JUE (*jue*): AN ENERGY-BASED CONCEPT FIELD FOR ONLINE RELATIONAL MEMORY ON FROZEN TRANSFORMERS

WHAT ADDRESSING SUFFICES FOR, WHAT PROPAGATION ADDS, AND A MATCHED-INTERFACE PROTOCOL FOR TELLING THEM APART

Anonymous authors
Paper under double-blind review

ABSTRACT

A deployed language model is frozen: retrieval and fine-tuning augmentations fetch or overwrite information, but nothing accumulates inside a system that could be said to be learning. We present **Jue**, an energy-based concept field attached to a frozen transformer that accumulates knowledge online through local Hebbian updates—no gradients reach the base model, and every contact with language is governed by explicit, discrete provenance contracts at the input and readout boundaries. Disabled, the system is exactly the base model. All results are reported under a falsification-first protocol: preregistered thresholds and conclusion wordings, frozen materials, constant provenance with automatic staleness, and a permanent generation-level endpoint CI; results are additionally layered by the interface at which they were measured (field level versus controller level).

The contribution is a three-act finding about what the *concepts* of such a memory must be. *Act one—naive readout fails structurally.* Across eight non-conditioned geometric readout variants (winner-take-all, top-k, and a softmax temperature sweep), accuracy on a divergent multi-relation testbed pins at exactly 0.333: the first-taught relation permanently captures a concept's output, and later knowledge about the same concept cannot reach generation. *Act two—relation-conditioned addressing suffices for disambiguation.* A relation-conditioned query key, the strongest baseline in our own ablation family, sweeps a complete six-cell capability matrix against exact tables, vector kNN, RAG-with-triples, and a learned router. The relational primitive's one measured weakness—a promotion-latency window producing a 0.16 replacement-error rate—is dissected to a single mechanistic cause and absorbed as a design input: with relation conditioning from node creation, the error falls to 0.000 with no regression elsewhere. The finding becomes the architecture. *Act three—what propagation adds is mechanistic, not an accuracy advantage.* Under isolation-merge teaching, cut-edge causality holds at the string level on $3/4$ original frozen chains and $16/18$ blind-registered chains spanning 2–4 hops, with zero observed shortcut contamination ($0/36$) and every successfully formed chain cut-sensitive; without isolation, chain formation itself collapses ($3/18$). On 20 blind-registered ambiguous-relay constructions, however, soft multi-candidate retrieval baselines reach $20/20$ where field propagation reaches $15/20$: our own preregistered comparison retires the accuracy-advantage claim. A final matched-interface comparison finds that a provenance-aware typed graph memory reproduces every measured property of the field on independent blind materials and exceeds it on provenance completeness. That comparison runs under a *matched-interface protocol* we state separately and consider reusable: identical write contract, readout, provenance and intervention interfaces, matched budget, and blind materials with both architectures frozen at registration—six conditions without which an equivalence result is uninterpretable. Three discriminative probes on ground where a continuous substrate should differ—attractor completion, geometric generalization, shortcut licensing—returned one tie (with the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

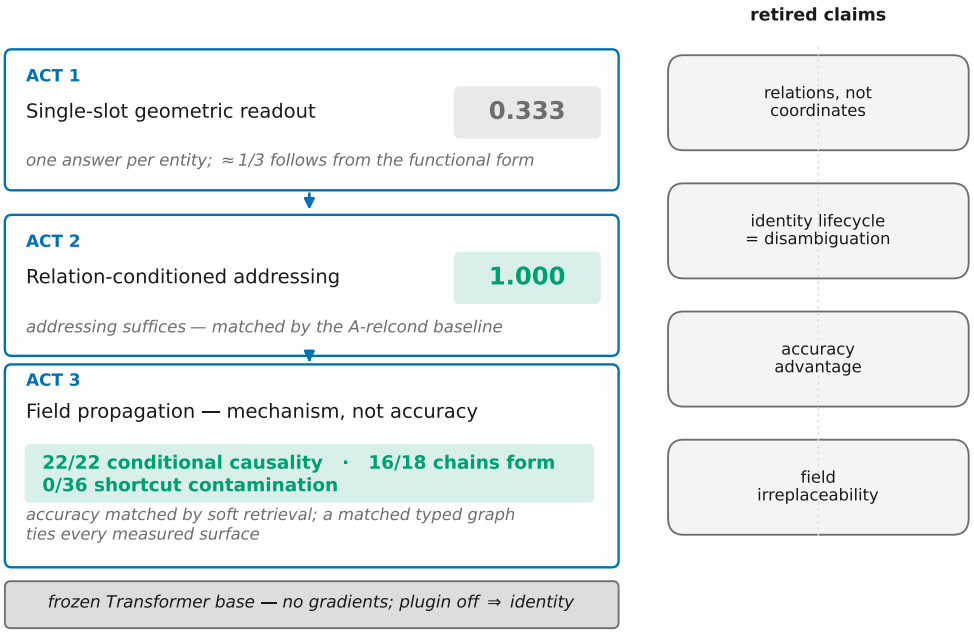

Figure 1: The three-act structure of this study. Each act reports what the architecture achieves and what it retires; the four falsified claims (right) are treated as part of the contribution rather than as removed history.

dynamics clause withheld by preregistration; a zero-relaxation ablation isolates the dynamics' contribution at $+3/117$) and two instrument blocks traceable to defects we localize here. The energy dynamics are not yet doing irreplaceable work on any surface we could measure, and the contribution of this paper is accordingly the audit discipline, these measured equivalences, and the complete falsification record.

Boundaries are reported at full strength. Field-level unification of the conditioned readout is bracketed by two preregistered failures from opposite sides—under-drive after candidate-set doubling, over-drive under a statically normalized compensation—which we formalize as the *drive–plasticity confound*: an open control problem, stated with three candidate directions and their falsification criteria, endorsed by none. Sparse real-word chains, endpoint-gain decoupling, and a distributed-pollution blind spot in the monopoly guardrail are itemized with mechanisms. A revision-phase audit found and repaired a symmetry leak between the implementation and the stated energy, with all published results triaged and reproduced; entity-scale operation remains a documented open limitation with a mechanism-level autopsy (runtime cluster entanglement under dense teaching, non-stationary teaching constants); and under conflicting updates the controller and field layers measurably diverge. The complete audit trail—nineteen stop-gates, four invoked hard stops, one retired flagship claim, and the author hypotheses struck down by the project's own data—is released as an inseparable part of the contribution.

1 Introduction

A deployed language model is frozen. Everything built to compensate— retrieval, agent scaffolds, external memory stores—works *around* that fact: information is fetched and re-presented, but nothing accumulates inside a system that could be said to have learned it. Fine-tuning accumulates, but offline, globally, and destructively. The gap between these two modes is the gap this work addresses:

can a frozen model be given a substrate that genuinely learns online—locally, without gradients to the base model, inspectably, and without forgetting what the base model already is?

We present **Jue** (*jue*, "awareness"), an energy-based concept field attached to a frozen transformer. The field holds named concept nodes whose activations relax under an energy function; knowledge is stored in learned, signed relation weights updated by local Hebbian rules; every contact with language—input coupling and token readout—is governed by an explicit, discrete provenance contract. Disabled, the system is exactly the base model. Jue draws openly on a lineage: modern Hopfield networks (Ramsauer et al., 2021) for relaxation-as-retrieval, Energy-Based Transformers (Gladstone et al., 2026) for thinking-as-energy- minimization, Hyper-SET (Hu et al., 2026) for energy as an architectural design principle, and Coconut (Hao et al., 2024) for the premise that language is not the medium of thought. It differs from all four in one commitment: they place energy and continuity *inside* the model—by training, replacing, or redesigning the transformer—whereas Jue asks what an energy substrate can do bolted *onto* a frozen one (§7).

The paper's central finding concerns a question none of that machinery answers by itself: when one concept participates in several relations, what makes a query reach the right one? Our answer, after falsifying our own hypothesis twice, is that *relation-conditioned addressing suffices*—and that the energy dynamics built to go beyond it are, on every surface we could measure, not yet doing work an explicit graph memory cannot do.

Our contributions are: **(i)** an architecture that learns relational structure online on a frozen transformer with no gradients to any base parameter, whose every taught edge is individually inspectable and cuttable; **(ii)** a three-act mechanism study—the structural failure of single-slot readout (0.333), its repair by relation conditioning (1.000), and field-level multi-hop with string-level causal ablation ($3/4$ original and $16/18$ blind-registered chains, $22/22$ conditional causality, $0/36$ shortcut contamination); **(iii)** a *matched-interface protocol* for comparing a continuous energy substrate against a symbolic graph memory—six conditions that make an equivalence result interpretable (§I.3)—and its first application, which retires our own irreplaceability claim and isolates the dynamics' net contribution at $+3/117$ by ablation; and **(iv)** a complete falsification record—nineteen stop-gates, four hard stops, four retired claims, and seven specification defects of our own—released as part of the contribution.

Section 2 opens with the design philosophy—written before any experiment and treated throughout as a set of testable claims, several of which did not survive contact intact.

2 Design Philosophy: An A Priori Chain and Its Fates

Jue began as a ten-step argument written before any experiment—algorithms over compute, interpretability as a prerequisite, continuous inside and discrete at the interface (language as protocol, not the medium of thought; a Wittgensteinian position reached independently, from an engineering direction, by Coconut (Hao et al., 2024)), an energy field, backward compatibility, unified memory. We treat the chain as testable claims; the text and a per-step verdict table are in Appendix H. Three fates carry the weight. Interpretability was *promoted* to an active design filter: a candidate energy form was rejected for corrupting causal edge-cut interventions. Dual-mode learning was *confirmed by failure*: consolidation alone produced edges two orders of magnitude too weak to propagate. The unified-memory claim was *rewritten*—"context deadlock is structurally impossible" became measured capacity and interference (§G.4), and we consider the measured boundary the stronger result. Under falsification pressure the architecture converged on the shape step six prescribes, and every representation-boundary failure we catalogue occurred where that boundary was crossed implicitly—an observation of convergence, not a proof, since philosophy and repairs share an author; what is checkable is that evidence preceded interpretation throughout (Appendix B).

3 The Jue Architecture

Jue attaches a small, continuously-relaxing concept field to a frozen transformer. The base model is never updated: with the field disabled, the system is exactly the original model. All learning happens inside the field through local, gradient-free updates; all interaction with the base model happens through three explicitly-owned boundaries—input coupling, node identity, and token readout. This

section describes the architecture in its final, validated form (v1.0). Section B recounts how each component was forced into this form by a specific experimental failure.

3.1 Energy Field

The field consists of N concept nodes with activations $a_i = \sigma(z_i) \in (0, 1)$, optimized in z-space by deterministic gradient relaxation with energy-delta early stopping (the implementation admits a Langevin noise term, but all reported runs use $T{=}0$, under which the dynamics are the exact gradient flow of Eq. 1; audit-verified per-step against the analytic gradient, §3.3). The energy is

$$E \;=\; -\tfrac{1}{2}\sum_{i\neq j} w_{ij}\, a_i a_j \;-\; \sum_i a_i \cos(v_i, \tilde{h}) \;-\; \sum_i a_i \cos(v_i, g) \;+\; \mu \sum_i a_i^2, \tag{1}$$

where $\tilde{h}$ is the centered input representation (§H.3), g an optional goal vector, and μ an activation regularizer. The interaction term deliberately contains *no* geometric factor: an earlier form, $-\sum w_{ij} a_i a_j \cos(v_i, v_j)$, multiplied learned affinity by prototype similarity, which made semantically dissimilar concepts impossible to associate strongly—the required effective pull was provably unreachable under the weight guardrail (§B). Decoupling yields the architecture's central division of labor: **geometry grounds; W stores knowledge**. A signed edge w_{ij} means association (> 0) or inhibition (< 0), independent of where the nodes sit in representation space. Relaxation over multiple rounds, with each round's terminal state seeding the next, provides test-time computation; every activation change is attributed by a CausalTrace record (§H.5).

3.2 Node Primitive, Coupling, and Learning (summary)

Four mechanisms complete the system (full specification in Appendix H). Nodes begin *ephemeral*, identified by a seed key (the centered hidden state of first mention); after $k{=}3$ distinct teaching events plus an adaptive edge-mass threshold they are *promoted* to *relational*, identified by their incident edge pattern, with the seed key frozen as an addressing key. A routing schedule (Eq. 2) transfers addressing weight from key to relation row as evidence accumulates. Language coupling grounds the field in the base model's own representation space: queries enter as centered hidden states, and an explicit distinction gate stratifies materials before any capability claim. Learning is dual-mode—weak ambient consolidation, which alone cannot build propagation-strength edges, and deliberate teach-to-criterion writes ($\alpha_{\text{step}}{=}0.2$, band $[0.6, 0.9]$) that pass a staging buffer. Five guardrails bound the system (drift $\cos > 0.95$, weight ceiling, degree cap, path decay, general-capability check); their measured blind spot is reported in §5.2.

3.3 Directionality and the Energy Function

The scalar energy sees only the symmetric part of W: $\partial E/\partial a_k$ contains $-\frac{1}{2}\sum_j (w_{kj} + w_{jk}) a_j$, so any antisymmetric component would be invisible to relaxation. A preregistered revision-phase audit (JW-AUDIT-1, Appendix) verified the implementation against this fact and found exactly one write path that could leak asymmetry when query and answer clusters overlap, while relaxation assumed the symmetric form—an internal inconsistency between code and the stated dynamics. A minimal repair (symmetrization at the write boundary) restored exact consistency: $\|W - W^\top\|_F / \|W\|_F = 0$ on replay, per-step updates match the analytic gradient to machine precision, and energy is non-increasing at every $T{=}0$ step. All published results were triaged for contamination: four of five never materialized asymmetry (ratio $= 0$); the multi-hop chains carried legacy asymmetry up to 0.019 and reproduce exactly under the repaired code (discrete outcomes unchanged; continuous shifts $< 0.1\%$).

W is therefore symmetric by construction. Arrow notation in this paper denotes *undirected association*, and directionality is carried entirely by the readout layer: the ordered $(\text{source}, \text{relation})$ key addresses the stored answer. A reverse query produces high field activation but matches no reverse key, hence selects no anchor and generates no answer (audit probe P2); edge-cut operations zero both matrix entries. The energy stores association strength; the contract layer stores direction.

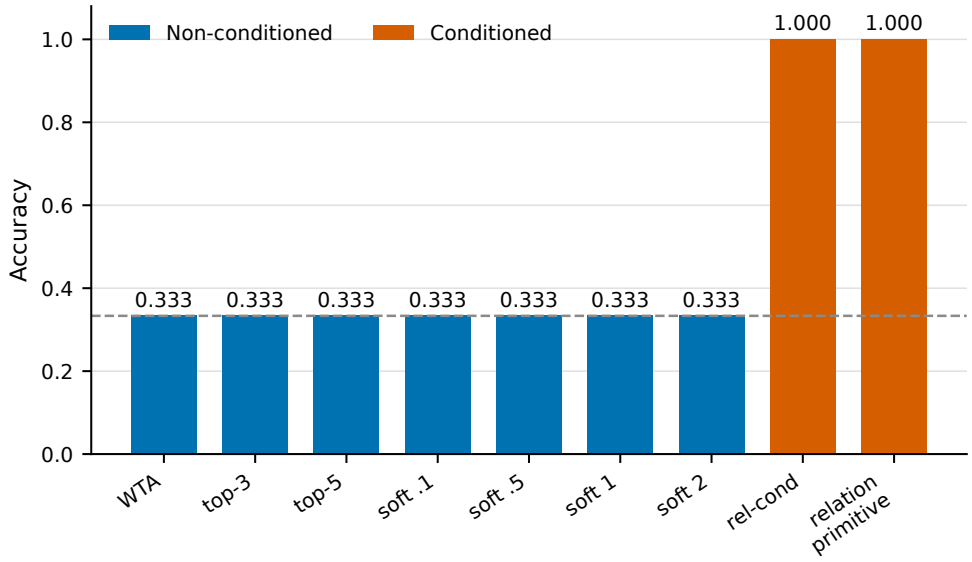

Figure 2: R6 mechanism signature (seed 19, train split). Route A answers every relation prompt about an entity with the first-taught relation's answer; Route B distinguishes all three relations per entity. Held-out paraphrase forms reproduce the same 0.333 vs. 1.000 pattern.

3.4 Provenance Readout

The output boundary is a lookup, not a projection. Every node carries a readout-anchor table: source text, the first continuation token of `tokenizer.encode("␣" + answer)`, the full answer token sequence, and an extraction commit. Token bias is activation-weighted anchor lookup; mapping hidden-space nodes through token-embedding similarity is deprecated, having scored $0/20$ on an alignment sanity check. Multi-token answers use a *sequence-forced* contract: the stored answer sequence is biased position by position at a frozen amplitude derived once from the base-model logit-gap distribution ($P_{90} + 2.0 = 12.44$) and never tuned against accuracy. The endpoint therefore verifies that an anchor can drive the complete target string; it does not claim the base model has internalized the spelling (§5). A permanent endpoint CI—teach one single-token real pair and one multi-token fictional pair, generate, and compare stripped strings—gates every capability experiment.

4 Validation

We validate Jue as a controlled mechanism study: every experiment is preregistered, runs on frozen materials, and is scored at a string-level generation endpoint. We report mechanism-level signatures with causal attribution, not benchmark statistics; the scale (a 0.5B base model, tens of associations) is deliberate and stated up front.

4.1 Setup

All experiments use a frozen Qwen3-0.5B decoder. *Route B* is the full architecture of §3; *Route A* is a matched control differing in exactly one commitment—fixed geometric prototypes with raw-cosine winner-take-all readout—sharing energy, teaching protocol, endpoint, amplitude, and materials. Route A is this architecture's previous version, repaired to pass the same coverage, selectivity, and endpoint gates before comparison.

Two frozen material sets are used. *Paired associations* (PX2B): five taught pairs with held-out paraphrase forms, plus four never-taught pairs as diagnostic-only hallucination probes. *Divergent multi-relation* (R6): four fictional entities $\times$ three relations, twelve triples plus twelve paraphrase

forms, disjoint from PX2B. All queries pass an input-distinction gate (pairwise cosine < 0.8, initial-node Jaccard ≤ 0.3, ≥ 3 activated candidates); ability judgments use train and held-out forms only. Scoring is stripped-string equality of the full generated answer under the sequence-forced contract at frozen amplitude, preceded by a permanent endpoint CI. Promotion parameters were calibrated on throwaway materials at R6 density (separation 1.0) before any capability measurement. The language path is deterministic (relaxation temperature 0), so the three seeds verify harness invariance, not statistical variation.

4.2 Supporting Surfaces

Acts one and two. On divergent materials (four entities $\times$ three relations), single-slot geometric readout answers every relation prompt with the first-taught answer: $0.333 \approx 1/3$, following from the functional form and reproduced on held-out paraphrases. Keying readout on the ordered $(\text{entity}, \text{relation})$ pair repairs it completely (1.000); an early version conditioned only after promotion mis-replaced 16% of delayed updates, and conditioning from node creation drove that to 0.000.

Supporting surfaces (full results in Appendix G). Paired associations: both routes reach 1.000 on taught and held-out paraphrase forms, so coupling addresses by meaning. Routing: the schedule of Eq. 2 transfers weight from seed key to relation row at promotion. Long-horizon: over 1000 rounds, promotion count and early-fact retention are stable while maximum degree grows unbounded, locating edge lifecycle as the next component. Synthetic dynamics: an $A{\to}B{\to}C$ chain activates C to 0.78 (threshold 0.55), and severing $B{\to}C$ suppresses it to 0.34.

4.3 What the Forced Endpoint Measures

The sequence-forced endpoint makes forced-realization accuracy track anchor selection, so we decompose readout into *M1* anchor retrieval, *M2* field-state quality, and *M3* free generation with forcing disabled. Free decoding under first-position bias rises monotonically with amplitude $(0/6 \to 2/6)$ as the answer-token logit deficit shrinks $(-8.33 \to -1.48)$; at full amplitude all five anchor-correct items place the gold *first* token at argmax, but only single-token answers complete $(2/5)$. Applying the same bias at every position without token scheduling recovers nothing $(2/6)$, so *scheduling, not bias coverage, is load-bearing* for multi-token answers. Full decomposition in Appendix I.

5 What Jue Does Not Yet Do

This section is written to the same standard as the results: every boundary below is measured, attributed, and scoped, and Table 3 binds each claim in this paper to its evidence and its edge. Nothing here is a disclaimer; several of these boundaries are findings.

5.1 Layering, Chains, and the Dead Injection Pathway

Every result carries a layer tag. *Field-level* results (the head-to-head matrix, the endpoint CI, the multi-hop causal cases) are measured through the hidden-space coupling of §H.3. *Controller-level* results (the readout ablation family, the capability matrix, the conditioned-from-birth repair) consume exact source strings above that coupling: they stratify materials by hidden geometry but cannot themselves establish robustness of the retrieval underneath. Unifying the two layers is what §K.1 reports as open, and until it closes the tags stand.

The chain boundary moved during this work and is reported in both states. Before teaching isolation, chain queries failed for both primitives; the audit attributed this to shortcut edges and relay-anchor dominance, both instrumented, and §I.4 reports $3/4$ original and $16/18$ blind-registered chains passing string-level causal ablation after the registered repairs. What remains is precise: the *sparse real-word* chain stays negative (its dense variant passes, so the boundary is sparsity of real-word clusters, not real words), and hop count is harness-declared.

Separately, the original design's residual pathway—injecting field state into the decoder's hidden stream under a scheduled gain—failed every arm of a four-arm preregistered test: with a frozen

Table 1: Remaining boundaries with their measurements.

Boundary	Measurement and scope
Layer separation	every result tagged field- or controller-level; unification open (§K.1)
Sparse real-word chains	dense variant passes, sparse negative: boundary is cluster sparsity, not real words
Hidden-state injection	four-arm test: gain locks to zero, LoRA neutralizes the signal; tested regime had redundant information
Edge lifecycle	max degree grows linearly, unbounded over 1000 rounds while stability holds
Confidence calibration	constant within regime (0.85 normal, 0.35 conflict): detection exists, calibration does not
Distributed pollution	monopoly guardrail silent while retention $\to 0.000$ at pollution 0.990
Conflict retraction	field anchor $1/10$, non-convergent $10/10$, stale edge weight 0.34; no retraction mechanism

decoder the gain locks to zero, and LoRA adaptation learns to *neutralize* the injected signal rather than exploit it. The tested regime is one in which the injected information was redundant with context; a redesigned protocol (memory facts withheld from context and adapter alike) is preregistered but deferred. Until it runs, hidden-state injection is dead under the tested protocol—a verdict independently reproduced by the geometric-generalization probe (§I.5)—and the provenance readout is Jue's only validated output pathway.

5.2 Supervision, Conflict, and Missing Components

"Online learning from language" admits four readings: (i) online database writing, (ii) online *supervised* structure writing, (iii) unsupervised continual learning, (iv) autonomous acquisition from raw language. Jue is *(ii)*: the controller supplies entity token, connection node, relation type, answer sequence, teaching target, and stopping condition. Two measurements pin this. Weak ambient Hebbian updates alone cannot build propagation-strength edges, so the supervised path is not ornamental; and the answer *suffix* lives controller-side (§I.1). Phrasing about learning from every contact with language is aspiration, not capability.

Table 1 states the remaining boundaries. One deserves emphasis: under conflicting updates the two layers *diverge*—the controller is last-write-wins ($10/10$) while the field selects the new anchor in $1/10$ cases and fails to converge ($10/10$), with no retraction mechanism.

5.3 Entity Scale: An Open Limitation with a Mechanism Autopsy

Entity-scale operation ($N \geq 50$) is not demonstrated. Five preregistered attempts terminated on instrument findings or hard gates (Appendix B); the surviving measurements localize the bottleneck in four layers, two now repaired. *(1)* A degenerate harness scorer max-pooled over shareable cluster nodes, making the query-dependent term 0.064% of the score; repaired, with production readout paths audited clean and $10/10$ regressions exact. *(2)* Material morphology produced token-level overlap; constrained at generation to Jaccard ≤ 0.2. *(3)* Overlap nonetheless regrew at runtime ($0.2 \to 0.667$). A read-only autopsy found the cause: cluster membership is fixed *before* W is written, as the top-k prototypes nearest the answer hidden state under an unconditional cosine rule, so distinct entities deterministically reuse prototypes; residual activation, allocation collision, and write-time spill were each excluded, and tightening k was ruled out by counterfactual. Exclusivity-competitive recruitment eliminated the entanglement (runtime Jaccard $0.667 \to 0.000$, clusters disjoint, regressions exact)—and anchor retrieval at $N{=}4$ nonetheless remained $1/7$. Entanglement was therefore one cause among at least two; a residual selection failure survives its removal. *(4)* Teaching constants are non-stationary across accumulated density (band hit 1.000 early-stream to 0.225 late-stream at $N{=}200$, replicated across three material generations), so a single α does not transfer, before or after the recruitment repair.

Layers (3) and (4), with the sequential-interference result ($3/18$ formation without isolation), identify *dense online writing* as this architecture's central open problem—the drive–plasticity confound

at network scale. One layer diagnosed and repaired with the gate still closed is a sharper statement than an unanalyzed failure: it tells the next attempt where not to look.

5.4 SCOPE

All results are from one frozen 0.5B base model on small frozen materials (tens of associations; four multi-relation entities); they are mechanism signatures with causal attribution, not statistics over task distributions. The language path is deterministic, so multi-seed runs verify harness invariance, not sampling variance. The sequence-forced endpoint verifies that a provenance anchor can drive a complete target string; it does not show the base model has internalized novel spellings, and tasks requiring spontaneous generation of taught surface forms are outside the validated envelope. Promotion parameters were calibrated at the evaluation's relational density; transfer across densities is untested. One further endpoint observation is reported against ourselves: outside the two-case endpoint CI, a teaching run that stalled far below the activation band still produced a correct string match—at the frozen amplitude, string correctness is partially decoupled from field activation strength, so the endpoint verifies anchor-driven control and activation-level claims must not lean on it. Finally, the original design document's benchmark projections (*e.g.*, ARC-AGI gains) are formally withdrawn: nothing in this paper supports them, and we prefer the measured envelope to the projected one.

The claim–evidence–scope bindings for every capability claim in this paper are given in Table 3 (Appendix D).

5.5 OPEN PROBLEM: THE DRIVE–PLASTICITY CONFOUND

Field-level consolidation fails from both sides. Under-driven, teaching stalls at activation 0.295 and edges form two orders of magnitude below propagation strength; normalized to reach the band, every unit overshoots and retention collapses to 0.000 under distributed readout pollution. Drive and plasticity are confounded in the current update rule, and we could not separate them within the registered stop-loss (two hard stops invoked). Three candidate directions with falsification criteria, and the full failure analysis, are in Appendix K. The dense-online-writing results of §5.3 are the same control problem at network scale.

6 METHODOLOGY: THE AUDIT DISCIPLINE

Every number above is governed by one protocol (Appendix E): thresholds, branch wordings, and stop conditions registered before execution, worst case included; constants carrying a derivation commit and regime, stale on regime change; a permanent CI validating to the consumed string; calibration materials disjoint from evaluation; retired claims recorded, not removed. Nineteen stop-gates, four hard stops, four retired claims, and seven specification defects of our own are catalogued in Appendix B.

7 RELATED WORK

Energy and associative memory. Modern Hopfield networks (Ramsauer et al., 2021) show attention implements Hopfield retrieval, so relaxation alone adds nothing a transformer lacks; Jue's departure is persistence and explicitness—attention weights are recomputed and discarded each pass, while W accumulates across episodes and every edge is individually inspectable and cuttable. Memory-augmented networks (Graves et al., 2014; Weston et al., 2015) learn read/write controllers by gradient descent; Jue writes by local Hebbian updates with no gradient to any component, which is what makes online writing possible on a frozen base. Test-time training and continual-learning methods (Behrouz et al., 2024; Sun et al., 2024) update parameters at inference; Jue's substrate is external and auditable, trading capacity for intervention. Weight editing (Meng et al., 2022; 2023) modifies base parameters directly and inherits their opacity.

Graphs, retrieval, and latent reasoning. Our matched-interface comparison places this work closest to typed relational memories and graph message passing: given the same contracts, a provenance-aware typed graph reproduces every property we measured and exceeds Jue on provenance completeness (§I.4), so we position the energy substrate as *not yet* distinguished from explicit graph

memory rather than as superior to it. Multi-hop RAG and agentic requery pipelines (Lewis et al., 2020) solve chains by extract-then-requery; our probes show a hard single-relay variant fails on ambiguous relays ($0/20$) while soft multi-candidate variants match or exceed the field. Coconut (Hao et al., 2024) reaches continuous-latent reasoning from a training direction, converging with this architecture's "continuous inside, discrete at the interface" commitment from the opposite side.

What remains structurally distinct (and untested). Our measurements license a narrow claim—no field advantage on any measured surface—and it is worth saying what that leaves open. Four properties distinguish the substrates by construction: the field is *continuous* (graded edges admitting partial match), *co-representational* (nodes in the base model's own space, queried by hidden states, where symbolic keys need a translation layer), *dynamical* (relaxation has attractors; our one qualitative asymmetry—failure by non-commitment rather than nearest-neighbour substitution—is a property retrieval lacks), and *self-organizing* (edges can grow from coactivation). A graph can acquire each, but only by implementing the corresponding mechanism. Whether any yields measurable advantage remains open: our probes ended in one tie and two instrument blocks attributable to defects we localize here.

A second reading of shortcut edges (conjecture). Semantic networks are measured small worlds (Steyvers & Tenenbaum, 2005), and learning in them includes compressing traversed long paths into direct edges—what ungoverned serial Hebbian teaching did. We removed shortcuts because causal measurement needs a clean graph; a mature memory would *license* them, provenance recording which path a shortcut compresses and edge-cut intervention verifying it carries inference rather than co-occurrence. Our probe could not test this on the field, but the graph arm separated the two classes completely ($24/24$), reframing the missing edge-lifecycle component as the governor of path compression. This is conjecture; its predictions are preregistered future work.

8 Conclusion and Roadmap

We set out to give a frozen transformer a substrate that learns, and returned with a humbler division of labor than we proposed. Addressing that carries the relation suffices for single-hop disambiguation—our own strongest baseline proves it. Soft multi-candidate retrieval is at least as good as propagation on multi-hop accuracy. And for the remaining combination—cuttable taught edges with string-level consequences and provenance—a matched typed graph reproduces all of it and exceeds us on provenance completeness. What we can show is a taught graph whose edges are individually cuttable with predictable consequences ($16/18$ chains, $22/22$ conditional causality, $0/36$ shortcutting), a boundary map with instrumented causes, two open problems measured from their failing sides, and an audit trail inseparable from the results. At the controller level Jue is a *supervised online structure-writing* system; at the field level a mechanism prototype; it is not yet a system that learns autonomously from language, and the two layers can measurably disagree under conflict.

The roadmap is preregistered: the *dense-online-writing* family this revision isolated—sequential interference, runtime entanglement (recruitment repaired, gate still closed), non-stationary constants—joins the drive–plasticity confound as one control problem at four scales, ahead of edge lifecycle, a distributed-pollution guardrail, a regime-calibrated confidence module, and the sparse real-word chain. The discriminative question returns once the recruitment residual closes, since two of three probes were blocked by it; a 9B replication is specified and scheduled. The measured envelope will grow or it will not; either way it will be measured.

REFERENCES

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024.

Alexi Gladstone, Ganesh Nanduru, Md Mofijul Islam, Peixuan Han, Hyeonjeong Ha, Aman Chadha, Yilun Du, Heng Ji, Jundong Li, and Tariq Iqbal. Energy-based transformers are scalable learners and thinkers. *International Conference on Learning Representations (ICLR), Oral*, 2026. arXiv:2507.02092.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Yunzhe Hu, Difan Zou, and Dong Xu. Hyper-set: Designing transformers via hyperspherical energy minimization. *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2502.11646.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2008.02217.

Mark Steyvers and Joshua B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78, 2005.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *International Conference on Learning Representations*, 2015.

AI USAGE STATEMENT

Large language models were used substantially in this work, in three roles. *Specification and writing*: an LLM assistant drafted the experiment specification documents (preregistered thresholds, branch wordings, stop-loss conditions) and drafted and edited the text of this paper. *Implementation*: a separate code-generation model wrote the implementation and experiment harness code. *Analysis support*: the assistant summarized returned result files and proposed candidate interpretations for adjudication.

Responsibility and decision authority rest entirely with the author. Every experimental design decision, every preregistered threshold and locked conclusion wording, every branch adjudication, every invoked stop-loss, and every retraction of a previously reported claim was made by the author, who takes full responsibility for all content of this paper including LLM-generated text. The separation

Table 2: Component birth certificates: each v1.0 mechanism and the failure that forced it. Evidence values are measured, with commits in Appendix B.

Failure (evidence)	Diagnosis	Component forced
Shared-template pooling collapses inputs (pairwise $\cos \approx 0.9997$)	field never received distinguishable input	target-token coupling, frozen centering (§H.3)
Single-letter symbols collapse to one node (all queries $\rightarrow$ node 412)	semantically vacuous tokens are indistinguishable in context space	input-distinction gate
Learned edges 0.04 vs. required pull 2.46	weak Hebb consolidates; it cannot build	deliberate-learning mode (§H.4)
Max effective pull $1.53 < 2.46$ even at the weight cap	multiplying W by prototype cosine strangles dissimilar-concept association	decoupled energy (§3.1)
Fixed-rate teaching: pull varies $25\times$ across associations, diverges superlinearly	Hebbian positive feedback makes open-loop rates uncontrollable	teach-to-criterion (§H.4)
Hidden$\rightarrow$token-embedding readout: $0/20$ top-5	another implicit cross-space projection	provenance lookup (§3.4)
Oracle bias arm scores 0%; `Nemor`$\rightarrow$`Nemesis`	wrong continuation-token convention; first-token bias cannot control spelling	sequence-forced endpoint + permanent endpoint CI
Stale pull target $[2.5, 4.0]$: feasible domain empty after energy change	constants outlive the regimes that derived them	constant provenance with automatic staleness
Geometric WTA: first-taught relation permanently captures readout (0.333)	coordinates cannot carry multi-relation identity	relational primitive (§H.1)
Promotion latency: 0.16 replacement errors, all in one pre-promotion window	lifecycle identity arrives too late for readout	conditioned-from-birth readout (§I.2)
Chain shortcut edges and relay-anchor dominance	serial teaching builds $X \rightarrow Z$ bypasses; unconditioned terminal readout	isolated-merge teaching (D-031) + conditioned terminal readout (§I.4)
Normalization double failure: under-drive 0.295, then distributed overshoot	teaching observable mixes drive and plasticity	open problem, stated not patched (§K.1)

of roles was itself part of the audit discipline described in §6: the assistant that wrote specifications did not write implementation code, and the model that wrote implementation code made no design decisions. Assistant-proposed interpretations were themselves subject to falsification and are recorded among the retired claims in Appendix B.

All specification documents, result reports, probe CSVs, and the decision registry are released with the paper (Appendix A), so that every claim can be checked against the evidence and commit that produced it, independently of how the text was drafted.

A Architecture Evolution History

B How Jue Took Shape: Falsification-Driven Design

Jue v1.0 was not designed in one pass. Between the first integration attempt and the head-to-head comparison of §4, the project stopped nineteen times at preregistered gates (twelve in the original campaign, seven across two revision campaigns); each stop produced a diagnosis with code-level or measurement-level evidence, and each major component of §3 exists because a specific failure forced it. We regard this record as part of the architecture's credibility: every component has an experimental death certificate for its predecessor. Table 2 maps failures to components; the full gate-by-gate chain, with commits, is Appendix B. Here we narrate the four arcs that taught us the most.

The blindness arc. The first full evaluation returned uniformly negative capability results—and internally contradictory diagnostics: two disjoint word sets produced digit-identical statistics. An

instrument audit traced the contradiction to input collapse: under shared-template mean pooling, six different words entered the field as nearly the same vector (pairwise cosine 0.9997). Every language experiment to that point had been administered to a blind system. The repair (target-token extraction, frozen centering) was straightforward; the lesson was not. A second collapse— literal single-letter placeholders all routing to one node—recurred *after* the first repair, because the fix had been propagated to the diagnostic harness but not the teaching harness. Both incidents are members of what we later formalized as the A6 family (§6): a representation boundary crossed on an implicit assumption of alignment.

The propagation arc. With inputs repaired, taught edges still failed to move their targets: weak Hebbian consolidation produced edges two orders of magnitude below the propagation threshold, establishing empirically that the consolidation mode cannot substitute for a deliberate-learning mode—the architecture's dual-mode design was vindicated by the failure of running on half of it. Supervised strong learning then hit a harder wall: an analytical and empirical scale audit (cross-validated to within 1%) showed that under the original interaction term $-\sum w_{ij} a_i a_j \cos(v_i, v_j)$, the maximum achievable pull with every taught edge at the weight cap was 1.53, below the required 2.46. The multiplicative coupling of learned affinity and geometric similarity made it *provably impossible* to associate semantically dissimilar concepts—precisely what associative memory is for. Decoupling ($w_{\text{eff}} = w + \lambda \cos$, with $\lambda{=}0$ selected because $\lambda{=}0.3$ corrupted edge-cut interventions) is the single deepest design change in the project, and the origin of the slogan *geometry grounds; W stores knowledge*.

The endpoint arc. The repaired system then failed everything again—including the oracle control arm that is told the correct answer and adds a direct logit bias. An oracle scoring 0% is not a capability result; it is an instrument confession. The audit found two stacked faults: the stored provenance token (bare word form) mismatched the continuation gold (leading-space form) for $20/20$ audited words, and the frozen bias amplitude (0.54) could flip $0/5$ base-model logit gaps. Repairing the convention and re-deriving the amplitude from the measured gap distribution (frozen before any accuracy measurement) still left one failure: biasing only the first token of `Nemor` let the base model's spelling prior complete it as `Nemesis`. The sequence-forced contract closed the arc, and a permanent generation-level CI now precedes every capability run. The general lesson we draw: *a pathway is validated only at its final consumed form*—a bias vector's top-5 is not an endpoint; the generated string is.

The primitive arc. With clean instruments, both the geometric and relational designs finally produced positive paired-association results—and the divergent multi-relation testbed then separated them (§G.2). Notably, the relational primitive itself had a falsification-driven correction before it ever ran: a dry run showed the promotion criterion, as first calibrated, could never fire on the evaluation topology (every node touched by exactly one event), forcing recalibration on density-matched throwaway materials. Even the winning component's parameters were set by an experiment that first proved the previous setting impossible.

Across all stops, no gate was passed by adjusting a threshold after seeing results, no experiment was rerun without independent evidence of an instrument fault, and one available knob—raising the oracle arm's amplitude to rescue a failing run—was explicitly declined as unregistered tuning. The revision sprint added a symmetric entry: two preregistered hard stops were invoked as written, and three of the authors' own working hypotheses—cluster-overlap confusion, hub-crosstalk interference, and goal-term interference—were struck down by the project's own controlled data before any repair was attempted. The falsification machinery does not distinguish between the architecture's claims and its authors'. The audit discipline that enforced this is described in §6.

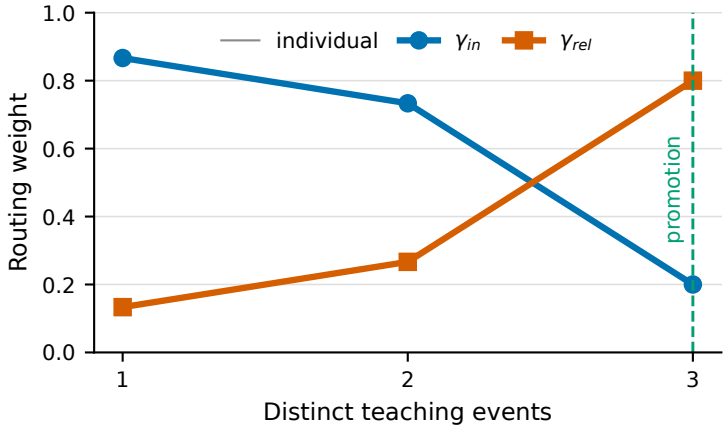

Figure 3: Routing schedule (Eq. 2) across distinct teaching events. All four R6 entities follow the identical trajectory; relations take over at promotion (k=3). This shows the designed D-022 schedule executing, not emergent dynamics.

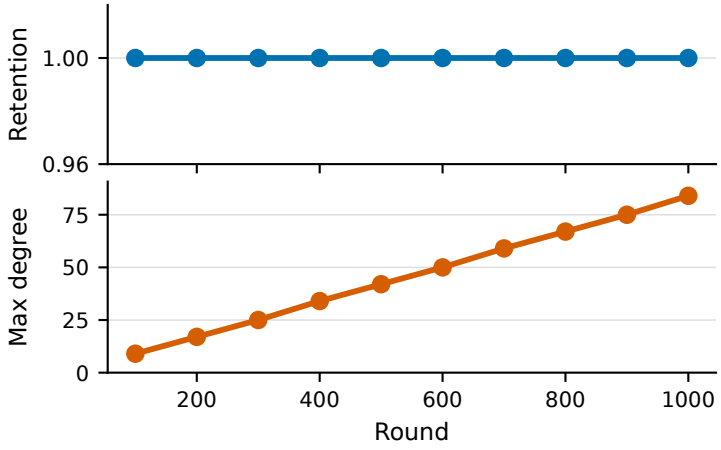

Figure 4: 1000-round mixed-stream run (seed 19; seeds identical under the deterministic path). Promotion count and early-fact retention are stable; maximum node degree grows unbounded, locating edge lifecycle as the next required component.

C Supplementary Figures

D Claim–Evidence–Scope Bindings

E Methodology in Full

The results of §4 are only as good as the instruments that produced them, and this project's instruments failed often enough that we built a discipline around catching them. We describe it as a reusable practice; the complete record is Appendix B.

Preregistration with frozen thresholds. Every gate—success criteria, stop conditions, and conclusion wordings—was written before the experiment ran and never adjusted after seeing results. When a preregistered conclusion table proved to lack a needed middle category, the gap itself was recorded rather than papered over.

Table 3: Claim–evidence–scope bindings. Every capability claim in this paper appears here with its measured basis and its boundary.

Claim	Evidence	Scope / boundary
Relational primitive beats geometric+WTA on divergent multi-relation	1.000 vs. 0.333; same-answer reuse; first-taught capture (§G.2)	mechanism signature, 4 entities; single 0.5B model; deterministic runs
No regression on paired associations	all non-chain axes 1.000 both routes; four-arm identification (§G.1)	small frozen materials; strata reported, not screened
Relations take over at promotion	identical γ trajectories, reverse gate clean (§G.3)	designed schedule executing, not emergent dynamics
Long-horizon stability with retention	1000 rounds; zero demotions; early facts 1.000 (§G.4)	degree grows unbounded; no pruning; monopoly instruments partially inactive
Field supports multi-hop propagation	synthetic $C{=}0.78$ intact, 0.34 cut (§G.5); after isolation-merge teaching, string-level causal ablation on $3/4$ original and $16/18$ blind-registered chains (§I.4)	accuracy matched or exceeded by soft multi-candidate retrieval ($15/20$ vs $20/20$); hop count harness-declared
Chain formation requires isolation	without isolation $3/18$ chains form; with it $16/18$; shortcut contamination $0/36$ (§I.4)	sparse real-word chain still negative; dense-online-writing family open (§5.3)
Endpoint validates string readout	endpoint CI; oracle arm restored to 1.000 (§3.4)	anchor-driven, not spontaneous spelling
Teaching is controllable	in-band stops; calibration separation 1.0 (§H.4)	calibrated at R6 density; cross-density transfer untested

Constant provenance and automatic staleness. Every frozen constant carries its derivation commit, the energy/parameterization regime it was derived under, and its upstream dependencies; any change to that regime marks dependents `STALE`, and experiment entry points reject stale constants at runtime. The rule exists because it was violated first: a pull target derived under one energy form was carried into another, where its feasible domain was empty.

Endpoint CI at the final consumed form. An evaluation pathway is validated only at the form a user consumes. A readout vector's top-5 passed sanity while the generated string failed; the smoking gun was an oracle control arm scoring 0%—impossible if the endpoint works. A permanent CI (teach, generate, compare stripped strings, for one single-token real and one multi-token fictional pair) now precedes every capability run.

The A6 taxonomy. Six instrument failures, initially appearing unrelated, reduce to one family: *a representation boundary crossed on an implicit assumption of alignment*—token-embedding prototypes compared with layer hiddens; template pooling collapsing lexical content; learned affinity multiplied by prototype cosine; hidden nodes projected through token-embedding similarity; vacuous symbols collapsing in context space; and stored token forms mismatching continuation gold (Appendix G). The prevention rule that closes the family: every representation space, unit, and provenance path must have an explicit owner, and every cross-boundary pathway an end-to-end sanity test at its final consumed form. The revision sprint added a second rule of the same rank, generalizing the normalization double failure: *a control loop's observable must be isolated from the mechanism it regulates*—a controller that watches a quantity contaminated by its own operating conditions will fail on both sides of its range.

One sentence from the record deserves preservation verbatim: the first end-to-end language measurement this project can call valid occurred at its final gate. Everything before it, positive and negative alike, was measured through at least one broken boundary—which is the argument for the discipline, not against the architecture.

Table 4: Capability matrix (controller level, frozen materials), positioned as a *mechanism diagnostic*, not a comprehensive system comparison. A cell is occupied at accuracy ≥ 0.95; D6 requires replacement-error ≤ 0.05 and retention ≥ 0.95. Online scores for vector kNN and RAG-triples were 0.000 in an earlier revision because those implementations lacked an incremental write path; with the write path added (a registered correction), both score 1.000—the column measures implementation affordance, not algorithmic impossibility. Jue's D6 value is before the F1 repair; see text.

System	Exact	Paraphrase	High-overlap	Online	Near-dup	Interference
exact table	1.000	0.000	1.000	1.000	1.000	clean
vector kNN	1.000	0.519	1.000	1.000	1.000	—
RAG-triples	1.000	0.962	1.000	1.000	1.000	—
learned router	1.000	0.421	1.000	0.000	1.000	—
A-relcond	1.000	1.000	1.000	1.000	1.000	clean
Jue (relational)	1.000	1.000	1.000	1.000	1.000	repl. err. $0.16 \rightarrow$ **0.000**

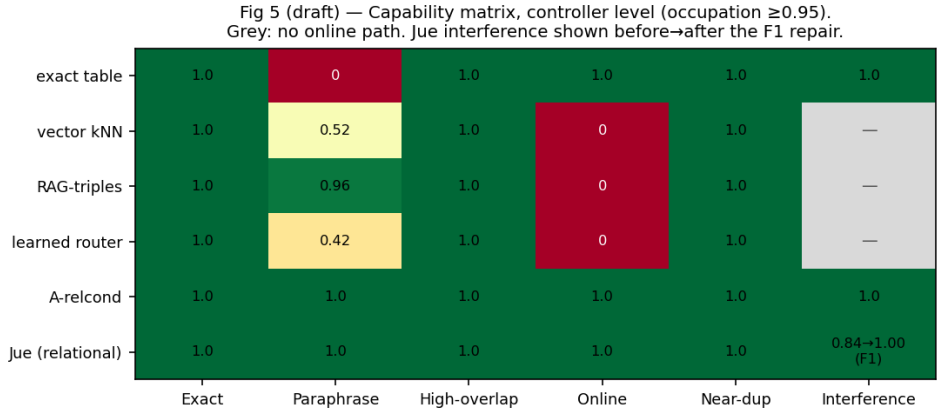

Figure 5: Capability matrix at controller level (occupation threshold 0.95). Grey cells: the system has no online-update path. Jue's interference cell shows the value before and after the conditioned-from-birth repair.

F Capability Matrix

F.1 The Baseline Capability Matrix

Act two is decided by a matrix, not an anecdote. Five baselines—an exact $(\text{entity}, \text{relation}) \rightarrow$ answer table, vector kNN over a triple store, RAG-with-triples, a small learned router, and **A-relcond**, the relation-conditioned geometric readout discovered by our own ablation family (§G.2)—are run against Jue's relational route on frozen XL materials (40 entities $\times$ 5 relations with paraphrase forms and distractors; 50+ pairs stratified by overlap), under the same sequence-forced endpoint and scoring. Six dimensions are measured: exact query, paraphrase, high-overlap disambiguation, a 50-step online stream with no index rebuilds or retraining, near-duplicate disambiguation on deliberately unscreened materials (overlap logged, not rejected), and serial interference over a 100-event stream (replacement errors, readout pollution, retention of early facts). These runs are controller-level probes (§5.1); the preregistered occupation rule—a claim of a unified substrate survives only if no baseline occupies every cell—was written before any system ran.

The full matrix is given in Table 4 (Appendix F).

The outcome (Table 4, Fig. 5) triggered the preregistered branch we least wanted: A-relcond occupies every cell. Relation-conditioned addressing is sufficient for every single-hop capability we could measure, and the paper reports this without qualification. The matrix also exposed the relational route's one weakness, and its dissection is the most instructive number in this section: all sixteen replacement errors occurred in a single mechanistic window—after an entity's second relation was taught but before its third event triggered promotion. Before promotion, the node had no

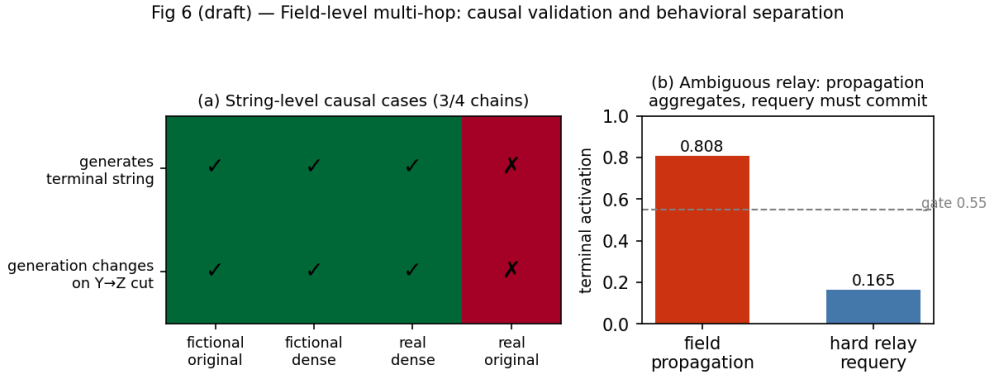

Figure 6: Field-level multi-hop. (a) String-level causal criterion per frozen chain: generate the terminal answer, and change it when the taught relay edge is cut. (b) On an ambiguous first hop, continuous propagation aggregates over both candidate relays; hard requery must commit to one.

relational identity, and first-taught capture (§G.2) recurred in miniature. The repair follows directly: relation-condition the readout from node creation, demoting promotion to routing, bookkeeping, and guardrails. On the identical frozen interference stream, the replacement-error rate falls from 0.16 to 0.000 with near-duplicate accuracy, retention, and the promotion reverse-gate all unchanged. The finding became the architecture: at controller level, the repaired route is a strict superset of the addressing baseline, and the question act three answers is what the superset's remainder—the field—buys.

G Supporting Surfaces in Full

G.1 Paired Associations and Paraphrase Generalization

On paired associations both routes reach 1.000 on taught queries and on held-out paraphrase forms, confirming that the coupling addresses by meaning rather than by surface string; never-taught probes are reported as diagnostics only, since an associative memory cannot be penalized for lacking what it was never taught. This surface is saturated for both primitives and therefore carries no discriminative weight; it exists to establish that the endpoint and coupling work before the divergent materials are introduced.

G.2 The Primitive Comparison: One Concept, Many Relations

The flagship experiment asks a question that paired associations cannot: what happens when one concept participates in several relations? Each R6 entity is taught three relations; queries name the relation type. Our preregistered prediction was that winner-take-all readout would answer exactly one relation per entity, yielding $1/3$ accuracy. The prediction landed exactly (Fig. 2): Route A scores 0.333 on train and paraphrase forms alike. The per-query record sharpens the mechanism beyond the prediction: A does not merely fail two relations—it *returns the same answer for all three relation prompts*, and the captured answer is in every case the *first-taught* relation's. Under a geometric primitive with winner-take-all readout, the first edge written on a concept permanently captures its readout; later knowledge about the same concept cannot reach the output. Route B, whose promoted entities re-rank candidates by relational context (Eq. 2), answers all twelve triples and all twelve paraphrase forms correctly. We read the failure as structural, not a tuning gap: it is invariant to teaching order and amplitude, follows from the readout's functional form, and was preregistered. What repairs it is the finding of this paper's second act: the readout must be *relation-conditioned*. Two designs achieve this—conditioning the query key (a geometric route we construct as our own strongest baseline in §I.2) or conditioning through relational identity (Jue's route)—and adjudicating between them is the matrix's job, not this experiment's.

G.3 Relations Take Over at Promotion

The routing schedule (Eq. 2) transfers addressing weight from the seed key to the relation row as evidence accumulates, with all four R6 entities following an identical trajectory and relations dominating at promotion (k=3; Fig. 3, Appendix C). This documents the designed schedule executing as specified—not emergent dynamics—and is included because the capability results depend on it having run.

G.4 Long-Horizon Stability, Proficiency, and Measured Capacity

A 1000-round mixed-stream run (multi-relation teaching, paired associations, unsupervised consolidation) exercises the promotion lifecycle at length (Fig. 4). Promoted-node count stays constant at nine across all seeds; no node demotes; no node oscillates (maximum promote/demote transitions $= 1$), so the hysteresis band does its job on first long-horizon contact. Early-taught facts are re-probed every 100 rounds and remain at 1.000 throughout—no long-horizon forgetting in this regime. Repeated tasks show a proficiency effect: relaxation converges in 9 steps late in the run versus 18 early, the architecture's "faster with practice" claim in its first measured form. Two boundary observations are reported with equal prominence. First, maximum node degree grows linearly (9 to 84) with no mechanism to shed edges—stability holds, but the missing edge-lifecycle component (§H.5) is visibly accumulating debt. Second, serial teaching on a shared field exhibits interference: later associations teach in 1–2 steps where isolated fields need 9–18, the third overshoots, and shared hub nodes climb from degree 10 to 24. We therefore replace the original design document's claim that context deadlock is "structurally impossible" with what the instruments actually show: capacity and interference are real, measurable, and bounded in this regime—a measured boundary being worth more than an asserted infinity.

G.5 Synthetic Field Dynamics

Independently of language coupling, the field itself supports multi-hop propagation: in the synthetic testbed, with all nodes visible and no staging, an $A{\to}B{\to}C$ chain activates C to 0.78 (semantic threshold 0.55) under the decoupled energy, and severing the $B{\to}C$ edge suppresses C to 0.34. The dynamics that indirect *language* retrieval requires exist in the field, and §I.4 shows that, once teaching contamination is removed, they carry chain queries to string-level causal validation; the residual boundary is analyzed in §5.

H Architecture Mechanisms in Full

H.1 Relational Node Primitive

A node begins *ephemeral*, carrying an input key v^{seed} (the centered hidden state of its first valid mention), a provenance record, and an empty relation row—its identity is where it sits. As distinct teaching events accumulate it is *promoted* to *relational* status: identity becomes the incident edge pattern and v^{seed} is demoted to a frozen addressing key. Promotion is evidence-gated and calibrated on throwaway materials disjoint from all capability sets (k=3 distinct teaching events plus an adaptive edge-mass threshold), reversible with a hysteresis band, and degree-guarded against collapsed-input hubs. No anonymous node may exist: a node lacking both a seed key and a readout anchor is invalid by construction.

H.2 Routing Schedule

Readout addressing interpolates between the seed key and the relation row as a node accumulates evidence:

$$\gamma_{\text{rel}} \;=\; \min\Big(1,\; \frac{n_{\text{events}}}{k}\Big), \qquad \gamma_{\text{in}} \;=\; 1-\gamma_{\text{rel}}, \tag{2}$$

so that relations take over at promotion (k=3). This is a designed schedule, not emergent dynamics.

H.3 Language Coupling

The input boundary is owned explicitly. For an association query, the input representation is the layer-L^* hidden state of the target token (final token for multi-token words), centered by a frozen mean vector estimated once on a 5000-word corpus disjoint from all task materials; shared-template mean pooling is prohibited, having been shown to collapse lexical content to near-collinearity (pairwise cosine ≈ 0.9997). L^* is selected by semantic class structure (within-category minus between-category cosine) rather than by isotropy alone. Every teaching run is preceded by an input-distinction gate: pairwise query cosine below 0.8, initial-node Jaccard at most 0.3, and at least three activated candidate nodes per query; a coverage gate requires each input word to activate at least three teachable node clusters.

H.4 Dual-Mode Learning

Learning is Hebbian and local throughout; the two modes differ in scale and license. *Weak learning* ($\Delta w_{ij} \propto \alpha_{\text{weak}} a_i a_j (1 - |w_{ij}|/w_{\max})$, $\alpha_{\text{weak}}{=}10^{-3}$) runs continuously and only consolidates; we show empirically that it cannot, by itself, build edges strong enough to propagate activation—the measured edge scale falls two orders of magnitude short of the propagation threshold (§B). *Deliberate learning* requires an external supervision signal (self-evaluated confidence is excluded as circular) and is executed as *teach-to-criterion*: one strong Hebbian step at a time ($\alpha_{\text{step}}{=}0.2$; an initial 0.4 calibration was re-registered to 0.2 under the constant-provenance protocol for the isolated-teaching regime), followed by a read-only relaxation probe of the target activation; teaching stops inside the band $[0.6, 0.9]$, flags overshoot above 0.95, and fails outside $[0.5, 0.95]$ at the step budget. Closed-loop control of the observable replaced open-loop rate calibration after the latter was shown to be uncontrollable under Hebbian positive feedback: at fixed rates, block-level pull varied by $25\times$ across associations and diverged super-linearly.

H.5 Interpretability, Guardrails, and Long-Horizon Mechanisms

Interpretability is a design filter, not a reporting feature: a candidate energy form was rejected because it corrupted causal edge-cut interventions, and every activation change carries a CausalTrace record naming its cause. Five guardrails bound the system—concept drift ($\cos > 0.95$ to seed), weight ceiling, degree cap, path decay, and a general-capability check—and a monopoly monitor watches single-edge dominance. Their measured blind spot (distributed rather than single-edge pollution) is reported in §5.2. Learning is staged: consolidation commits directly, deliberate teaching passes a staging buffer, mirroring a propose–verify–commit discipline rather than immediate write-through.

I Readout Metric Decomposition

I.1 What the String Metric Measures: A Three-Way Decomposition

The sequence-forced endpoint makes forced-realization accuracy track anchor selection, so we decompose readout into three preregistered metrics: *M1* anchor retrieval, *M2* field-state quality (terminal activation), *M3* free generation with forcing *disabled*. On six frozen items (base floor $0/6$; forced endpoint reproduces its labels $5/6$), free greedy decoding under first-position bias rises monotonically with amplitude ($0/6 \to 2/6$) as the answer-token logit deficit shrinks ($-8.33 \to -1.48$). At full amplitude all five anchor-correct items place the gold *first* token at argmax, but only single-token answers complete ($2/5$): every multi-token item fails at the unforced suffix. Applying the same bias at *every* position without token scheduling recovers nothing ($2/6$)—the model repeats the anchored token—so *scheduling, not bias coverage, is load-bearing* for multi-token answers. This also elevates a previously buried observation—a chain whose teaching missed its activation band still generated the correct string—from anomaly to design fact: the string metric certifies the anchor-and-schedule pathway, and claims about field learning quality must cite M2, not M3.

I.2 The Baseline Capability Matrix

Against five baselines—exact table, vector kNN, RAG triple store, learned router, A-relcond—the matrix is a *mechanism diagnostic*, not a system comparison. Two results matter, both against us. A-relcond, our own strongest addressing baseline, ties Jue on every cell: exact match, paraphrase, high-overlap, online, near-duplicate (all 1.000). And the online column originally scored 0.000 for kNN and RAG because those implementations had no incremental write path; with the write path added—a registered correction—both reach 1.000, so that column measured implementation affordance, not algorithmic impossibility. What remains distinctive at this layer is only interference behaviour: Jue's delayed-update replacement error is $0.16 \rightarrow 0.000$ after conditioning from birth. Full matrix in Appendix F.

I.3 The Matched-Interface Protocol

Comparing a continuous energy substrate against a symbolic graph memory is usually confounded: the two are wrapped in different controllers, judged at different endpoints, and given different write budgets, so any difference may belong to the wrapper rather than the substrate. We therefore specify a protocol under which the comparison isolates the substrate, and we state it separately because it is reusable beyond this paper. Six conditions must hold simultaneously. *(i) Identical write contract*: both systems receive the same controller-parsed entity, relation, answer, teaching target, and stopping condition. *(ii) Identical readout*: both consume their selected record through the same endpoint, judged by one string-comparison function. *(iii) Identical provenance interface*: both record which event wrote each edge, at the same granularity. *(iv) Identical intervention interface*: both expose edge cutting with the same semantics, so causal claims are testable on either. *(v) Matched budget*: propagation depth and candidate breadth are matched to the task, not to whichever system they favour—in our first attempt a round budget fixed for two-hop chains was carried over to four-hop materials and silently handicapped the graph, which the instrument gate caught and we repaired before adjudication. *(vi) Blind materials with both architectures frozen at registration*: evaluation materials come from an independent generator with no shared vocabulary, and neither system may be repaired after materials are seen.

The protocol makes ties informative. Without it, an equivalence result is uninterpretable, because any missing interface on the comparison arm inflates the measured advantage of the other; with it, a tie licenses the statement that the substrates are not distinguished *on the measured surfaces*, and a residual—such as the zero-relaxation ablation of §I.5—can be attributed to a specific mechanism rather than to the wrapper.

I.4 Act Three: Field-Level Multi-Hop and the Matched Graph Baseline

Chains are taught edge-by-edge under isolation-merge teaching and judged by a two-part criterion: generate the terminal answer, *and* change that string when the designated relay edge is cut.

Formation and causality. On 18 blind-registered chains (six each at 2, 3, and 4 hops), isolated-merge teaching forms $16/18$ chains and every formed chain is cut-sensitive ($6/6$, $5/6$, $5/6$ by hop count); without isolation only $3/18$ form. Isolation therefore protects *formation under sequential write interference*, not causality: conditional on formation, cut-edge causality is $22/22$ across arms, hop counts, and the four original frozen chains ($3/4$, the failure being the sparse real-word variant). Shortcut contamination is $0/36$.

Ambiguous relays. On a chain whose first hop admits two candidate relays, continuous propagation keeps both active and reaches terminal *activation* 0.808 where a hard single-relay requery reaches 0.165 (activations under gate 0.55, not accuracies).

The matched comparison. On 20 blind-registered constructions judged by a common string endpoint, soft top-k marginalization and candidate-subgraph message passing reach $20/20$ (hard single-relay requery $0/20$) where field propagation reaches $15/20$; our claim that propagation supplies what the lookup family structurally lacks is retired. We then tested whether the remainder—cuttable edges, string-level consequences, provenance—is field-specific. A provenance-aware typed graph memory (PGM), given the *identical* controller contract, sequence-forced readout, provenance records, edge-cut interface, and hop-matched propagation budget, evaluated on independently generated blind materials with both architectures frozen at registration, reproduces every measured prop-

erty: formation and cut-causality tie ($11/18$ each; shortcuts $0/18$ both), ambiguous-relay and single-hop endpoints tie within margins ($20/20$ vs. $19/20$; $15/20$ each, the same five items failing on both at a shared endpoint boundary), and the graph is *superior* on provenance completeness ($68/68$ vs. $52/68$) and retraction ($7/10$ vs. $2/10$). The preregistered conclusion is adopted verbatim: given the same contracts and intervention interfaces, a typed graph reproduces every measured property of the field on these materials; *the energy dynamics are not yet doing irreplaceable work*.

I.5 Discriminative Probes on the Field's Native Ground

That equivalence is measured on discrete, exact-cue tasks—the graph's home ground. We therefore preregistered three probes where a continuous substrate should differ, each with the graph (plus its strongest variant) as null hypothesis and the field required to exceed a margin. None produced a field advantage; two could not be measured.

Attractor completion (measured; tie). Under degraded cues—partial cluster activation and uniform damping, three levels each—field relaxation reached $114/117$ against $114/117$ for fuzzy retrieval over the typed graph (margin 12; clean-cue gate $40/40$ for all arms). The preregistered dynamics clause was not licensed: an ablation holding the cue fixed but running zero relaxation steps reached 111, so relaxation contributed a net 3 items, below the registered threshold of 6. Grounding, not dynamics, carries the completion. One asymmetry is worth recording: all three field failures were NO-BASIN (no answer committed), none WRONG-ATTRACTOR, where nearest-neighbour retrieval always returns its closest key.

Geometric generalization (not measured). A probe of field-level addressing on unseen surface forms triggered its instrument gate: the field-native arm scored $2/40$ on *canonical* queries, reproducing a previously archived negative result (hidden-state injection fails under the current protocol) rather than testing generalization. It did quantify a readout boundary: for one item the gold token's base-logit deficit was 18.05, exceeding the forcing amplitude 12.44, so anchor-correct items can fail realization—which accounts for the shared endpoint failures above.

Shortcut licensing (not measured). A probe of provenance-gated path compression failed its registered premise: the field grew no causal shortcuts to license ($0/12$). Diagnosis converges with §5.3: source and terminal clusters overlapped on 31.75 of 32 nodes with source–terminal activation difference exactly zero, leaving no distinct pair between which a shortcut could form. The same recruitment defect that closes the scale gate makes this probe unmeasurable. The graph arm, given the same licensing gate, separated the two shortcut classes completely ($24/24$).

Two of three probes were blocked by instrument boundaries the campaign itself identified, and the one that ran ended in a tie whose dynamics clause the preregistration withheld: on every surface we could measure, the energy dynamics are not load-bearing.

J Real-Task Qualitative Demonstration

Three real-task scenarios were registered before execution—prompt text, taught facts, expected answers, and interference passages frozen and hashed—and each was run exactly once under greedy decoding at the registered amplitude. All three are reported here in full, including every failure; they are an illustration, not a sample, and no rate or capability claim is drawn from them.

Scenarios. *(S1)* A single new fact is taught from context and queried after an interference passage. *(S2)* Two relations of the same entity are taught and queried separately, illustrating relation conditioning. *(S3)* A two-hop chain is taught under isolated merge and queried before and after cutting the relay edge.

Outcome. All three scenarios are failure cases at the string endpoint. The decomposition is consistent across them and matches §I.1: *anchor selection was correct on all four queries* (the intended record was retrieved in every case), and the only query that realized correctly was the one whose answer is a single token (S2, `Veyra`). The three multi-token answers fell back to base-model continuation of the interference passage. This is the forced-realization boundary reported in the body, reproduced on materials for which no parameter was tuned: retrieval reaches the right record, and the endpoint fails to spell it out under long-context interference.

The multi-hop case is worth stating precisely. In S3 the string output was wrong both before and after the relay edge was cut, so the string endpoint could not display any difference; but anchor selection changed from the correct record to *none*. The intervention therefore acted on retrieval as designed, while the readout layer masked it at the level where we usually measure. Read together with the $22/22$ conditional causality of §I.4—which is conditioned on chains that form and realize—this locates the gap precisely: the causal machinery works where the endpoint permits observation, and these scenarios are outside that permission.

Complete registered prompts, generated strings, per-token generation records, and full CausalTrace excerpts for all three scenarios are released with the artifacts.

K The Drive–Plasticity Confound in Full

K.1 An Open Problem: The Drive–Plasticity Confound

One boundary in this paper deserves more than a limitation entry, because two independent failures bracket it from opposite sides and reduce it to a well-posed control problem. We state it formally, report the empirical bracket, and lay out three candidate solution directions with falsification criteria. We deliberately do not endorse a solution: this project has already buried one elegant a-priori repair, and the record argues for stating the problem precisely rather than decorating it with an untested fix.

Setting. Let a query or teaching probe recruit an active candidate set A of field nodes, and let $a_i(A)$ denote the relaxed activation of a target node i. Informally, $a_i(A)$ mixes two contributions:

$$a_i(A) \;=\; \Phi\big(\; \underbrace{d_i(A)}_{\text{input drive}}\,, \qquad \underbrace{e_i(W)}_{\text{learned-edge contribution}} \;\big), \tag{3}$$

where $d_i(A)$ depends on the size and composition of A (which varies with field load: isolated, shared, or consolidated regimes), and $e_i(W)$ is the quantity that teaching is actually meant to control. The closed-loop teaching rule of §H.4 stops when $a_i(A)$ enters a frozen band. The confound is now visible: *the controller observes $a_i(A)$ but intends to regulate $e_i(W)$*. Whenever $d_i(A)$ shifts—because A doubled, or because a consolidated field recruits differently than an isolated one—the same band corresponds to different learned-edge strengths.

Problem statement. *Design a teaching control law whose controlled quantity is the learned-edge contribution $e_i(W)$, and whose stopping decision is invariant to the size and composition of the active set A.*

The empirical bracket. Two preregistered failures delimit the solution space from both ends (Table 5). Under a consolidated topology that doubled $|A|$ from 64 to 128, per-node drive was diluted and teaching stalled far below the band (under-drive). A scale compensation calibrated on isolated throwaway fields then over-drove the majority of associations in shared and full fields (over-drive), collapsing retention to 0.000 with readout pollution 0.990—while the monopoly guardrail stayed silent, because the damage was *distributed* across edges rather than concentrated in one. A static, open-loop compensation therefore fails on both sides of the operating range; the failure is structural, not a miscalibrated constant. We note the precedent inside this same project: open-loop teaching rates failed for the same reason and were replaced by closed-loop control of an observable—except that here the observable itself is what is contaminated.

Why this is not Jue-specific. We conjecture the confound applies to any external memory that combines (i) similarity-driven candidate recruitment, whose recruited-set size varies with memory load, and (ii) local plasticity regulated by an activation-level observable. Fast-weight and Hopfield-style test-time memories that adopt activation-referenced write rules would face the same mixed observable as their stores fill. We state this as a conjecture, not a result: our evidence is one architecture, bracketed twice.

Candidate directions and falsification criteria. *(a) Observable separation.* Measure the drive-only response by a zero-edge forward pass (W masked) and control the residual $\hat{e}_i = a_i(A) -$

Table 5: Two failures bracketing the drive–plasticity confound. Both were preregistered runs; neither was tuned post hoc.

Regime	Compensation	Teaching outcome	Downstream
Consolidated, $\lvert A\rvert$: 64→128	none	stall at 0.295 (band $[0.6, 0.9]$)	endpoint CI blocked
Shared/full fields	static normalization calibrated on isolated fields	overshoot on $94/100$, $191/200$, $41/48$ of associations	retention 0.000; pollution 0.990; monopoly guardrail blind (top-1 edge share < 0.0015)

$a_i^{W=0}(A)$. The energy decomposition instrumented in our diagnosis is a first step toward this. *Falsified if*: teaching controlled on $\hat{e}_i$ still exits the band across isolated, shared, and consolidated regimes on frozen materials. *(b) Closed-loop drive scaling.* Before teaching, servo the drive scale itself to a criterion—a fixed reference activation on a frozen neutral probe set—so that teaching always operates at a standard drive; the project's own teach-to-criterion philosophy applied one level down. *Falsified if*: no drive scale simultaneously satisfies the retrieval operating point and the teaching reference across regimes. *(c) Load-invariant reparameterization.* Replace additive input drive with a competitive (budget-conserving) form so that recruited-set growth redistributes rather than dilutes drive. *Falsified if*: budget conservation suppresses the multi-answer coexistence that divergent multi-relation queries require, or breaks the frozen synthetic-propagation regressions.

Any of these, validated under the discipline of §6, would close the last gap between this paper's controller-level repairs and a fully field-level online concept memory. Until then, the problem stands as stated—measured from both sides, owned by no solution.

A Artifact Availability

All specifications, execution reports, probe CSVs, frozen materials with generation seeds, decision registry, and stop-gate ledger are released at an anonymized repository: `https://anonymous.4open.science/r/jue-docx-55D2`. Every number in this paper can be traced to the report and commit that produced it; the repository README indexes claims to evidence files. Retired claims and instrument failures are included on the same footing as the results.

B Stop-Gate Audit Chain

Table 7 lists the stop-gates with rulings and commits (early audits predate version control; their rows carry stage tags instead of stable hashes). Three additional PX-era stops (promotion dry-run zero on paired topology; endpoint-CI baseline STOP; Step-0 calibration infeasibility) are documented in the released artifacts with the same schema.

Revision-Sprint Gates

C Claim Ledger: Whitepaper v0.3 Verdicts

D Decision and Constant Registries

The full decision registry (D-001–D-026, with supersession annotations) and the constant registry (values, derivation commits, regimes, dependencies, and freshness states) are released as `DECISIONS.md` and `CONSTANTS.md` in the project artifacts. The live-constant quick reference reproduced at the head of the decision registry is the authoritative index; STALE values are retained for the historical record and rejected at experiment entry points.

Table 6: Revision-sprint stops, rulings, and repairs (all preregistered).

Gate	Result (evidence)	Ruling / repair	Commit
RS-1 readout family	eight non-conditioned arms 0.333; A-relcond 1.000	flagship claim repositioned: conditioning suffices	91d66a19
RS-2 density probe	chain 0.000 after density fillers	density-only explanation rejected	91d66a19
RS-2 two-round wiring	chain 1.000; attribution audit: VERSION_B_REQUERY	downgraded: harness requery, not mechanism	91d66a19
RS-2C field propagation	1/4 strict; addendum: pre-cut generation already wrong (L L)	partial-pass branch invoked; breakpoints mapped	91d66a19
RS-4 matrix	A-relcond occupies all four cells	claim repositioned (preregistered branch)	ffe781a
RS-4B near-dup + interference	Jue replacement-error 0.16, all in pre-promotion window	serial-interference disadvantage branch	a4f632be
F1 conditioned-from-birth	$0.16 \to 0.000$; D5/retention/reverse-gate unchanged	PASS; promotion demoted to guardrails/routing	1875f435
F2 isolated multi-hop	$3/4$ chains string-level causal; ambiguity 0.808 vs 0.165	FULL_POSITIVE	02be3f8b
F3 field consolidation	teaching stall 0.295; then normalized variant: distributed overshoot, retention 0.000	two hard stops invoked; open problem stated (§K.1)	72b65a89
Revision campaign (post-review):			
JW-AUDIT-1 W symmetry	one asymmetry-leaking write path; row-form relaxation; E load-bearing	A-INC: implementation inconsistent with stated energy	f7a123d2
JW-FIX-1 symmetry repair	$\|W-W^\top\|=0$; gradient error 0; $4/5$ states clean, T2 rerun exact $(3/4)$	FIX-S; all published numbers retained	6a92f54a
ANOM-1 α provenance	0.2 registered (D-016); 0.4 was initial calibration	closed: registered protocol change, not tuning	f7301318
RS-7/7b natural generation	M3 monotone $0 \to 0.333$; per-position bias without scheduling: $2/6$ unchanged	N-GRAD + S-FORCE: scheduling load-bearing	289b9784
RS-8 r1	activation compared with accuracy	*voided*: spec unit error, recorded	—
RS-8 r2 multi-candidate	field $15/20$; soft top-k / subgraph MP $20/20$; hard requery $0/20$	G-LOSE: accuracy-advantage claim retired	db001b08
RS-9 online write path	kNN / RAG online $0.000 \to 1.000$	F-SCORE: column semantics corrected	a1f1d7a0
RS-12 isolation ablation	ISO $16/18$ vs NOISO $3/18$ formed; shortcuts $0/36$; hops $6/6, 5/6, 5/6$	D-NEED: isolation protects formation	c6d0dbd7
RS-10b conflict updates	controller LWW $10/10$; field anchor $1/10$; confidence 0.35 const	MIXED; two-layer divergence measured	feef1194
RS-10a r1–r3 entity scale	r1 prototype-cache collision; r2 token-ID mismatch $1507/1507$; r3 gate stop $1/7$	three instrument stops; FINAL invoked	15477ba7
JW-DIAG-2 collapse autopsy	scorer max-pool tie (spread 0.0, node 851 in all clusters); query term 0.064%	DIAG-LOC: harness-local cause	d68dcaf5
JW-FIX-2 scorer + morphology	production paths audited clean; regressions $10/10$ exact; Jaccard ≤ 0.2 frozen	FIX2-OK	1dc8b48b
RS-14 gated rescan	runtime cluster Jaccard 0.667 from frozen 0.2; gate $1/7$	FINAL-2: open limitation stands (§5.3)	5232a188
RS-15/15b matched graph baseline	PGM ties every surface; superior on provenance $68/68$ vs $52/68$, retraction $7/10$ vs $2/10$; B2 budget + runner route were instrument-repaired, shared endpoint boundary classified capability	P-LOSE: energy dynamics not yet irreplaceable; adopted verbatim	d7fff80e
Discriminative campaign:			
V1-GEO geometric generalization	field-native arm $2/40$ on canonical queries (archived negative reproduced); logit deficit $18.05 > 12.44$ amplitude	GEO-ODD: not measurable; readout boundary quantified	5e412408
V1-ATTR attractor completion	$114/117$ vs $114/117$; zero-relaxation ablation 111 (net $3 < 6$)	ATTR-TIE; dynamics clause withheld	0c525bf2
V1-LICENSE shortcut licensing	Jue causal shortcut growth $0/12$; source/terminal overlap $31.75/32$; PGM separation $24/24$	LIC-ODD: premise failed; converges with recruitment defect	feda5592
Scale campaign (recruitment):			
JW-DIAG-3 entanglement autopsy	membership = unconditional cosine top-k before W write; H2/H3/H4 excluded; k tightening ruled out by counterfactual	DIAG3-LOC	7a5e96dc
JW-FIX-3 exclusive recruitment	runtime Jaccard $0.667 \to 0.000$; regressions exact	FIX3-OK	e0be550e
RS-14 r2 gated rescan	Jaccard 0.000 but anchor $1/7$; $N=200$ late-stream band hit 0.225	FINAL-R2: repair insufficient; second cause established (§5.3)	fcc58a7a

E MATERIALS

F REVISION-SPRINT ABLATIONS AND MATRICES IN FULL

Readout family (frozen four-entity testbed). All arms share materials, teaching, endpoint, and amplitude; only the readout varies.

Table 7: Stop-gate chain: reason, ruling, commit.

Stage	Stop reason (evidence)	Fix / ruling	Commit
P1.9-AUDIT	P1.9 all fail/tie/dead contradicted P1.5; E3 diagnostic showed real and fictional max-cos both about 0.135, hidden pair cos about 0.99999	Harness bug ruled in for E3 interpretation; architecture discussion narrowed to language-field alignment; no capability rerun authorized	`p19-audi`
P1.9-DIAG	Layer-20 shared-template segment pooling collapsed content; pooled median cosine 0.999697/0.999824	H1 pooling collapse primary and H2 anisotropy amplifier; input extraction moved toward target-token centered hidden with centering	`50d90a5d`
P1.95-R1	R1 first-hop stopped: target Y round-1 activation stayed at floor 0.010000 while gate required ¿=0.5	Attributed to weak edge scale; escalated to P1.96 propagation-scale/strong-learning audit	`b0893058`
P1.96-CAP	Strong-edge route still structurally capped: frozen pull target 2.463084 but current max 1.190633 and theoretical max 1.532916	Identified multiplicative W*cos ceiling; moved to P1.97 energy interaction decoupling	`dcfd090e`
P1.97-READOUT	Route B provenance readout sanity failed 0/20 top-5 hits	Replaced geometric readout mapping with provenance/token endpoint audits; output path treated as separate boundary	`0abc29e1`
P1.98-ENDPOINT-AUDIT	Endpoint audit found B-static impossible failure: bare token biased while continuation gold differed; 20/20 token convention mismatches; amplitude flipped 0/5	D-017 continuation-token convention and D-018 final-string CI introduced; matrix v2 suspended	`90da9c5b`
P1.98-OUTPUT	After token fix, old output amplitude was stale; P90 logit gap 10.443748 so AMPLITUDE_V2 set to 12.443748; teaching CI still exposed route defects	Registered AMPLITUDE_V2 once before endpoint accuracy measurement; no iterative tuning	`0a2f5a45`
PX-0	PX-0 endpoint baseline intentionally STOP: single-token real passed but multi-token fictional stopped before route repairs	D-018 kept as hard gate; PX-0 allowed design work only, not capability experiments	`px0-clos`
PX-3B-PREFLIGHT	Promotion criterion did not activate on PX2B: 5 train pairs touched 10 nodes with max distinct events 1; promoted nodes 0; adaptive threshold inactive	Added R6 divergent multi-relation topology and R6-density calibration; k/N refreshed to 3/4	`91d66a19`

Table 8: Every claim of the original design document and its verdict.

v0.3 claim	Verdict	Evidence
Continuous reasoning = energy minimization (§4 of v0.3)	Holds in-field; language-level multi-hop open	synthetic $C{=}0.78$/cut 0.34; R4 chain $0/0$
Online learning = local gradient-free Hebb (§5)	Holds, rewritten: weak = consolidation only; deliberate mode requires external signal + closed loop	edge 0.04 vs. 2.46; teach-to-criterion
Dual-mode learning (§5.1/5.2)	Confirmed by failure of the single mode	P1.95/96 R1 record
Interpretability first (§6)	Promoted: active design filter	$\lambda{=}0.3$ rejected by edge-cut pollution
Unified memory; deadlock "structurally impossible" (§7)	Rewritten: capacity and interference measured; stability and retention hold	serial 18/1/2; hubs 10→24; 1000-round retention 1.000
Interaction energy $W \cdot \cos$ (§4.2)	Overturned and replaced: $\lambda{=}0$ decoupling	ceiling $1.53 < 2.46$
Node = sample point of semantic space (§4.1)	Replaced: relational primitive wins	R6 0.333 vs. 1.000; first-taught capture
Residual α interface / progressive takeover (§8)	Dead under tested protocol; E4-v3 deferred to 9B	four-arm α zero-lock record
Benchmark projections (§11)	Formally withdrawn	—
(added in revision) "Relations, not coordinates" (this paper's own v1 flagship)	Retired: conditioning suffices for disambiguation; identity lifecycle is guardrails, not the disambiguation source	RS-1; RS-4B; F1
(added in revision) Language multi-hop unreachable	Revised: $3/4$ chains pass string-level causal ablation after shortcut-isolated teaching	F2

Arm	Overall	Train	Held-out form
A-wta	0.333	0.333	0.333
A-top3 / A-top5	0.333	0.333	0.333
A-soft $\tau \in \{0.1, 0.5, 1.0, 2.0\}$	0.333	0.333	0.333
A-relcond	**1.000**	1.000	1.000

Table 9: Paired-association materials (PX2B). Ability rows use train and held-out-form; held-out-content is diagnostic only.

ID	Suite	Split	Type	Source	Answer	Role	Jaccard band
PX2B-01	R2_R3	train	fictional	Kelvar	Nemor	ability	¡0.15
PX2B-02	R2_R3	train	fictional	Tarsen	Caldis	ability	¿0.3
PX2B-03	R2_R3	train	fictional	Veyra	Orvane	ability	¡0.15
PX2B-04	R2_R3	train	real	cat	red	ability	¿0.3
PX2B-05	R2_R3	train	real	dog	blue	ability	¿0.3
PX2B-06	R2_R3	heldout_form	fictional	Kelvar	Nemor	ability	¡0.15
PX2B-07	R2_R3	heldout_form	fictional	Tarsen	Caldis	ability	¿0.3
PX2B-08	R2_R3	heldout_form	fictional	Veyra	Orvane	ability	¡0.15
PX2B-09	R2_R3	heldout_form	real	cat	red	ability	¿0.3
PX2B-10	R2_R3	heldout_form	real	dog	blue	ability	¿0.3
PX2B-11	R2_R3	heldout_content	fictional	Morin	Selith	diagnostic	¡0.15
PX2B-12	R2_R3	heldout_content	fictional	Peldor	Rynas	diagnostic	¡0.15
PX2B-13	R2_R3	heldout_content	real	river	green	diagnostic	¿0.3
PX2B-14	R2_R3	heldout_content	real	memory	yellow	diagnostic	¿0.3
PX2B-15	R4	direct	fictional	Kelvar	Nemor	ability	¡0.15
PX2B-16	R4	direct	fictional	Veyra	Orvane	ability	¡0.15
PX2B-17	R4	direct	real	cat	red	ability	¿0.3
PX2B-18	R4	chain_edge	fictional	Zerith	Lomar	ability	0.15-0.3
PX2B-19	R4	chain_edge	fictional	Lomar	Velora	ability	¿0.3
PX2B-20	R4	chain_query	fictional	Zerith	Velora	ability	0.15-0.3
PX2B-21	R4	chain_edge	real	horse	justice	ability	¿0.3
PX2B-22	R4	chain_edge	real	justice	white	ability	¿0.3
PX2B-23	R4	chain_query	real	horse	white	ability	¿0.3

Table 10: Divergent multi-relation materials (R6): four fictional entities $\times$ three relations, train and paraphrase forms, all D-015 gates passed.

ID	Split	Entity	Relation	Answer	Jaccard band	D-015
R6-01	train	Alvorn	capital	Corvax	0.15-0.3	True
R6-13	heldout_form	Alvorn	capital	Corvax	0.15-0.3	True
R6-02	train	Alvorn	currency	Praxen	0.15-0.3	True
R6-14	heldout_form	Alvorn	currency	Praxen	0.15-0.3	True
R6-03	train	Alvorn	neighbor	Quenlor	0.15-0.3	True
R6-15	heldout_form	Alvorn	neighbor	Quenlor	0.15-0.3	True
R6-04	train	Halven	capital	Brevik	¡0.15	True
R6-16	heldout_form	Halven	capital	Brevik	¡0.15	True
R6-05	train	Halven	currency	Cendril	0.15-0.3	True
R6-17	heldout_form	Halven	currency	Cendril	0.15-0.3	True
R6-06	train	Halven	neighbor	Mivara	0.15-0.3	True
R6-18	heldout_form	Halven	neighbor	Mivara	0.15-0.3	True
R6-07	train	Qadira	capital	Korath	0.15-0.3	True
R6-19	heldout_form	Qadira	capital	Korath	0.15-0.3	True
R6-08	train	Qadira	currency	Seldin	0.15-0.3	True
R6-20	heldout_form	Qadira	currency	Seldin	0.15-0.3	True
R6-09	train	Qadira	neighbor	Atheron	0.15-0.3	True
R6-21	heldout_form	Qadira	neighbor	Atheron	0.15-0.3	True
R6-10	train	Ostrel	capital	Norvik	0.15-0.3	True
R6-22	heldout_form	Ostrel	capital	Norvik	0.15-0.3	True
R6-11	train	Ostrel	currency	Tavora	0.15-0.3	True
R6-23	heldout_form	Ostrel	currency	Tavora	0.15-0.3	True
R6-12	train	Ostrel	neighbor	Wexlor	0.15-0.3	True
R6-24	heldout_form	Ostrel	neighbor	Wexlor	0.15-0.3	True

Interference dissection (RS-4B). Jue's sixteen replacement errors all occurred after an entity's second relation event and before its third (promotion at k=3); retention of facts 1–20 was 1.000 at steps 50 and 100 for both A-relcond and Jue; diagnostic top-5 pollution was 0.320 for both. After the F1 conditioned-from-birth repair, on the identical frozen stream: replacement-error 0.000, immediate accuracy 1.000, D5 strata 1.000/1.000, retention 1.000/1.000, promotion reverse-gate 1.000/1.000.

Multi-hop chains (F2). Isolated-merge teaching (D-031) with zero X-to-Z-exclusive shortcut weight asserted before scoring.

Chain	Terminal string generated	Changes on $Y \to Z$ cut	Pass
fictional original	yes	yes	✓
fictional dense	yes	yes	✓
real dense	yes	yes	✓
real original (sparse)	no	—	×

Ambiguous-relay diagnostic: field propagation terminal activation 0.808; hard relay requery 0.165 (activation gate 0.55).

G A6 Failure Taxonomy: Six Cases

Table 11: Six representation-boundary failures, one family: a boundary crossed on an implicit assumption of alignment.

Case	Key evidence	Commit	Source
A6-1 representation-space mismatch and E3 collapse	fictional mean max-cos 0.1353397; real mean max-cos 0.1353402; X/water hidden cos 0.99999297; Y/city 0.99998963; Z/music 0.99999440	p19-audit	experiments/results/p19_audit/p19_audit_report.md; audit1_e3_word_cos_dump.csv
A6-2 pooling collapse and anisotropy	layer-20 pooled median cos 0.999697/0.999824; target-token median cos 0.552327; pooled top-3 energy share 99.776%; centering median -0.067990	50d90a5d93263d0b5b	experiments/results/p19_diag/p19_diag_report.md
A6-3 weak propagation scale / edge strength	R1 Y round-1 activation 0.010000; frozen pull target 2.463084; current max pull 1.190633; theoretical max pull 1.532916	b089305810b1771476 / dcfd090e352a809ddc	experiments/results/p195/p195_r1_first_hop.md; experiments/results/p196/p196_r1_cap_diagnosis.md
A6-4 geometry-knowledge multiplicative coupling ceiling	P1.96 target 2.463084 unreachable under W*cos; P1.97 selected lambda=0; synthetic G2 lambda=0 jump C=0.780539 and cut C=0.335162 under semantic threshold 0.55	91d66a1957be2552c8	experiments/results/p197/p197_synthetic_regression.md; experiments/results/p196/p196_r1_cap_diagnosis.md
A6-5 readout-space mismatch	P1.97 provenance/readout sanity top-5 hits 0/20 against ¿=18/20 gate	0abc29e1028102e78c	experiments/results/p197/p197_readout_sanity_initial_zero_scale.m
A6-6 output endpoint contract failure	B-static biased bare token 45 while continuation gold began 62168; stored/bare mismatched continuation gold for 20/20 words; static amplitude 0.542411 flipped 0/5 generation-gold targets; D-018 initially STOP	90da9c5b282818eb3a	experiments/results/p198/p198_endpoint_audit/p198_endpoint_audit_repo

H The Design Philosophy in Full

Jue began as a philosophical argument, written in April 2026 before any experiment existed. We reproduce that argument here as it was made, then report what three months of adversarial validation did to each of its steps. We include this section not as preamble but as data: a design philosophy is itself a set of claims, and ours was tested.

H.1 The A Priori Chain

The original design document reasoned in ten steps. (1) *Algorithms over compute*: scaling has diminishing returns; the bottleneck is how efficiently a system learns. (2) *Learning over patches*: retrieval, agents, and memory systems work around a frozen model rather than teach it. (3) *A real-time model*: brains do not separate training from inference; every experience is both. (4) *Interpretability as a prerequisite*: to learn online, a system must know what it used in order to know what to update. (5) *The brittleness trap*: classical interpretable systems equated transparency with discrete rules and inherited their fragility—a fault of implementation, not of the concept. (6) *Continuous inside, discrete at the interface*: cognition is continuous; only communication is discrete. Language is a protocol, not the medium of thought—a position we took from Wittgenstein, and which Meta's Coconut (Hao et al., 2024) later reached independently from an engineering direction. (7) *An energy field* satisfies the resulting requirements: continuous, locally updatable, naturally inspectable, with reasoning as relaxation. (8) *Backward compatibility*: the field must be a plugin; disabled, the base model is untouched. (9) *Unified memory*: context and memory should be one system. (10) *Dual-mode learning*: everyday consolidation and deliberate study are different processes and need different mechanisms.

Table 12: The a priori chain (April 2026) and its post-validation fates (June 2026). Evidence pointers reference sections of this paper.

#	A priori claim	Verdict	Evidence
1	Algorithms over compute	practiced: decisive test needed 4 entities	§G.2
2	Learning over patches	sharpened: injection died; learned edges live	§5
3	Real-time model	partial: online loop yes; sleep consolidation absent	§H.5
4	Interpretability first	**promoted to design filter**	§H.5
5	Brittleness trap	**inverted confirmation**: interior robust; all fragility at discrete boundaries	§6
6	Continuous inside, discrete interface	**converged upon** (§H.3)	throughout
7	Energy field	retained, restructured: geometry expelled from knowledge	§3.1
8	Backward compatible	intact: field off = base model	§4.1
9	Unified memory	rewritten: asserted infinity → measured boundary	§G.4
10	Dual-mode learning	**confirmed by failure** of the single mode	§B

H.2 Fates

Table 12 gives each step's verdict after validation. Three deserve comment. Step 10 was confirmed *by failure*: running the system on consolidation alone produced edges two orders of magnitude too weak to propagate (§B)—the insufficiency of the everyday mode is the empirical argument for the deliberate mode. Step 4 was not merely confirmed but *promoted*: interpretability began as a stated prerequisite and ended as an active design filter, rejecting a candidate energy form because it corrupted causal edge-cut interventions (§H.5). Step 9 was honestly *rewritten*: the original document claimed context deadlock was "structurally impossible"; the instruments instead delivered measured capacity and interference (§G.4), and we consider the measured boundary the stronger result.

H.3 The Wittgenstein Closure

The most striking fate is step 6's. The final architecture converged, under falsification pressure, to exactly the shape that step prescribes: knowledge lives in continuous, learned relations (W); every contact with language is governed by an explicit, discrete contract—seed keys and distinction gates at the input, provenance anchors and a sequence-forced protocol at the output. The sequence-forced endpoint is the position made literal: output is a protocol to be driven, not thought to be expressed. The negative evidence is sharper still. All six representation-boundary failures catalogued in §6 occurred where a boundary was crossed on an implicit assumption—pooled inputs, borrowed embedding spaces, mismatched token conventions. The system failed every time it violated the philosophy and was driven, failure by failure, back to the position the philosophy specified. We state the epistemic status precisely: this is an observation of convergence, not a proof of the philosophy.

H.4 Self-Limitation

The philosophy and the repairs share an author, so the convergence of §H.3 cannot be fully separated from a survivorship narrative: repair directions were inevitably guided by the same priors being evaluated. What we can claim is narrower and checkable: each boundary violation was struck down by an experiment before it was explained by the philosophy—the evidence preceded the interpretation in every case in Appendix B—which distinguishes this record from pure confirmation bias without elevating it to a demonstration.