Long Context Is Not Memory

A falsifiable proposal for intelligence that compounds. Neither a longer window nor a nearest-neighbour index decides what an experience meant, or whether it should return.

Ockams, Inc. — WHY Research Reference application: **why.com**

Version 1.0 — August 2026

MEASURED
— observed in the deployed why.com system; source and method stated inline.

MODELED
— derived analytically or by simulation under stated assumptions. Not an observation.

PROPOSED
— design not yet built or evaluated.

Every quantitative claim in this report carries one of these markers. Nothing modeled is reported as measured.

§0 Abstract

A foundation-model call does not, by itself, provide durable governed memory: it maps a context to a distribution over continuations and carries nothing forward between calls. Deployed systems bolt on state — caches, fine-tuned weights, retrieval — but none of these decides *what an experience meant and whether it should return.* The prevailing remedies, longer context windows and embedding retrieval, extend how much a model can *see* without addressing what it should *keep*. We argue that scaling either one *alone* does not produce intelligence that accumulates, and we are precise below about which parts of that argument are proven and which are conjecture.

We show that context stuffing has per-query cost linear in lifetime history and is therefore not lifelong-viable; that embedding similarity *need not preserve memory-value ordering*, so nearest-neighbour ranking is not value ranking; and that under an extreme-value model, the embedding separation required to hold recall constant grows as `v(2 ln N)`, so — under that model's assumptions — retrieval degrades as a corpus accumulates.

We then propose Engram: a persistent typed graph in which the *significance* of a node is defined as its expected marginal predictive value over a person's forward trajectory, estimated from four topological observables, updated by consequence-driven credit assignment, and retrieved by constrained selection against a submodular surrogate rather

than top-*k* ranking. We are explicit throughout about epistemic status: the negative results on existing systems are argued analytically; the significance mechanism is a *research hypothesis*, and its learning loop is specified but not yet deployed. We describe the deployed reference application, separate what has been measured from what has only been modeled, and give the experiment that would falsify the central claim.

§1 Introduction

A model with a trillion parameters and no memory begins every interaction at the same place its smallest predecessor did: with nothing behind it. Scaling improves the quality of a moment. It does not make moments accumulate.

This distinction is easy to lose because the symptoms of missing memory look like symptoms of insufficient capability. A system that forgets a stated constraint looks careless. A system that re-derives a conclusion it reached last week looks slow. A system that cannot tell which of two contradictory user statements is current looks confused. None of these is a reasoning failure. Each is the absence of a substrate that decides what an experience meant, whether later events changed that meaning, and when the meaning should return.

Two remedies dominate practice. The first is to enlarge the context window until the relevant past fits inside it. The second is to embed the past into a vector space and retrieve nearest neighbours at query time. §2 argues that the first is an economic dead end and the second optimizes the wrong objective. §3 gives the formal alternative. §4 describes the running system. §5 separates what we have observed from what we have only modeled.

THE CENTRAL CLAIM

An experience is significant to a person in proportion to how much it changes what that person does next — not in proportion to how often it appears, how recently it occurred, or how closely it resembles the present query.

Frequency, recency and similarity are three measures of *attention*. None is a measure of *significance*, and the gap between them is where memory systems currently fail.

§2 The Insufficiency of Stateless Inference

2.1 Context windows are a band-aid, not a substrate

Three independent arguments, in increasing order of severity.

(A) THE ECONOMIC ARGUMENT

Self-attention over a sequence of length `n` costs `O(n²d)` in compute and `O(n·d·L)` in key-value cache. Sub-quadratic approximations reduce the exponent; none removes the

dependence on `n`. Let `T(t)` be the total history a user has generated by time `t`. Under context stuffing, per-query cost is `Ω(T(t))` even granting perfect prefix caching.

```
Define: a memory system is lifelong-viable iff
per-query cost is independent of lifetime history —

    Cost(q, t)  =  O(b)        b = retrieved budget

Context stuffing: Cost(q, t) = Ω(T(t)),  T growing in t.
∴ context stuffing is not lifelong-viable.
```

The bound is not softened by cheaper tokens. Any monotonically growing per-query cost eventually dominates any fixed budget; falling unit prices change the constant, not the exponent.

(B) THE DILUTION ARGUMENT

Suppose `k` tokens of an `n`-token context are relevant. The signal fraction `k/n` tends to zero as the window grows with `k` fixed — but a shrinking fraction does not, by itself, imply worse retrieval; that is an empirical matter, not an identity. The empirical evidence is what carries the point: positional degradation in long contexts is documented [4] — accuracy varies with position and adding distractors reduces it. So enlarging the window trades a hard failure (the fact is absent) for a soft one (the fact is present and not attended to), which is harder to detect and to debug. This is a robustness argument, not a proof that attention quality falls with `k/n`.

(C) THE CATEGORICAL ARGUMENT

This is the one that matters. Decompose memory into four operations:

```
Memory  =  ⟨ Selection, Consolidation, Revision, Retrieval ⟩

Context window provides:
    Selection      —  identity (everything is kept)
    Consolidation  —  none (raw transcript)
    Revision       —  none (contradictions coexist)
    Retrieval      —  attention over the undifferentiated whole
```

A context window is the degenerate case in which selection is the identity function. Re-reading one's complete diary before every decision is not an implementation of memory; it is the absence of memory compensated for by bandwidth. Human memory is valuable precisely because it is *lossy in a structured way* — it discards, compresses, and revises, and the policy governing what survives is where the intelligence lives [5].

2.2 Similarity need not preserve value order

Naive retrieval-augmented generation scores a memory `e` against a query `q` by embedding similarity, typically `cos(φ(q), φ(e))`, and returns the top `k`. We take that objective seriously and show it is misspecified — with the caveat, developed at the end of this section, that the mature RAG systems worth comparing against are no longer naive top-*k*.

What we actually want is the memory whose inclusion most changes what the model will conclude. That quantity has a natural information-theoretic form:

```
V(e | q)  =  D_KL( P(y | q, e)  ‖  P(y | q) )
```

The value of a memory is the divergence it induces in the output distribution. A memory that does not change the answer has zero value regardless of how relevant it looks.

OBSERVATION 1 — SIMILARITY NEED NOT PRESERVE VALUE ORDER

Embedding similarity $\cos(\varphi(q),\varphi(e))$ does not determine $V(e|q)$: two memories at equal similarity to a query can carry very different value. Therefore ranking by cosine need not rank by value, and top-k cosine is not, in general, the value-optimal selection.

An illustrative pair, not a proof. Consider two memories at comparable, high similarity to a query. One is a near-duplicate the query already implies — "how is my company doing?" beside a memory that merely restates the question — and adds little: V small. The other, "how is my company doing?" beside "the company filed for bankruptcy," sits equally close in embedding space yet carries decisive information the query lacks: V large. If $\cos$ determined value, equal similarity would force equal value; it does not. So similarity ranking need not preserve value ordering.

What this does and does not establish. This is a qualitative argument, not a theorem: we do not exhibit an explicit distribution and embedding geometry with $\cos(q,e_1) > \cos(q,e_2)$ yet $V(e_1|q) < V(e_2|q)$, which a formal claim would require, so no ∎ is asserted. The claim it supports is only the weak one the rest of the paper uses — cosine need not preserve memory-value ordering — not that V vanishes at maximal similarity nor that it peaks in the interior. The true shape of V against similarity is empirical (§5.2).

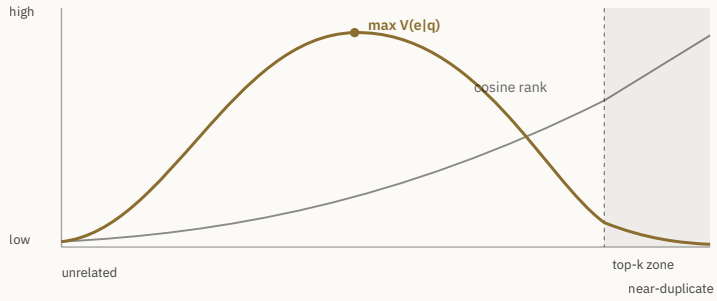

Figure 1. One conjectured shape for information gain against embedding similarity: cosine ranking rises monotonically while value need not, so top-k can draw from a high-similarity, low-information tail.

MODELED

The curve is a hypothesis illustrating Observation 1's non-monotonicity, not a theorem and not fitted to data; the true shape, and whether informative memories concentrate at intermediate distance, is the empirical question of §5.2.

The practical tendency practitioners report is that naive RAG over-returns the obvious — the passage that restates the query. Observation 1 explains why that failure is *possible* and why it is not corrected by better embeddings alone; how often it dominates in a given corpus is empirical, not settled here.

FOUR PROPERTIES A BARE METRIC CANNOT SUPPLY

The following are limits of a *pure symmetric similarity index*. Each can be, and in mature systems is, addressed by bolting on additional machinery; the point is that the bare metric supplies none of them, so each is a component a memory substrate must add rather than inherit.

Frequency bias. If an entity `v` occurs m_v times with clustered embeddings, then `P(retrieve v) ≈ 1 − (1−p)^m_v`, increasing monotonically in m_v. Significance carries no such dependence. Similarity retrieval is thus a biased estimator of significance, with bias growing in occurrence count. A person may raise their work five thousand times and a difficult family relationship twice; counting concludes the work dominates, and it is wrong in the direction that matters most.

Symmetry cannot encode direction. Cosine is symmetric: `cos(a,b) = cos(b,a)`. Causal and dependency relations are directed — and, in real systems, not necessarily transitive or acyclic (feedback loops exist; a direct causal edge does not compose into another direct edge). What matters here is only the direction: a single undirected score per pair cannot distinguish "A caused B" from "B caused A". A raw vector index therefore cannot represent directionality at all; typed directed edges are the minimal structure that can.

No supersession. When a user's state changes — "I moved to Berlin" superseding "I live in London" — the two statements are *topically almost identical* and receive nearly equal similarity scores. Contradiction is a form of proximity. A bare similarity index returns both and has no operator by which the later one wins; correctness needs a revision rule keyed on

temporal validity, which the metric does not carry. Production RAG stacks add this through metadata and versioning — which is precisely the point: it is added, not native.

Naive chunking severs relations. Fixed-window chunking splits an antecedent from its consequent and retrieves them independently, destroying the relation at ingestion. Semantic chunking, parent–child retrieval, overlap windows and graph indexing can preserve or reconstruct it — again, by adding structure the bare metric lacks.

THE HONEST COMPETITOR IS NOT TOP-K COSINE

Observation 1 and the four properties target the *bare metric*. The retrieval systems actually worth beating already move past it: maximal-marginal-relevance and other diversity selectors attack redundancy directly; cross-encoder rerankers rescore beyond cosine; hybrid lexical–dense retrieval, metadata and temporal filtering, query expansion, learned relevance and knowledge-graph RAG each add one of the missing pieces. Engram's claim is therefore *not* that these fail where cosine fails. It is narrower and testable: they rank on variants of *relevance* and add structure per-application, whereas Engram ranks on *significance and consequence* and treats direction, supersession and revision as first-class primitives of one substrate. Whether that difference predicts behaviour better is the empirical question of §5.2, not a claim settled by argument.

2.3 Retrieval saturation: a model of degradation with scale

Under a stylized model, similarity retrieval gets worse as the corpus accumulates — backwards for a system meant to support a lifetime. We give the model, then state plainly what it does and does not assume.

Model a corpus of `N` memories, one relevant, `N−1` distractors, with unbounded Gaussian scores of common variance σ^2 and means separated by $\Delta = \mu_r - \mu_d$. By extreme value theory the largest distractor score concentrates near $\mu_d + \sigma\sqrt{2 \ln N}$. The relevant item is retrieved at rank 1 only if it exceeds that maximum:

```
P(success)  ≈  Φ( Δ/σ  −  √(2 ln N) )

To hold P(success) = 0.90 as N grows:

    Δ/σ   ≥   √(2 ln N)  +  1.28
```

CORPUS SIZE N	√(2 LN N)	REQUIRED Δ/Σ FOR 90% RECALL
10^3	3.72	5.00
10^4	4.29	5.57
10^6	5.26	6.54
10^9	6.44	**7.72**
10^{12}	7.43	8.72

MODELED

Under this idealization, embedding separation would have to improve by roughly 55% between 10^3 and 10^9 memories simply to hold recall constant — while encoder quality is fixed and does not improve with corpus size. Within the model, a similarity-indexed store becomes less reliable the longer it is used.

What the model assumes, and where it can fail. The derivation assumes one relevant item, i.i.d. distractors, shared variance, and effectively unbounded scores. Real cosine scores are bounded, and real corpora are highly structured. Those structures do not merely shift constants — hierarchical retrieval, deduplication, learned embeddings and candidate pre-filtering can blunt or, in favourable regimes, defeat the growth. So this is *not* an assumption-free impossibility result. It is a directional pressure: absent countermeasures, more distractors crowd the threshold as `N` grows. The countermeasures a mature system already uses are, notably, ways of *not* ranking similarity over the undifferentiated whole — which is the same move significance makes.

THE CONSEQUENCE

The lesson is a design constraint, not a theorem: do not rank by similarity over an undifferentiated store. Reduce what is *eligible* for retrieval to a bounded working set. A *constant fraction* of history does not achieve this — an eligible set of size `ρN` still grows with a life, and the saturation pressure returns, merely delayed. What bounds it is an absolute eligibility budget (or a fraction that shrinks with `N`): a fixed cap on the significant, retrievable working set, with the rest retained but ineligible by default. Significance is the criterion proposed for choosing what occupies that fixed budget.

2.4 Training is population memory, not personal memory

A third remedy is raised often enough to answer directly: *memory is just data — the incumbents dissolve amnesia by continually training on proprietary corpora, so no separate substrate is needed.* Continual training is real and it is not what this paper proposes to replace. But parametric memory and personal memory are different objects, and four properties separate them.

Granularity. Gradient updates write regularities that hold across a corpus. What a given person decided last Tuesday is, at population scale, noise to be averaged away — the training objective is built to discard exactly the signal a personal memory exists to keep.

Latency. Personal memory must incorporate a fact the moment it occurs. Weight updates arrive on a training cadence, and no one retrains a frontier model per user per turn. The economics forbid it, and — as the serving literature makes plain — per-user weight state is precisely what breaks the shared-weight batching that makes inference affordable. *The same property that makes stateless serving efficient is what makes parametric personal memory impractical.*

Revisability. Supersession must be cheap. "I moved to Berlin" has to retire "I live in London" on contact. In a graph that is an edge operation (§3.5). In weights it is machine unlearning,

an open research problem, and there is no reliable way to remove one fact without disturbing others.

Governance. Provenance, retention policy, deletion on request and audit are enforceable over records. They are not enforceable over a weight: one cannot point to where a fact lives, prove it was deleted, or explain which experience produced a belief.

THE RELATIONSHIP IS COMPLEMENTARY, NOT COMPETITIVE

Parametric memory gives a model competence about the world. Non-parametric memory gives a system continuity about a person. Training more does not produce the second, and a substrate does not replace the first — every design in this paper assumes a strong foundation model doing the reasoning. The claim is narrow: the ability to decide what an experience meant, and whether it should return, does not arrive by scaling the training corpus.

§3 Engram: A Formalization

We define the substrate in four parts: what is stored, how significance is defined and estimated, how retrieval selects, and how the system learns from consequence.

3.1 Data structure

An event is a typed, provenance-bearing observation:

```
e  =  ( id, source, type, t, payloadHash, parent,
        entities, policy, sig )

  type   ∈ { OBSERVE, THINK, ACT, COMMIT }
  parent   hash of prior event — tamper-evident chain
  policy   retention, retrieval and disclosure rules
  sig      signature over canonical content
```

Events induce a typed directed graph `G_t = (V_t, E_t, W_t)` whose vertices are people, artifacts, decisions, goals and concepts, and whose edges are typed relations — `authored`, `depends-on`, `contradicted`, `caused`. Edge type is preserved rather than collapsed into a scalar, which is what §2.2 showed a metric space cannot do.

DEFINITION — ENGRAM

An Engram is a vertex together with its significance state: an experience that has earned the right to influence future inference. Storage is not sufficient for Engram status; significance is.

3.2 Significance

Let T_p denote the observed forward trajectory of person p — the questions pursued, directions chosen, paths returned to. Define the salience of a vertex as the predictive information about that trajectory which is lost when the vertex is withheld:

$$\mathrm{Sal}(v) \;=\; \mathbb{E}_{(p,G,T)\sim\mathcal{D}}\left[\, I(T;G) \;-\; I(T;G\setminus v)\,\right]$$

An estimand, not a computable quantity. The salience of a vertex type is its expected marginal predictive value over the population distribution $\mathcal{D}$ of persons, graphs and trajectories.

This is an estimand, and it is deliberately not presented as computable. Two caveats are load-bearing. First, mutual information is defined over random variables; a single person's realized graph is one draw, so the quantity is well-defined only as an expectation over the population $\mathcal{D}$, estimated empirically — not computed on one realized G. Second, deleting a vertex is not by itself a valid causal intervention; Sal is an *associational* predictive-lift quantity, and any causal reading requires assumptions we do not claim to have discharged. What we actually compute is a surrogate $\hat{\mathrm{Sal}}$, fit to predict held-out trajectory (§5.2) from four topological observables — each computable from graph structure alone, *without the content of what was asked*, which matters for the privacy discussion in §6.

$$\begin{aligned}
&\mathrm{depth}(v) && \mathbb{E}[\,\ell \mid v \text{ active}\,] - \mathbb{E}[\,\ell\,] \\
&&& \text{expected path-length gain when } v \text{ is active} \\[1ex]
&\mathrm{return}(v) && \frac{1}{n_v}\sum_{i}\left[\,1-\exp\!\left(-\Delta t_i/\tau\right)\right] \\
&&& \text{mean over the } n_v \text{ returns to } v \text{ — bounded in} \\
&&& (0,1).\ \tau \text{ sets the gap that "counts" as} \\
&&& \text{decay-resistant; } n_v \text{ tracked separately, with} \\
&&& \text{shrinkage toward prior for low-evidence vertices} \\[1ex]
&\mathrm{convergence}(v) && H(\mathrm{origins}(v)) \\
&&& \text{entropy of the distribution of path origins} \\
&&& \text{terminating at } v \text{ — many unrelated routes in} \\[1ex]
&\mathrm{displacement}(v) && D_{KL}\left(\pi(\cdot \mid v) \;\|\; \pi_0(\cdot)\right) \\
&&& \text{divergence of choice behaviour from baseline} \\
&&& \text{preference when } v \text{ is available}
\end{aligned}$$

We conjecture displacement is the most diagnostic of the four — no comparison has established this yet. The intuition: when a normally dominant interest is abandoned in favour of a rarely mentioned one, the rarely mentioned one may be carrying more weight than any frequency count reveals. Of the four, it is the one that most directly separates significance from attention in principle; whether it does so in practice is part of what §5.2 measures.

$$\hat{\mathrm{Sal}}(v) \;=\; w^{\top}\left[\,\mathrm{depth},\ \mathrm{return},\ \mathrm{convergence},\ \mathrm{displacement}\,\right]$$

The weight vector w is fit by regression against held-out next-selection prediction (§5.2), not chosen by hand.

THE STRUCTURAL ANALOGUE, AND WHERE IT BREAKS

PageRank [1] observed that a document need not repeat a term to be authoritative on it — authority is latent in the structure of links pointing toward it. The corresponding claim here is that a memory need not recur to be significant: significance is latent in the structure of paths that bend toward it.

The analogy is precise in mechanism and *imprecise in scope*, and the difference is the harder problem. PageRank operated over a single shared graph, so every participant benefited from every link from the first query onward. A meaning graph is singular to one person and begins empty. The proposed response — and this is a *bet, not a result* — is to not learn an individual's meaning from their own history alone, but to learn a meta-structure across many shallow graphs (which topological signatures indicate salience in general) and apply that prior to a sparse one. Whether such signatures transfer across people is exactly the cold-start risk in §6; if they do not, this reduces to slow per-user learning.

3.3 Retrieval as constrained selection against a submodular surrogate

Given non-monotonicity, top-*k* is the wrong selection rule for a second reason beyond scoring: the top *k* individually-best memories are often near-duplicates, so the set carries less information than its members suggest. The right formulation optimizes the *set*. The ideal objective is:

```
M*(q,t)  =  argmax   I( Y ; M | q )
            M ⊆ 𝒞(q,t)
            |M| ≤ b
            policy(M) = allowed
```

𝒞(q,t) is the candidate set, not the whole graph V_t (see §4). It is bounded and small, which is what keeps this tractable.

A caveat the earlier draft skipped. Conditional mutual information is *not* submodular in general — memories can be synergistic (jointly informative beyond their parts), which makes `I(Y;M|q)` supermodular in exactly the cases where interaction matters. So the Nemhauser–Wolsey–Fisher [2] `(1 - 1/e)` guarantee does *not* attach to the true objective. What we do instead is optimize a submodular surrogate $\hat{F}$ — a coverage / facility-location objective over embeddings plus explicit type-diversity — for which greedy carries the `(1 - 1/e)` bound *with respect to* $\hat{F}$. We buy tractability and a guarantee against the surrogate, and accept an unquantified gap $\hat{F} \approx I(Y;M|q)$. That gap is honest and it is the price of running in real time.

CONSEQUENCE — WHY THE BOARD IS NOT A TOP-3 LIST

Selecting the three highest individually-scored candidates maximizes a *modular* objective, $\Sigma\ \text{score}(e_i)$. Selecting the best *board* maximizes a set

objective. These differ whenever candidates overlap, which they typically do when generated from one answer. The diversity requirement in a slate is not a decoration; it is what set-level selection looks like at `b = 3`. This is a modelling argument, not a claimed optimality result.

3.4 Learning rule: behavioral credit assignment with decaying eligibility

Significance is not fixed at write time. A detail can become decisive months later; an event that seemed pivotal can become irrelevant. Credit for a later outcome must flow back to the memories that were active earlier, which is a credit-assignment problem. We borrow the *eligibility trace* mechanism from temporal-difference learning [3], but the per-step signal below is not a TD error — there is no reward-plus-bootstrapped-value term — so we call it behavioral credit assignment rather than overloading "TD".

The per-step signal is a scalar outcome score over observed behavioural features, requiring no counterfactual model inference:

```
δ_t  =  w₁·continuation  +  w₂·normDepthGain  +  w₃·delayedReturn
        −  w₄·correction   −  w₅·abandonment

z_t(v)      =  γ λ · z_{t−1}(v)  +  1[ v ∈ M_t ]     sparse eligibility
S_{t+1}(v)  =  S_t(v)  +  η · δ_t · z_t(v)
```

δ_t is a single number read from what the person actually did after M_t was shown, not a leave-one-out forward pass through the model. The weights w are part of what the proposed experiment (§5.2) must calibrate; the sign pattern — reward continuation and delayed return, penalize correction and abandonment — is the design prior. The trace z is nonzero only on the active path, so each update touches a handful of nodes.

Two definitions, kept distinct. The *conceptual* target of credit is the counterfactual influence of `v`, which one could write as a log-probability difference `log P(a|q,M) − log P(a|q,M\v)`. Computing that per active memory would require a separate model pass per memory per turn — latency-fatal, and we do not propose it. The *proposed* estimator is the scalar behavioural `δ_t` above, which costs nothing beyond observing the user, and whose weights are unfit until §5.2. This learning rule is designed, not deployed (§4).

Outcome records are compact and can remain local:

```
r  =  ( experience, selection, latency, continuation,
      depthGain, correction, return, t )
```

A caution the estimator cannot escape: even these signals measure *engagement*, not decision quality. A chosen door records attraction or predicted preference; continuation and return record that a direction held someone's interest. None of them establishes that the memory *improved the eventual decision* — that is a stronger claim requiring a different study (§5.2). The signals are a better proxy for significance than a raw click, and they are still a proxy.

3.5 Revision, supersession and forgetting

A store that only accumulates becomes internally inconsistent. When two Engrams assert incompatible values of the same predicate, the later assertion with adequate provenance supersedes the earlier:

```
supersede(v_old, v_new)  ⇒
      status(v_old) ← SUPERSEDED_BY(v_new)
      v_old retained, excluded from default retrieval
```

Retained rather than deleted: "what did I used to believe, and when did that change" is itself a legitimate query, and deletion makes it unanswerable. Belief revision here follows the AGM tradition — contraction and revision as operators, not as garbage collection.

Forgetting is then the policy consequence of low significance: an Engram whose `S(v)` stays below threshold across a decay horizon becomes ineligible for default retrieval. A threshold alone does not bound the eligible set — millions of memories could sit above any fixed threshold — so what bounds `N` in the §2.3 saturation analysis is thresholding plus an explicit top-`B` eligibility budget: significance thresholds prune the clearly-irrelevant, and the budget caps what remains eligible at any time, with the rest retained but off the default retrieval path. Forgetting is thus a feature of the substrate rather than a limitation of it.

§4 System Anatomy: why.com

why.com is the first-party reference application and the research instrument. Its loop is deliberately narrow: a question yields one compact causal answer and three typed directions; the person chooses one; the choice, its depth and any return are recorded.

The loop is chosen because it produces an unusually clean signal. Search observes consumption. Feeds observe engagement. Conversational assistants observe dialogue. A three-door board observes deliberate inquiry under a forced, legible choice — the person selects one direction and simultaneously declines two known alternatives, which is precisely the `displacement` observable of §3.2 in its most measurable form.

WHAT THE DEPLOYED LOOP ACTUALLY IS, TODAY

MEASURED

The current production system is deliberately minimal, and narrower than an earlier draft of this report claimed. A single structured generation emits the answer and exactly three questions — a plain ordered array, schema-constrained to `minItems: 3, maxItems: 3`, described to the model as "underneath, twist, beyond." There is no candidate pool, no typed-role enum, no anchor/gap fields on the wire, and no separate selection or judging pass in production. The model is asked to write three distinct directions and the system hydrates them; a light validity check rejects a turn only if it fails to return three usable questions. That is the whole live mechanism.

An earlier version of the codebase did generate a larger candidate set with typed roles, anchor/gap fields, and a board-selection step, and an earlier draft of this paper measured that version. A refactor removed that machinery. Reporting the five-candidate pipeline as the

current deployed system was an error, now corrected: the selection stage the paper theorizes in §3.3 is *not* presently in production.

THE CONSTRAINT THE REMOVED VERSION ENCODED — AS A DESIGN PROPOSAL

PROPOSED

The prior version's validity contract is worth preserving as a design target, because it operationalizes the intuition behind Observation 1. Its load-bearing conditions were: an `anchor` phrase that must appear *verbatim* in the answer (a direction cannot exist unless it is grounded in something the answer said), and a `gap` that must contain at least one concept *absent* from the answer (rejecting a paraphrase disguised as a question). These are a *syntactic operationalization* of "grounded but not redundant" — not an empirical validation of Observation 1, which they cannot provide.

PROPOSED

To those we add a reserved-role constraint: of the three directions offered, at least one must be *divergent* — a role that departs from the parent answer's line of reasoning rather than deepening it. Enforced as a constraint on the slate rather than as a weight in a score, so that set-level diversity cannot be competed away by a relevance term that will always favour the most similar direction. Re-introducing the validity contract together with this constraint is the near-term product proposal, and §7.1 gives the independent reason it matters.

WHAT IS DESIGNED AND UNBUILT

MEASURED

The client is not stateless. It keeps persistent local threads, active-path context, a local curiosity graph, outcome records, and a single earned earlier-path cue that can influence the third question. What it does *not* have is the piece this paper is about: no learned cross-session significance model, no consequence-driven ranking, and no persistent Engram substrate shared across the population.

PROPOSED

So the gap between the product and Engram is specific: the persistent *shared* graph, the significance estimator, the consequence ledger and learning rule of §3.4, the supersession operator, and any selection stage. Outcome signals are recorded client-side but rank nothing, and the test suite enforces that ranking stays off the critical path. The deployed product today is a locally persistent curiosity system with no learned significance model; the substrate this paper formalizes is a research programme built on top of it, not a description of running code. Every section above should be read against that boundary.

§5 Observations and Evaluation

5.1 What has been measured

MEASURED

Over a 72-hour organic traffic event in August 2026, the deployed system served:

QUANTITY	VALUE	METHOD
Unique visitors	677,000	Analytics, 72h window
Tokens processed	317.8M	Provider accounting
Inference cost	≈ $67	Provider accounting
Implied requests	≈ 265,000	Derived at ~1,200 tok/request
Cost per visitor	**$0.0001**	Derived
Requests per visitor	0.39	Derived

What the cost figure establishes is narrow. At 1×10^{-4} per visitor, *generating three directions* is cheap enough that selection experiments are economically feasible — one can afford to try many selection policies over the same traffic. It does not establish the cost profile of the proposed substrate: graph construction and maintenance, significance estimation, any counterfactual evaluation, and the learning loop are all unbuilt and unmeasured, and the $67 says nothing about them.

The 0.39 requests-per-visitor figure is reported because it is unfavourable and material: a majority of arrivals did not complete a single question, so the loop described in §4 was exercised by a minority of the traffic. Any claim about the substrate rests on that minority.

5.2 Experiment 1 — does significance predict the next selection?

PROPOSED

The central claim of §3.2 is a prediction about behaviour and can be refuted. The weak form of the experiment — significance versus recency, frequency and cosine — is not enough, because §2.2 already conceded that the honest competitor is *mature* RAG, not bare cosine. The experiment must therefore run a full ladder of baselines, all conditioned on the same underlying model and the same histories:

RETRIEVAL POLICY	WHAT IT ISOLATES
Recency	time alone
Frequency	count alone
Cosine top-k	bare similarity
Cosine + metadata / temporal filter	similarity + validity
MMR / diversity retrieval	redundancy control
Cross-encoder reranking	learned relevance
Hybrid (lexical + dense)	coverage

RETRIEVAL POLICY	WHAT IT ISOLATES
Long context (no retrieval)	see-everything baseline
Strong learned ranker	a fair modern opponent
Engram significance `Ŝal`	**the proposal**

Protocol. Hold out the final selection of each path; rank candidate directions under every policy; compare next-selection accuracy and mean reciprocal rank, within a single session and across 30- and 90-day intervals. To pass, Engram must beat *every* baseline — including the mature-RAG conditions — at every horizon. If it wins only within a session, significance is a session-local artifact and the durable-memory claim fails.

Guarding against leakage and false positives. The four observables (`depth`, `return`, `convergence`, `displacement`) must be computed *only* from events available *before* the held-out selection — otherwise the target leaks into its own features. The design is preregistered with a minimum lift fixed in advance; splits are at the *user* level, not the event level, so no person appears in both train and test; and every reported difference carries a confidence interval.

EXPERIMENT 2 — AND DOES IT ACTUALLY HELP?

Passing Experiment 1 proves *personalization*, not *intelligent memory*. Predicting the next selection is necessary and not sufficient: an engagement-optimizing feed would also pass it, and `depth` and `return` can equally track compulsion, confusion or entertainment. Experiment 2 is therefore the load-bearing one — the test that can fail after Experiment 1 has already passed — and it is specified here rather than gestured at.

EXPERIMENT 2 — PROTOCOL

PROPOSED

Design. Within-subject, blind, randomized order. Each participant receives answers to matched questions under two conditions using the *same* model and the same underlying history.

```
A.  Engram-conditioned : context selected by Ŝal
B.  Control            : context of EQUAL TOKEN LENGTH,
                         selected by the strongest §5.2 baseline
```

Matching token length is the critical control. Against an empty or shorter control, any positive result would only show that more context helps — a conclusion this paper does not need and cannot use. The comparison must isolate which memories were selected, holding how much constant.

Measures, in descending order of evidential weight:

- Task outcome, where ground truth exists — accuracy on questions with a verifiable answer.

- Correction rate — how often the person had to restate or repair what the system assumed.
- Decision confidence, self-reported before and after, with the delta as the statistic.
- Blind pairwise preference — weakest of the four, since preference tracks fluency as readily as usefulness, and reported last for that reason.
- Seven-day recall — whether the person can still reconstruct what they concluded, which distinguishes an answer that landed from one that merely pleased.

Controls and preregistration. Raters and participants blind to condition; presentation order randomized; splits at the participant level; a minimum effect size fixed in advance; confidence intervals on every reported difference. Engagement measures (session length, doors opened) are recorded as *diagnostics only* and explicitly excluded from the success criterion — a variant that raises engagement while lowering task outcome counts as a failure, and stating that in advance is what separates this from an engagement study.

THE FALSIFYING OUTCOME

If Engram-conditioned answers are preferred at chance and produce no improvement in outcome, correction rate or retained understanding — *while Experiment 1 passes* — then significance predicts behaviour without improving it. That is the specific result that would refute the thesis while leaving a working personalization system behind, and it is the outcome this paper most needs to rule out. Neither experiment has been run.

5.3 Modeled projection

MODELED

Under the §2.3 saturation model with $\Delta/\sigma = 5$ fixed by the encoder, top-1 recall at a lifetime corpus of 10^9 events depends on how large the *eligible* set is kept. The rows below fix the eligible set to an absolute size B (a budget, per the §2.3 correction — a *constant fraction* would not bound it, since $\rho \cdot 10^9$ still grows):

ELIGIBILITY POLICY	ELIGIBLE SET	Δ/Σ NEEDED	OUTCOME AT Δ/Σ = 5
Similarity over full history	10^9	7.72	recall collapses
Fixed budget, B = 10^6	10^6	6.54	degraded
Fixed budget, B = 10^3	**10^3**	**5.00**	**recall held at ~90%**

These are consequences of the §2.3 model, not observations, and they inherit every one of its idealizations (§2.3 lists where it can fail). Read as a design target rather than a prediction: if the eligible working set is held to a fixed budget rather than a growing fraction of a life, the saturation term stops growing with lifetime history. The open question is not whether a

small budget helps — arithmetically it must — but whether significance is the right criterion for choosing what occupies it. That is what §5.2 tests, and it is unproven.

§6 Limitations

Cold start. A meaning graph begins empty, and the meta-structure prior of §3.2 is unproven. If topological salience signatures do not transfer across people, the approach reduces to per-user learning on sparse data, which will be slow to the point of impracticality.

The estimator may not track the definition. The four observables are a hypothesis about what predicts `Sal`. Nothing guarantees the surrogate correlates with the population estimand it stands in for. §5.2 is designed to detect exactly this failure.

Ranking is unbuilt. §3.4 is specified, not deployed. No claim in this report should be read as asserting that the running system currently learns from consequence.

Grounding. Directions classified as requiring external grounding do not currently receive grounded generation in the same pass; source attribution is performed as a subsequent retrieval step. Sourced-truth guarantees should not be claimed until that path is unified.

Topology reduces, but does not deliver, privacy. Salience is estimated from the *structure* of behaviour — depth, return, convergence, displacement — so content minimization is possible: structure can be recorded where the content of what was asked is deliberately withheld. This *reduces* disclosure; it does not constitute privacy, and the earlier draft's "architectural privacy" claim was too strong. Behavioural topology is itself sensitive — the shape of what someone pursues can reveal interests, conditions and traits as surely as the words can. Content minimization is one control among several (retention limits, local-first storage, policy boundaries), not a guarantee. A second, softer commitment: an inferred salience model is not surfaced to the person it describes — it may shape what the system offers, but announcing what it believes is a product choice we make against, because a correct inference delivered without invitation is not insight to the person receiving it.

Cost claims are narrow. The §5.1 figure of `$0.0001` per visitor establishes that *generation* is cheap for today's locally persistent curiosity product. It says nothing about the cost of the unbuilt system — graph construction and maintenance, significance estimation, and any future selection or learning stage each carry their own cost, none of which the $67 measures.

Selection is not in production. The board-selection stage this paper theorizes (§3.3) was present in an earlier build, removed in a refactor, and is not currently deployed (§4). The paper argues for re-introducing it; it does not describe a running selector.

§7 Failure Modes and Mitigations

PROPOSED

The limitations in §6 concern what has not been proven or built. This section concerns something different: how a substrate of this design breaks once it is running, and who bears the cost. Persistence and learned selection introduce failure modes that a stateless system does not have, and we treat the mitigations below as release requirements rather than optional hardening.

7.1 Self-reinforcement and narrowing

`Ŝal` is estimated from behaviour, then shapes which directions are offered, which shapes subsequent behaviour. That is a closed loop, and the update rule of §3.4 contains no damping term. Left unchecked it concentrates eligibility mass on an increasingly narrow band — the failure the substrate exists to oppose, arrived at by a different route. A system whose stated purpose is to widen what a person considers must be built so that it cannot quietly narrow it.

Mitigation. The reserved-role constraint of §4 — at least one divergent direction per slate, enforced structurally rather than scored — is the primary damper, and this is the independent justification for it. Beyond that: a diversity floor in the selection objective, and a ceiling on the share of the eligibility budget `B` that any single cluster of related Engrams may occupy.

7.2 Asymmetric cost of false negatives

The two error directions are not symmetric. Surfacing something irrelevant is cheap and self-correcting: the person ignores it. Failing to surface something that mattered is expensive and *invisible* — the person cannot know what was withheld, so the error generates no signal and never enters the learning rule of §3.4. The eligibility budget of §2.3 makes such omissions structural rather than accidental, and nothing in the current design prices them.

Mitigation. Note first what §3.5 already guarantees: low significance makes an Engram *ineligible for default retrieval*, never deleted, and superseded assertions are retained rather than discarded. Nothing is lost; it becomes unreachable by default. The gap is therefore not retention but recovery: an explicit path by which a person can reach beyond the eligible set, an audit log of what the selector declined so omissions become inspectable, and thresholds calibrated against *recall* on held-out follow-up tasks rather than precision alone.

7.3 Adversarial manipulation of behavioural topology

§3.2 treats behavioural topology as evidence precisely because it is hard to fake casually. It is not hard to fake deliberately. `depth`, `return`, `convergence` and `displacement` are all functions of traversal, and traversal is a control surface: a motivated party can walk a graph to inflate the significance of chosen nodes, and injected content can shape which directions are offered in the first place. The untrusted-context boundary of §3.1 governs what may enter the graph; it does not govern what a sequence of legitimate-looking traversals can do to the ranking once inside.

Mitigation. Rate-limit significance updates per source; detect anomalies in update magnitude and in the source distribution feeding a node's trace; require provenance before injected content may influence traversal; and give the person direct controls over their own memory, including inspection and deletion of high-significance nodes. User-initiated deletion is distinct from the system-internal supersession of §3.5 — the former is a right, the latter a policy — and the substrate must honour both.

7.4 Error compounding across sessions

This is the failure class the architecture genuinely introduces. In a stateless system a bad inference dies with the turn. Here, an incorrect significance estimate persists in `S(v)`, shapes the next retrieval, and is reinforced by the behaviour that retrieval produces. Errors

do not merely survive; they accrue interest. The eligibility trace of §3.4, which exists to let a detail gain significance long after it was recorded, propagates a mistaken credit assignment by exactly the same mechanism.

Mitigation. A decay term reducing `S(v)` in the absence of confirming behaviour, so significance must be re-earned rather than merely retained; periodic re-evaluation of high-significance nodes under fresh context; and versioning of the graph sufficient to roll back a bad update window. The tamper-evident event chain of §3.1 already provides the substrate for that rollback.

7.5 Instrumentation coupling and surface drift

This is the failure mode that most threatens the substrate claim itself, and the current design is exposed to it. The four observables of §3.2 are computed from interaction events. If those events are defined against *one product's interface* — a three-door board, a particular path length, a specific abandonment gesture — then significance is not a property of the person's history; it is a property of why.com's front end. Three consequences follow. A UI redesign shifts the baselines and silently invalidates the fitted weights `w`. A change in user habit does the same without any code change. And most seriously, observables defined this way are not comparable across surfaces: a desktop client or a third-party application through an API would emit incommensurable signals, so the "one substrate, many surfaces" claim would not survive contact with a second surface. A memory layer that works only behind one interface is a product feature, not a substrate.

Mitigation. Define the observables over an abstract interaction model rather than over UI events. The model requires only three things from any surface: that a bounded set of alternatives was *presented*, that one was *selected* (or none), and that what followed is *observable*. `depth` becomes continuation length in that surface's own units, normalized per surface; `displacement` becomes selection against the prior-preferred alternative *within the offered set*, which is well-defined wherever a choice set exists; `return` and `convergence` are already defined over graph topology and time, not over gestures. Then: version the instrumentation schema alongside the graph, refit `w` per surface with a shared prior rather than assuming transfer, and hold a control cohort on the previous surface through any migration so drift is measured rather than inferred.

Two honest notes. Normalization across surfaces is an empirical question this paper does not settle — whether a "deep" path on one interface is commensurable with a "deep" path on another must be measured, not assumed, and §5.2's protocol would need a per-surface stratum to test it. And this constraint should be read as a *discipline on the design* rather than a repair: an observable that cannot be stated without naming a button does not belong in the substrate.

WHY THESE ARE RELEASE REQUIREMENTS

Each mitigation above is cheap relative to the failure it prevents, and each becomes far more expensive to retrofit once a graph has accumulated. A memory system that decides what deserves to return is not neutral infrastructure: it is a system with a point of view about a person, operating

where they cannot easily audit it. The obligations that follow are load-bearing, not decorative.

§8 Conclusion

The next layer of the stack may not be a better model. It may be a better memory for models to reason from.

We have argued two things carefully, and no more than they support. Context stuffing has per-query cost that grows with a life, which no unit-price decline removes — that argument is robust. And embedding similarity *need not preserve memory-value ordering*, so nearest-neighbour ranking is not value ranking; under a stylized model, similarity retrieval also degrades as `V(2 ln N)` as a store accumulates. The saturation result is a directional pressure under stated idealizations, not an assumption-free impossibility — scaling similarity alone does not solve selection, consolidation, revision and consequence-learning, which is the defensible claim, weaker than "impossible."

The proposed alternative treats significance as the primitive — an estimand defined as expected marginal predictive value over a forward trajectory, estimated from behavioural topology, selected against a submodular surrogate rather than a modular score, and revised by consequence. We are explicit that this is a research programme: the significance estimator, the learning loop, and the persistent shared graph are designed and unbuilt. The deployed product today is a locally persistent curiosity system — threads, active path, a local graph, outcome records — with no learned significance model on top. A database retrieves what was stored; a vector index retrieves what looks similar; what we propose to build would retrieve what matters — if the falsifying experiment of §5.2 holds, which it has not yet been run to decide.

why.com is where the claim gets tested first, because its three-door loop produces an unusually clean observation of deliberate inquiry: a person selects one direction and declines two known alternatives, in the open, many times a day. The near-term work is concrete — re-introduce a grounded-but-divergent selection stage with a reserved slot for the unexpected direction — and the standard by which it succeeds is not elegance but a number: whether significance predicts what a person does next better than recency, frequency and cosine. Until that number exists, this is a thesis with a mechanism, honestly labeled as one.

§9 References

[1] S. Brin and L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine.* Computer Networks and ISDN Systems, 1998.

[2] G. Nemhauser, L. Wolsey and M. Fisher. *An analysis of approximations for maximizing submodular set functions — I.* Mathematical Programming, 1978.

[3] R. Sutton. *Learning to Predict by the Methods of Temporal Differences.* Machine Learning, 1988.

[4] N. Liu et al. *Lost in the Middle: How Language Models Use Long Contexts.* TACL, 2024.

[5] E. Tulving. *Episodic and Semantic Memory.* In Organization of Memory, 1972.

[6] C. Alchourrón, P. Gärdenfors and D. Makinson. *On the Logic of Theory Change.* Journal of Symbolic Logic, 1985.

[7] A. Vaswani et al. *Attention Is All You Need.* NeurIPS, 2017.

[8] P. Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* NeurIPS, 2020.

[9] C. Packer et al. *MemGPT: Towards LLMs as Operating Systems.* arXiv, 2023.

[10] J. Park et al. *Generative Agents: Interactive Simulacra of Human Behavior.* UIST, 2023.

WHY / Engram — Technical Report ET-01 Ockams, Inc. Version 1.0 — August 2026

Reference application: why.com