Coherence is selected, anchored, then pruned: a literature synthesis on logical coherence in LLM world models

W. Schulz

Nov 10, 2026

Abstract

A literature synthesis on whether LLMs hold a logically coherent model of reality. They hold many local, blended models instead, with accuracy near ceiling and cross-question consistency near 60%. Methods that propagate new facts generate, filter and retrain; iterating that loop converges to whatever the selector rewards. Many sources are unreviewed 2025–26 preprints.

W. Schulz, NeuroGenesys, November 2026. A literature synthesis, not an experiment: the numbers are the cited authors' own and were not reproduced here. Many sources are 2025–26 preprints, several unreviewed, and some figures were read from abstracts or search snippets rather than full papers; those are flagged where they appear. Figures marked as this report's estimate are not sourced. Research notes: `research_notes/Logical coherence of LLM world models/` *in the NeuroGenesys repository.*

Since this review. Three of its proposals have now been tested. Belief repair, on llama3.2 3B and llama3.1 8B, measured the coherently-wrong rate (65% and 31% under confidence-weighted repair), found that entrenchment, not confidence or evidence priority, best decided which belief gives way, and found the solver about 10^5 times cheaper than the LLM work. Two more ran on the lab's toy knowledge-base transformer. Iterated self-closure (the anchored-versus-unanchored chewing test) converged in every arm; unanchored, it settled on a coherent but partly wrong fixed point, and anchoring plus pruning made it true. Gradient kernels and edit ripple (the kernel-plus-knowledge-base pairing) found the kernel worse than low confidence at predicting forgetting in a drifting stream, useful for predicting consequence propagation and for flagging contradictions by its sign, and harmful as a rehearsal selector. The last two are toy-scale results; the text below is unchanged from the review.

No current LLM has a single logically coherent model of reality, and it cannot get one from prediction alone. Pretraining produces many context-indexed predictive models (personas, task heuristics, local simulators). They combine knowledge by similarity-weighted blending. That machinery has no native way to detect a contradiction, apply a negation or carry an implication. The result is a stable gap between accuracy and coherence. A 2026 benchmark found a frontier model 98.9% accurate on single questions but only 60.3% consistent across logically equivalent rewrites (CRTBench). When a new fact is written into weights, its downstream consequences hardly move: 0.97 → 0.01 accuracy on facts a Bayesian agent would revise (Hase et al. 2024). A "logos" can only be built. It comes from a loop: generate the consequences of what is believed, check them against each other with an NLI model, solver or knowledge base, keep the most coherent subset weighted toward external evidence, forget what loses, write the survivors back, and stop on held-out ground truth. Iterating that loop ("chewing") does converge, but to whatever the selector rewards. With no external anchor it converges to a self-consistent error such as collapse, sycophancy or a template answer. With an anchor it converges toward truth, within

the limits of what can be checked. The formal limits are computational, not Gödelian: exact consistency checking is NP-hard and deductive closure is infinite. So the only achievable target is bounded coherence: no detectable contradiction within a stated budget of checks. Two parts of the question have narrow answers. Kernels predict which existing beliefs a new fact will *move*, not which it *should* move. Physical computing speeds up the solver, which is the cheap step; the bottleneck is extracting beliefs and relations with LLM or NLI calls. For NeuroGenesys this reframes the "logos" problem as a selection problem: choosing which belief survives a contradiction is forgetting-as-selection, and the open question is which selector to use.

An LLM holds many local world models and blends rather than checks

"World model" means at least five different things in current AI. Only one of them makes logical coherence an explicit goal. Treating them as one thing is the main source of confusion in the "do LLMs have world models?" debate.

Kind of world model	What it predicts or scores	Judged by	Examples
Simulative	next state or observation given state and action	rollout accuracy	Ha & Schmidhuber, DreamerV3, Genie 3
Decision-equivalent	only reward, value and policy	planning performance	MuZero
Energy / representation-predictive	compatibility of x and y in latent space; inference minimizes energy	prediction in embedding space	JEPA, V-JEPA 2
Implicit in a sequence model	whatever next-token prediction needs	probes and behavior	LLMs, Othello-GPT
Declarative belief set	propositions plus entailment and contradiction relations	consistency and truth	ECHO, knowledge bases, AGM belief revision

The first four are trained on prediction error, reconstruction or reward. None has a loss term for non-contradiction across propositions. Where "consistency" appears, it means something else. DeepMind's Genie 3 keeps generated worlds "largely consistent for several minutes," with visual memory reaching back about one minute, and that consistency is emergent rather than built in (DeepMind). This is spatial persistence, not logic. MuZero's latent state need not match the real environment at all, only its value predictions (Schrittwieser et al.). The energy-based family is the conceptual bridge. In LeCun's JEPA framing, inference means *minimizing an energy over latent variables* (LeCun 2022). "Low energy" is formally the same as "high coherence" in Thagard's constraint-satisfaction model of explanatory coherence (Thagard 1989). A logos in the sense the question uses is the fifth kind: a declarative belief set that settles. An LLM has to have one built on top of, or distilled into, its implicit model.

The implicit LLM world model is real, but it is local. Othello-GPT develops a board state that probes can decode and that interventions can change (Li et al.). Llama-2 linearly encodes geographic coordinates and dates (Gurnee & Tegmark). Vafa and colleagues showed that decodability is not coherence. A transformer trained on shortest-path Manhattan taxi routes scored 1.00 on next-token validity and 0.91

on a state probe, but only 0.10 compression precision under Myhill-Nerode tests, which check whether the same state reached by different paths yields the same continuations. With 1% random detours its valid-route rate fell to 0.69, and at 10% to 0.08. A model trained on random walks stayed at 1.00. The Othello model trained on championship games had compression precision 0.00 against 0.98 for one trained on synthetic games, even though the two looked equal on standard diagnostics (Vafa et al., NeurIPS 2024). Their inductive-bias probe found an orbital-mechanics transformer that predicts trajectories well while implying a force law with "no resemblance to Newtonian mechanics" (Vafa et al., ICML 2025). A February 2026 rebuttal argues that three inductive biases (spatial smoothness, spatial stability, temporal locality) are enough to recover the right law (Liu et al.). That moves the debate from "can coherent world models emerge" to "which priors and data are needed." Two lessons hold either way. Coherence of the implied world needs its own tests, separate from accuracy and probe decodability. And diverse data (random walks, synthetic games) produced coherent models where curated "optimal" data did not. That matches the lab's own toy result that exploration-coverage data gave 3.9× lower world-model error than narrow policy data (findings ledger 2.5). It also warns against forgetting-as-selection that prunes toward narrow, high-quality data: such pruning can hold accuracy while degrading coherence.

There is also no single model to be coherent. Anthropic's Persona Selection Model holds that pretraining learns "a distribution over personas," and post-training sharpens one Assistant persona. It explicitly does not claim that persona is "a single, coherent persona that is consistent across contexts" (Marks, Lindsey & Olah, Feb 2026). Behavior fits that view. In a 20-questions game where the model secretly holds a target, GPT-4o changed its target on 43.4% of turns in number guessing (Luo et al. 2026). So "is the LLM consistent?" is ill-posed until it names whose beliefs and in what context. A logos needs cross-context invariance of factual commitments. Role-play legitimately violates that.

The mechanism explains why contradiction is not caught natively. Softmax attention is a normalized kernel smoother, Nadaraya-Watson regression over the context (Tsai et al.). One study reconstructed in-context-learning outputs from a kernel-regression equation with over 80% accuracy (Han et al., TMLR 2025). The same update is the retrieval rule of a modern Hopfield network. When stored patterns are similar, retrieval settles into "metastable" states that *average* over them, and many BERT heads operate in that averaging regime (Ramsauer et al.). Weight matrices behave as outer-product associative memories over near-orthogonal keys (Bietti et al.). Put two conflicting statements into this machinery and the output is a similarity-weighted mixture, with no operation that flags a contradiction. Any contradiction detection has to come from learned nonlinear computation on top of the blend, or from outside the model.

Accuracy near ceiling coexists with coherence near 60%

Consistency is measured by asking logically linked questions and checking that the answers obey the link. The linked forms include paraphrase, negation, contrapositive, transitivity, implication, symmetry, and probabilistic identities such as $P(A) + P(\neg A) = 1$. Every family of check shows the same gap. In CRTBench, GPT-5.4-mini answered contrapositives consistently only 72.4% of the time and double negations 84.6%, against 94–100% for surface rephrasing. Reasoning effort helped unevenly. o4-mini reached 96.9% family consistency, and higher effort lifted GPT-5.4-mini to 85.4%. GPT-5.4 showed no net change, because gains on nested negation were offset by new quantifier failures (CRTBench; unreviewed preprint, 1,750 questions). Consistency also degrades combinatorially as more items are jointly constrained. On LogiCon-Bench the best model's generative consistency fell from 83.6% to 42.7% as group size grew (ICLR 2026). Scale and reasoning RL narrow the gap but do not close it. The hardest checks are the ones that require computing logical structure rather than recognizing surface paraphrase.

The most valuable result in this literature is that consistency predicts truth when no ground truth exists yet. Paleka and colleagues scored LLM forecasters by arbitrage, meaning how much a bettor could earn by trading against incoherent probabilities. Those scores correlated with Brier scores that resolved later. The same paper found that attempts to *improve* consistency tended to overfit the specific checks used (Paleka et al., ICLR 2025). Taken together, the two findings give the operating rule for using consistency as a reliability signal. It is a valid ground-truth-free proxy, and it gets Goodharted as soon as it becomes the target. Check families have to be rotated or held out.

Sampling consistency, as used in SelfCheckGPT and semantic entropy, measures something weaker: *stability*. Semantic entropy flags confabulations with average AUROC 0.790 against 0.691 for naive entropy (Farquhar et al., Nature 2024). By construction, though, it cannot catch an error the model makes every time. Mündler and colleagues found self-contradictions in 17.7% of ChatGPT sentences. 35.2% of those could not be checked against online text at all, so consistency checks catch errors that retrieval misses (Mündler et al., ICLR 2024). A 2026 result shows why stability is not enough. Facts answered with *perfect* self-consistency can collapse under mild contextual interference, while facts that are coherent across a neighborhood of related questions resist it (Neighbor-Consistency Belief, 2026). For forgetting-as-selection, neighborhood coherence is a candidate test for which facts are load-bearing (keep them) and which are isolated and brittle (re-verify or retire them).

Inside the network there is a second, white-box channel. True and false statements separate linearly at scale, and moving activations along that direction flips how the model treats a statement (Marks & Tegmark). Earlier probes failed on negations because truth lives in a two-dimensional subspace: one general truth direction plus a polarity-dependent one. Modeling both lifts lie detection to 95% on simple lies and 82% on real-world lies (Bürger et al., NeurIPS 2024). Two cautions matter. First, unsupervised discovery is unreliable. Contrast-Consistent Search enforces $p(x) \approx 1 - p(\neg x)$, but arbitrary binary features satisfy that constraint, and CCS latches onto salient non-knowledge features (Farquhar et al., DeepMind). Second, probes appear to track the model's *belief*, not the truth, and those belief signals predict whether the model retracts an answer (2025 retraction study). That is exactly what a coherence check wants. It also gives a useful diagnostic split. When output and probe disagree, the failure is unfaithful *reporting*: sycophancy, prompt pressure or deception. When probe and world disagree, the failure is *knowledge*. These call for different fixes: faithfulness training in the first case, filtering or forgetting in the second.

Philosophy supplies the integrating standard. Herrmann and Levinstein argue that a representation counts as a belief only if it meets four criteria jointly: accuracy, coherence, uniformity across contexts, and use, meaning it causally drives behavior (Minds and Machines 2025). On this account coherence is part of what makes something a belief, not just a quality score. Each criterion maps to a measurable test: probe against ground truth, Paleka-style arbitrage over linked probes, probe transfer across negation and domain, and causal intervention. No paper yet scores a frontier model on all four.

New facts land; their consequences and retractions do not

Belief revision gives the vocabulary for "does new information mesh." In AGM theory, *expansion* adds a belief, *revision* adds one while restoring consistency, and *contraction* removes one. Rational revision obeys minimal change and protects entrenched beliefs. Truth maintenance systems add the missing engineering piece: a record of which belief depends on which, so that retracting a premise withdraws everything that rested only on it (Doyle 1979; Almeida & Casals 2026). LLMs handle expansion well and revision-with-consequences and contraction badly, whether the new information arrives in weights or in context.

In weights, the clearest measurement is Hase and colleagues' semi-synthetic Wikidata testbed, whose gold labels come from a Bayesian agent. A rank-1 LoRA edit raised direct-edit accuracy from 0.96 to 1.00. On downstream facts that *should* change, accuracy fell from 0.97 to 0.01, and the rate of true/false coherence violations rose from 0.38 to 0.53 (Hase et al., TMLR 2024). The edit landed and its consequences did not follow. Their list of open problems names the root causes. Edits arrive as bare input-output pairs with no source or evidence ("missing context"). Gradient updates may not be rational revision at all ("no learned belief-update mechanism"). And coherence has a compute cost. NeuroGenesys has already reproduced the core pattern at 1.5B: adding new dependent facts fixed covered entities +77.4 pp and held-out ones only +1.9 pp (findings ledger 3.15). This report does not treat that as new. What the literature adds is the TMS explanation. The model derived "this person lives in the old city" as a parametric belief with no recorded link to the edited premise. Pruning the premise therefore cannot retract the conclusion, and only a trained replacement for that specific conclusion moves it.

In context, the failures are subtler but have the same shape. AGM-Bench tested seven frontier models on 2,400 scenarios. All satisfied Success and Consistency but "systematically violate" Inclusion (minimal change) and Preservation. Under iterated revision they showed "severe belief inertia (retaining retracted information)." The reasoning-trained models revised better in one step but "degrade faster under iteration" (AGM-Bench, ICLR 2026). Belief-R found a trade-off: models good at updating when a new premise defeats a conclusion get worse at maintaining beliefs when no update is warranted (Wilie et al., EMNLP 2024). How a model arbitrates between context and parametric memory depends on the shape of the evidence. A single coherent counter-memory passage is readily accepted. Mixed evidence triggers "a strong confirmation bias" toward parametric memory (Xie et al., ICLR 2024). Even so, in-context editing remains the propagation ceiling. On RippleEdits it reaches about 82 ripple consistency, against 45.5–60.8 for ROME, MEMIT and MEND (Cohen et al., TACL 2024). That is why the lab's CCU study benchmarks weights against it.

The methods that do propagate share one structure: generate implications, filter them for consistency, train on the survivors. Deductive Closure Training has the model generate statements implied by seed documents, selects the most probable *consistent subset*, and fine-tunes on it. It gained 3–26% across CREAK, MQuAKE and the Reversal Curse, and 12% on CREAK fully unsupervised (Akyürek et al., ACL Findings 2024). PropMEND meta-trains an editor on propagation questions and roughly doubles multi-hop accuracy on RippleEdits. Its margin shrinks on unseen entity-relation pairs (Liu, Durrett & Choi 2025), which repeats the lab's covered-versus-uncovered split. EntiGraph re-expresses a small corpus as relations between entity pairs, generating 455M synthetic tokens from 1.3M. That raised Llama 3 8B closed-book QA from 39.49% to 56.22%, about 80% of the gain from retrieval (Yang et al., ICLR 2025). Two 2026 results identify where integration breaks. After 20 sequential fact writes into Qwen3, earlier facts were answered 1% of the time when trained as repeated statements, against 46% with varied restatements, "regardless of training method." The forgotten facts still sat in the log-probabilities (O'Neill et al. 2026). Weight-level contraction is suppression, not deletion. In TMS terms, the weights can mark a node OUT but never remove it, so a later cue can bring it back IN; that is AGM-Bench's "belief inertia" from the other side. FACTPROP adds that correct facts about *highly connected* entities are the most likely to be corrupted by updates (Zhang et al. 2026; abstract only). That challenges any rehearsal rule that selects only by low confidence. Confident hub facts can be the most exposed.

Every working "truth maintenance for LLMs" therefore keeps provenance outside the weights. BeliefBank keeps a persistent store of model answers and uses weighted MaxSAT to revise "beliefs that significantly clash with others" (Kassner et al., EMNLP 2021). Zep's Graphiti stores facts as graph edges with validity windows and bi-temporal timestamps, and it *invalidates rather than deletes* superseded facts, tracing every derived fact back to source episodes (Graphiti). This is non-destructive AGM contraction

with a JTMS justification structure. It is also what CCU's ledger reproduces inside a training pipeline. The answer to "how can new information be checked against existing knowledge" thus has three layers, and the layers have different strengths. Weights alone can say how *surprising* a new fact is but cannot say what depends on it. A belief readout (probes) can check local coherence of what the model currently holds. Only an external store with provenance can perform dependency-directed retraction.

Solvers buy coherence cheaply; accuracy needs independent evidence

The inference-time recipe that reliably raises coherence is: extract commitments, detect relations with an NLI model or a symbolic translation, and solve for the most coherent assignment. ConCoRD builds a factor graph over candidate answers for a batch of questions, using model likelihoods and NLI entailment and contradiction scores, and picks the jointly most coherent answers with MaxSAT. It needs no fine-tuning and added +5% absolute accuracy for LXMERT on ConVQA (Mitchell et al., EMNLP 2022). REFLEX builds a belief graph by backward chaining and improves consistency "by 8%-11% absolute without harming overall answer accuracy" (Kassner et al. 2023). The 2026 version extracts symbolic commitments from each answer and checks the accumulated state with SAT/SMT. Unsat cores localize conflicts, and a bounded repair loop runs on them. DeepSeek-R1's set-level satisfiability rose from 0.56 to 0.94, and accuracy from 0.80 to 0.85, at 1.55× cost. Plain self-consistency with 20 samples reached only 0.63 satisfiability at 2.75× cost. Most of the gain (+0.30 to +0.35) came from checking alone (Salla et al. 2026). The paper has real weaknesses. Commitment extraction was 92.75% accurate, solver timeouts left 16–19% of cases "Unknown," only 50 of 390 labels were actually solver-verified, and the same solver both enforced and scored consistency. The direction of the result is still robust across the whole family. Consistency gains run several times larger than accuracy gains.

That gap is the central risk. A solver can make a belief set coherent by flipping the *correct* belief when the wrong one carries higher model confidence. The 2026 paper's rule of retracting the fewest past commitments also entrenches early errors. This is the "confident wrong vote" problem moved onto belief graphs. No paper reports how often consistency repair resolves conflicts toward the false side, a quantity this report calls the coherently-wrong rate. It is the most important unmeasured number in the field. Formal-logic pipelines move the trust problem rather than removing it. Logic-LM and LINC give large gains on synthetic deduction benchmarks (Pan et al.; Olausson et al.). But about 17% of LINC's FOLIO generations fail to parse (2026 translation study). A 2025 reality check found LLM-as-formalizer *underperforming* LLM-as-solver in 15 of 24 model-dataset combinations, and reasoning models sometimes wrote the answer straight into the "formalization" (Reality Check 2025). That is a verifier exploit in its purest form.

Self-correction separates the methods that work from those that don't. Huang and colleagues found that without external feedback LLMs "struggle to self-correct," and "at times, their performance even degrades" (ICLR 2024). Kamoi's survey found no successful self-correction from prompted-LLM feedback outside specially suited tasks. Self-correction works with reliable external feedback or large-scale fine-tuning (Kamoi et al., TACL 2024). Multi-agent debate's gains come mostly from its majority-vote step. Debate alone forms a martingale over agents' beliefs, with no expected improvement (Choi, Zhu & Li, NeurIPS 2025). What works, then, is *independence*: independent samples, independently answered verification questions, executed tests, solver feedback. The lab reached the same conclusion from the selection side, where a reject threshold worked only on a score independent of the expert being judged (findings ledger 9.3). The lab has also already seen the set-level advantage. Deciding over a batch of related inputs cut selection error from 0.095 to 0.002 (findings ledger 9.9). Deciding jointly over a batch is the same move ConCoRD makes. Reasoning models help *within* a context: o4-mini detects inconsistent relations in a prompt 98.95% of the time, against under 15% for GPT-4o (2025 self-consistency

study). They do not secure coherence *across* separate queries, which still needs memory plus a solver.

Coherence can also be trained in, but only for the constraints it is trained on. LoCo-LMs fine-tune with a semantic loss, the negative log of the probability mass on assignments that satisfy compiled logical constraints. A 770M model reached 0.99 implication consistency, against 0.91 for ConCoRD's post-hoc MaxSAT, while plain cross-entropy on the same data collapsed to 0.03 F1. The implication-only variant scored 0.00 on negation consistency (Calanzone, Teso & Vergari, ICLR 2025). Coherence does not generalize across constraint types, and the method assumes its rules are correct. Anthropic's Internal Coherence Maximization goes further. It searches for a self-generated label set that is logically consistent and mutually predictable, fine-tunes on it, and matches gold-label training on GSM8K verification, TruthfulQA and reward modeling. Its v2 claims it beats human labels on strongly superhuman tasks (Wen et al. 2025–26). An ICML 2026 paper treats answer-distribution drift under repeated asking as a martingale violation and trains it away with no loss of accuracy (Kim et al.). ICM is, in effect, selection under a coherence fitness. It inherits CCS's weakness, since a coherence objective finds the most salient consistent feature and not necessarily truth. Two questions have no study yet: whether RL with verifiable rewards on math changes off-domain negation or implication consistency, and whether anyone has trained on a negation-consistency reward directly.

Chewing converges, but the selector decides to what

Gödel's theorems do not block a coherent world model. They apply to effectively axiomatized systems strong enough for arithmetic, and they rule out completeness and self-proved consistency, not consistency itself (SEP). The binding limits are computational. Propositional consistency is NP-complete (Cook 1971). Deductive closure is infinite, which is the logical-omniscience problem no bounded agent escapes (SEP). Thagard and Verbeurgt proved that maximizing coherence, partitioning beliefs into accepted and rejected so as to maximize satisfied constraints, is NP-hard (Cognitive Science 1998). The realistic target is the one logical induction formalizes. Credences come to resist exploitation by efficient checkers over time and approach coherence only in the limit (Garrabrant et al.). Stated operationally, a model has a logos *relative to* a checker class, a query budget and a constraint language. It holds no Dutch book on the queried sets and is martingale-stable under re-asking. Paraconsistent tolerance, which localizes contradictions instead of letting them explode, covers the remainder (SEP). LLMs already sit far from even that bound: GPT-4 and LLaMA-2 violate probability identities in human-like ways (Zhu & Griffiths).

Coherence is not truth, and philosophy settled why long ago. The isolation objection to coherentism says a fully coherent system can be cut off from the world, and the alternative-systems objection says many incompatible systems can each be coherent (SEP). Thagard's ECHO answered both with data priority: evidence units receive extra activation, so the settling process cannot reject observations as cheaply as hypotheses (Thagard 1989). No LLM training loss implements data priority. LoCo has no evidence term, and ICM's mutual predictability is internal. That gap is the precise formal reason unanchored consistency training is dangerous.

Every classical settling mechanism converges under conditions that real belief sets break. Its fixed points are local, and it converges by *suppressing* what doesn't fit. AC-3 always terminates because domains only shrink, but it guarantees only local arc consistency (Mackworth 1977). Hopfield dynamics always descend in energy, but they land in spurious mixture attractors as well as stored memories (Hopfield 1982). Crick and Mitchison proposed REM sleep as "reverse learning" that removes such parasitic attractors (Nature 1983), and Hopfield, Feinstein and Palmer showed that unlearning stabilizes stored memories (Nature 1983). Loopy belief propagation can oscillate or settle on initialization-dependent

fixed points (Heskes 2002). Deep equilibrium models need Jacobian regularization to converge reliably (Bai, Koltun & Kolter 2021). Recurrent-depth language models such as Huginn-3.5B improve with more latent iterations (Geiping et al. 2025), but independent probing found only marginal gains and no interpretable latent reasoning (2025 probe). Two design rules follow. Convergence is guaranteed when the process is monotone: an energy that cannot rise, or a store that only prunes. And the suppressed set (ECHO's negative units, AC-3's deleted values, unlearned attractors) is the mechanism of convergence, not a side cost. Forgetting-as-selection is built into settling.

The evidence on offline "chewing" says it helps when it makes implicit relations explicit and an anchored selector filters the output. EntiGraph's entity-relation re-expression scales log-linearly, while repeated paraphrase "quickly" saturates (Yang et al.). SEAL reinforces only the self-generated implications that improve held-out, no-context QA, reaching 47.0% against 46.3% for GPT-4.1-generated data (Zweiger et al. 2025). STaR keeps only rationales that reach the correct answer (Zelikman et al.). BoLT's EM over latent thoughts lifted data-constrained MATH from 5.7% to 25.4% (Ruan et al. 2025). Sleep-time compute pre-digests context and needs about 5× less test-time compute for equal accuracy (Lin et al. 2025). None of these papers measures whether chewing raises *logical consistency*. They report QA accuracy only, which is the key missing measurement.

Without an anchor, the documented end state is a self-consistent error. Recursive training on a model's own outputs erases the tails of the distribution (Shumailov et al., Nature 2024). Majority-vote self-reward rises, peaks, then collapses into high-entropy text followed by one template answer, which is *perfectly self-consistent and wrong*. A small validation set reliably detects the peak (Shafayat et al. 2025). Accumulating real data instead of replacing it keeps error bounded (Gerstgrasser et al. 2024). RLHF models drift toward agreeing with the user (Sharma et al.). EM gives the right theoretical frame. It never decreases the observed-data likelihood but converges only to a stationary point, and its value comes entirely from the likelihood term tied to observed data (Wu 1983). Shumailov's "tail loss" is forgetting-as-selection with the wrong selector: an unanchored pass prunes minority truths that conflict with the majority. The lab has two small instances of the same pattern. Making 30% of generated supplements wrong cut retention 19.8 pp (findings ledger 3.12). At 1.5B, a consequence-aware pipeline that rehearsed the model's *own* answers did worse than gold replay, with a 30.7% stale rate, under the untested explanation that it was re-entrenching old values (findings ledger 6.5). The claim that unifies these results, and that NeuroGenesys is placed to test, is that the selector, not the forgetting, decides whether consolidation converges to truth. Ranked by evidence, the anchors are: accumulate real data, filter by verifier or ground truth, stop early on held-out data, clamp evidence ECHO-style, cross-check across models, and gate consolidation on provenance.

A full re-chew is also expensive. EntiGraph's roughly 350× amplification, applied to a billion-token knowledge base, would mean hundreds of billions of generated tokens, on the order of a pretraining run. That is this report's extrapolation from the published ratio, not a reported cost. Feasible consolidation is incremental. It touches only new facts and their conflict neighborhood, as arc consistency does, and it runs a held-out stopping probe every pass, which is the cheapest safety spend available.

Kernels map coupling and physics accelerates the cheap step

Kernels are useful for this problem in one specific way. The empirical neural tangent kernel and its cheap proxies predict, before or during an update, *which stored facts a new fact will drag along or disrupt*. GradSim, the cosine between the gradient of an edited fact and that of a related fact, correlates strongly with ripple-effect success. Its three failure classes, negation, over-ripple and multilingual, all have very low GradSim, and facts stored in similar parameters ripple onto each other "even when the

two are logically unrelated or contradictory" (Qin et al., EMNLP 2024). CLaRE scores entanglement from single-layer forward activations. It reports 62.2% better Spearman correlation with observed ripple effects than gradient baselines, at 2.74× the speed, and proposes building preservation sets this way (Baser et al., ACL Findings 2026). A June 2026 function-space theory derives a forgetting predictor from a cross-task kernel before any gradient step, with forgetting concentrated in a few old-task NTK eigenmodes (Hidekel & Raviv). Forecasting which examples will be forgotten and replaying them beats random replay (Jin & Ren, ICML 2024).

The same results mark the limit. Kernel coupling is not entailment. A positive-definite kernel is symmetric, so it cannot score A → B differently from B → A, which is the reversal curse at the kernel level. Anthropic's influence-function work found that influence decays to near zero when key phrases are reordered (Grosse et al. 2023). P and ¬P share nearly all their tokens, so similarity treats contradictions as neighbors. Global consistency is a constraint over sets, and no pairwise kernel can express it. The lab's own toy results agree. Gradient-cosine rehearsal targeting did *worse* than random (−1.5 / −2.6 pp), similarity-based forgetting predictors reached AUROC only 0.65–0.69, and the Jin & Ren forecaster scored 0.573 (findings ledger 3.8, 3.11, 4.7). A kernel used alone as a rehearsal selector is therefore not a promising lead for NeuroGenesys. The untested idea is to pair a model-side coupling kernel (what *will* change) with a world-side knowledge-base relation (what *should* change). The pairing gives four cells. "Entailed but low-kernel" items are the predicted propagation failures and need explicit consequence training. "Independent but high-kernel" items are predicted collateral damage and should be rehearsed. "Superseded" items are pruned whatever their kernel score. "Neither" is ignored. Logic-aware geometry can supply the missing asymmetry. Order embeddings score entailment asymmetrically (Vendrov et al.), and cone embeddings handle negation (Zhang et al., NeurIPS 2021). On the reliability side, conformal filtering over a *deducibility graph* reaches 90% factuality under a strict coherent definition while keeping at least 80% of claims (Rubin-Toles et al. 2025). That is the closest published template for reliability guarantees over dependent claims rather than independent ones.

Physical computing fits the problem in form but not in practice. Weighted MaxSAT over beliefs maps onto Ising and QUBO problems, and relaxation of a physical system finds energy minima. But published coherence instances are tiny and structured. ConCoRD's batches have tens to hundreds of variables, which the RC2 solver handles "efficiently" (Mitchell et al.). The headline physical speedups were measured on the wrong problem class. The NTT coherent Ising machine solved 100,000-node Max-Cut 1,000× faster than CPU simulated annealing (NTT 2021). Belief coherence is sparse, dominated by implications, and full of hard constraints, which is where CDCL solvers' unit propagation wins. A 2025 study on a real CMOS Ising chip concluded that Ising machines "cannot exploit backbones or detect unsatisfiability" and need classical preprocessing to reach even 190 variables (Efe et al.). That also means they cannot produce the UNSAT certificates and minimal conflict sets that *checking* requires. Thermodynamic chips are pre-commercial. Extropic's scaled Z1 targets early access in 2027 (Extropic), and its energy claims rest on projections (Sacra). Intel's Hala Point "100× less energy" figure is a vendor claim about an unnamed domain, on a non-commercial prototype (Intel).

The decisive point is where the time actually goes. Pairwise NLI over 10^4 beliefs is 10^8 transformer calls before pruning. With nearest-neighbor pruning it might fall to about 2×10^5 calls, minutes on one GPU, while the resulting sparse MaxSAT solves in seconds. These figures are this report's estimate, since no published pipeline reports the timing split. On that estimate, even a 1,000× faster physical solver would change end-to-end time by well under 1%. Physical substrates become interesting in only two cases, both 3–5+ years out. One is continuous re-settling of a 10^6+-belief store under streaming edits. The other is *sampling* many coherent worlds to get calibrated marginal credences, the workload p-bit and thermodynamic samplers target (Nikhar, Kannan & Camsari). For a small lab, the physics-inspired

algorithms (block Gibbs, parallel tempering, simulated bifurcation) run on rented GPUs or the M2 Max, and the hardware money is better spent on NLI throughput.

Five tests a two-machine lab can run now

The directions below were chosen because each fills a measurement gap the notes identified, extends a NeuroGenesys result rather than repeating it, and fits an M2 Max for 0.5–3B models plus rented single GPUs for 7–8B. One engineering constraint applies to all of them. On the M2 Max at 1.5B, identical re-runs varied by −8 to +6 pp per seed (findings ledger 7.8), so every comparison needs at least three seeds and pre-registered thresholds above that noise floor.

Test	Question it answers	Setup and compute	Primary metric and kill rule
Entrenchment-rule ablation	When repair forces a belief to give way, which selector minimizes coherently-wrong resolutions?	BeliefBank and EntailmentBank constraints, 1–3B model, LoCo semantic-loss repair, with the yielding side chosen by lowest confidence, retrieval support (BM25 over Wikipedia), entity degree, or random. LoRA, a few A100-hours.	Coherently-wrong rate alongside negation/implication consistency and factuality F1. Kill if evidence-weighted entrenchment does not beat random on coherently-wrong rate.
Anchored versus unanchored chewing	Does iterated consolidation of a declarative KB converge to truth or to a self-consistent error, and which anchor decides?	DCT-style generate→select→fine-tune passes on a 1.5B model over a fixed fact set, with arms: no anchor, accumulate-real, KB-verified filter, held-out early stop. M2 Max.	Consistency *and* accuracy each pass. The prediction is that the unanchored arm's consistency rises monotonically while accuracy rises then falls.
Kernel × KB four-cell logging in CCU	Does a coupling kernel, conditioned on the KB relation, predict per-probe flips where kernels alone failed in the toy?	Log CLaRE activations or LoRA-gradient GradSim for every (edit, probe) pair in the CCU Phase 1 dev pilot. No new arm; record as a §17 exploratory deviation.	AUROC per cell, against the toy baseline of 0.65–0.69. Report negative results.

Test	Question it answers	Setup and compute	Primary metric and kill rule
Coherence-selected Best-of-N	Does choosing the most internally consistent answer *set* across linked questions beat plurality and judge selection, and can it be gamed?	Paleka-style linked forecasting tuples or BeliefBank clusters, N sampled answer sets, MaxSAT/ICM-style set selection versus the plurality-over-top-half rule. Inference only.	Accuracy and calibration, plus exploitability: the win rate of injected consistent-but-wrong sets. Emit as `selection/v1` events for the reliability product.
Off-domain consistency after RLVR	Does GRPO on math change negation and implication consistency on general facts?	1.5–3B model, GRPO on GSM8K versus matched SFT, measured on BeliefBank and ParaRel before and after. One rented GPU with TRL.	Change in negation/implication consistency. No prior study exists, so either sign is publishable.

The first two tests matter most, because together they test this report's central claim. The entrenchment ablation also follows up an anomaly the lab already found. Lowest-confidence rehearsal beat random in the toy but was worse than random at 1.5B (−6.7 pp, findings ledger 3.16), and FACTPROP's hub-corruption result suggests degree-based entrenchment as the missing comparison. The chewing test is the first direct measurement of "coherent but wrong" drift on a declarative knowledge base, a gap every note flagged. The coherence-selected Best-of-N test extends the lab's finding that a plurality vote over the top-scored half beat top-1 selection in all six math cells (findings ledger 8.14). It moves that finding from per-item selection to set-level selection, and its exploitability arm reuses the judge-exploitability protocol. Two cheaper add-ons fit inside these runs. One is the Kim et al. martingale-drift metric as a label-free early warning of interference during sequential fine-tuning. The other is a cross-persona invariance check: do the same 5–10k facts survive under five system personas? It tests whether consistency training generalizes through logic or through a shift toward a "careful Assistant" persona. The lab should not spend on Ising, thermodynamic or neuromorphic hardware for this program. An RC2/PySAT baseline on the laptop and a GPU parallel-tempering sampler for credences cover everything those machines could add within five years.

Conclusion

Answering the "logos" question changes what the problem is. Whether an LLM *has* a coherent model of reality turns out to be the wrong question. The right one is what selection process maintains coherence and what anchors that process to the world. The model supplies candidates and a weak internal belief readout. A checker supplies the constraints. A selector decides which belief yields. Forgetting carries out the decision. Each step has a measured failure mode: extraction errors around 7%, coherence gains outrunning accuracy, self-consistent collapse without anchors, and suppressed beliefs that come back. This is the same structure NeuroGenesys found in continual learning, where selection error, not forgetting, carried most of the variance between methods (findings ledger 8.15). The two programs are one program, and the lab's reliability product already has the right vocabulary for it.

The most consequential gap is the coherently-wrong rate: the fraction of repaired inconsistencies resolved toward the false belief. Every method surveyed reports consistency going up. None reports which side won. Until someone measures that rate, "more coherent" cannot be read as "more correct." It is a cheap, decisive number for a small lab to publish first, and a direct test of whether evidence-weighted entrenchment, ECHO's old idea of data priority, is the missing ingredient that turns chewing from self-agreement into learning.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/logical-coherence-llm-world-models