the negated form. Its AUROC for correctness was 0.76–0.82 on the direct forms but 0.54 on the negated form for 3B, and both models answered "not wrong" to about 90% of false atomic claims.

Most of the gain is paraphrase agreement (post hoc)

Making each proposition's three forms agree, with no cross-proposition constraints, already gave 80.6% (3B) and 89.7% (8B) under confidence weights. Adding the gold composition and exclusion rules changed accuracy by −1.2 pp [−2.1, −0.3] (3B) and +2.4 pp [1.2, 3.6] (8B) under confidence, and by +3.5 and +4.5 pp under entrenchment. For the 3B model, confidence-weighted logical repair on top of form agreement made beliefs slightly worse. Logic pays only when the rule for giving way is right.

LLM-judged constraints halve the gain

On the judged subsets, the gain of confidence repair over no repair was +4.8 pp (3B) and +4.2 pp (8B) with LLM-judged constraints, against +9.3 and +12.0 pp with gold logic on the same chains (H8 differences −4.5 [−6.3, −2.8] and −7.8 [−10.4, −5.4]). Coherently-wrong rates rose to 75.8% (3B) and 53.8% (8B).

Judge quality against gold relations	3B	8B
Contradiction (pairs): precision / recall	0.71 / 0.34	0.44 / 0.42
Implication (triples): precision / recall	0.80 / 0.61	0.79 / 0.82

Entrenchment still beat confidence under judged constraints, by +2.9 pp [0.9, 5.0] and +3.9 pp [0.9, 6.9].

Where the time goes

Stage (per chain)	3B	8B
LLM belief elicitation	8.5 s	13.6 s
LLM relation checking	14.2 s	24.9 s
Graph and component building	0.03 ms	0.03 ms
RC2 solve, gold-logic store	0.10 ms	0.08 ms

LLM stages were more than 99.998% of wall time in every configuration, and relation checking alone was 1.1×10^5 (3B) and 2.3×10^5 (8B) times the solve on the same judged instances. Exhaustive pairwise checking of the 2,280-proposition store would need 2,598,060 calls, about 14 days at the measured 0.46 s each; RC2 solves that store in under 0.5 s. The latencies are CPU-only on a contended machine. A GPU might cut LLM time 3–20× (not measured), which would leave the ratio above 10^4.

The one solver-side cost appears when an instance is not decomposed. Merged into one formula, RC2 took 0.47 s at 2,280 propositions, 23.3 s at 22,800 and 282.5 s at 68,400; solved per connected component it stayed linear, at about 1.2 s for 68,400.

Ising-style heuristics were worse

Rule (a), per connected component; accuracy over all nodes.

Solver	Optimum reached, 3B / 8B	Accuracy, 3B / 8B	Median time per component (3B)
RC2 (exact)	105/105 · 90/90	79.5 / 91.3	0.12 ms
Simulated annealing	76/105 · 71/90	71.1 / 82.5	3.6 ms
Simulated bifurcation, discrete (tuned)	72/105 · 74/90	66.5 / 75.4	12.7 ms
Simulated bifurcation, ballistic (default)	23/105 · 10/90	54.9 / 63.7	–

On merged instances annealing beat merged RC2 on time at 22,800 propositions (10.5 s vs 23.3 s), but with 2.3× the objective and 2 hard-constraint violations; discrete SB had 187. These instances are sparse, implication-dominated and full of hard clauses, the regime where exact solvers' unit propagation wins.

Pre-registered hypotheses

	Prediction	Verdict
H1	Evidence priority has the lowest coherently-wrong rate and the highest accuracy	Not supported: it beat confidence and random on coherently-wrong rate but lost to entrenchment on both metrics
H2	Confidence beats random on accuracy	Supported for 3B only (+1.3 pp); a negligible effect

	Prediction	Verdict
H3	Entrenchment beats random	Supported (+7.6 and +4.4 pp)
H4	Repair gives $\geq$ 99% consistency, $<$ 5 pp gain, $\geq$ 25% correct→wrong flips	Partly: consistency held; the gain was larger (+8.5, +13.1 pp); the flip share held for 3B only
H5	$\geq$ 30% of chains end coherently wrong under confidence repair	Supported (65.0%; 8B 31.2%, at the threshold)
H6	LLM stages $\geq$ 99% of time; relation checking $\geq 100\times$ the solve	Supported ($>$ 99.998%; about $10^5\times$)
H7	SA and SB reach the RC2 optimum on $\geq$ 95% of components, within $\pm$1 pp accuracy	Not supported on quality (69–82%; -8 to -16 pp); RC2 was also faster
H8	LLM-judged constraints help less than gold logic	Supported

No kill criterion fired: there were no log-prob fallbacks in 11,310 elicitation calls, and Q-form accuracy (82.4% and 90.5%) stayed inside the pre-registered range.

What this does and does not show

- Choose what gives way by structure, and check where the model is weak. Confidence-ordered retraction tied random, as low-confidence targeting failed in earlier NeuroGenesys work at 1.5B. The structural rule won here because the least-connected beliefs (2-hop conclusions) were also the ones the model got wrong.
- The coherently-wrong rate is now measured. It is the default outcome of repair, 65% (3B) and 31% (8B) under the usual rule, not an edge case.
- Physical solvers would not change end-to-end cost. The solve is about 0.001% of it, and the digital Ising stand-ins also lost on quality. Scale should come from decomposition.
- Limits. One 2-hop composition template over functional relations; in stores where conclusions are hubs, or premises are the weak point, the ranking could reverse. Two small quantised Llama models, one deterministic run, log-prob confidence only. CPU-only on a shared machine, with the 3B and 8B runs concurrent, so absolute latencies are inflated. The 8B model ran on 250 chains and was judged on 80 (3B on 120). The negated form dominates the consistency metric (no chain starts with all forms agreeing); the cleaner measure is that 30.7% (3B) and 37.2% (8B) of chains violate a gold rule on the canonical form alone. MQuAKE's 2022 gold values carry some noise, and shared beliefs make the case bootstrap somewhat optimistic. The judge prompt was chosen in a 5-chain pilot, and no dedicated NLI model was tested. The annealers used a large exact penalty, which is known to hurt them. The mechanism analyses (role table, atomic-first, within-chain degree, uniform weights, logic vs form agreement, the scaling test) were added after seeing partial 3B results and are labelled post hoc.

Deviations

Logged in the results file: SB defaults were changed to the tuned discrete variant before any main-run data, with default ballistic SB kept as a secondary arm; the exploratory rules and diagnostics above were

added after seeing partial 3B data, without editing the hypotheses; the two models ran concurrently with 4 threads each on a shared machine; because every chain started inconsistent, the coherently-wrong rate among initially inconsistent chains equals the overall rate; solver-comparison accuracy is over all nodes; pilot data is excluded.

Prior work

- Kassner et al. 2021, arXiv:2109.14723: BeliefBank, MaxSAT over stored model beliefs.
- Mitchell et al. 2022, arXiv:2211.11875: ConCoRD, factor-graph consistency with NLI relations.
- Zhong et al. 2023, arXiv:2305.14795: MQuAKE, multi-hop questions over Wikidata chains.
- Thagard & Verbeurgt 1998, Cognitive Science: coherence as constraint satisfaction, and its NP-hardness.
- Efe et al. 2025, arXiv:2511.21046: Ising hardware on SAT-like instances needs classical preprocessing.

Data and code

The pre-registration, code, aggregate results and full tables are in the study folder `research_notes/Logical coherence of LLM world models/experiments/E1_belief_repair/` in the NeuroGenesys repository. Raw per-call LLM outputs 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/belief-repair