AUTOSCIENTIST: AN AUTONOMOUS AGENT FOR LONG-HORIZON SCIENTIFIC DISCOVERY VIA CONNECTOME MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Accelerating the rate of scientific discovery requires autonomous agents capable of continuous literature assimilation, rigorous hypothesis generation, and empirical verification over extended horizons. Existing autonomous research agents rely on linear flat-file scratchpads or unindexed vector databases for memory. These representations suffer from context window saturation, catastrophic forgetting of early observations, and an inability to detect contradictions across multi-hop causal chains. To resolve these limitations, we introduce AutoScientist, an autonomous discovery system organized around a biologically inspired, multi-relational graph connectome. AutoScientist treats memory as an active graph, modelling scientific knowledge as an evolving network of typed concepts, causal relations, and empirical findings whose structure persists as it grows. The architecture executes diurnal discovery cycles: active phases acquire literature, formulate structured hypotheses, and run computational simulations in code sandboxes, while nocturnal phases apply multi-pass consolidation algorithms that resolve logical contradictions, prune unreinforced associations, and synthesize structural analogies. We evaluate AutoScientist across 20 longitudinal discovery cycles and three disparate scientific domains: non-equilibrium statistical mechanics, systems biology gene regulatory circuits, and non-binary error-correcting codes. Empirical evaluations show that the agent's memory develops a sparse, modular small-world topology that no degree-matched random null reproduces. In controlled multi-hop retrieval and falsification benchmarks, graph memory holds every chain step and resolves every recorded contradiction, whereas dense vector retrieval recovers only a third of a chain's interior, never exceeds chance on a recorded contradiction, and fills nearly half of its retrieved context with decoys.A sweep of working-memory capacity shows that bounded focus strikes an optimal balance between prompt efficiency and thrashing avoidance, and neuromodulatory state transitions prevent stagnation on intractable problems. The full agent runs on a single locally hosted 32B open-weight model, and matches the suite scores reported for frontier commercial systems on the same long-memory sets, so strong long-horizon scientific discovery is within reach of modest, fully local hardware. These results establish connectome-based memory as an effective computational foundation for long-horizon autonomous scientific research.

1 INTRODUCTION

An autonomous research agent is bounded less by what it can read than by what it can remember. A long-horizon campaign generates thousands of intermediate results while a context window holds a few hundred, so the agent must choose every cycle which subset of its own history to reason over, and the memory representation makes that choice.

Two representations dominate practice, and they fail in the same place. Flat-file logs append entries and retrieve by lexical overlap, so a log returns a claim and the result that refuted it with no indication that they disagree. Dense vector retrieval recovers passages resembling the query but cannot follow a chain through intermediate items, and ranks an unsupported sentence highly whenever it

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.

shares vocabulary with the question. Neither stores the difference between A supporting B and A contradicting B, which we take to be the central object an epistemic memory must hold.

We present AutoScientist, a research agent whose memory is one directed, weighted, typed graph revised in place. The design borrows from a different literature: biological brains sustain function under continuous synaptic turnover, with wiring that is sparse, clustered and short-ranged, and a homeostatic process that prunes during sleep. We ask whether an agent's knowledge store, under an objective that rewards only progress on a scientific question, converges on the same organization and what that buys. Contemporary autonomous research agents reach their results through frontier commercial APIs; we ask what a structured memory makes possible when every role runs on a single locally hosted open-weights model. Our contributions are:

- **Typed signed edges decide what is retrievable** (§4.2). Across nine seeded memories the graph recovers multi-hop chains in every case, while iterative dense retrieval recovers 32% of chains and almost no opposing pair; vocabulary-matched but unsupported content fills 46% of its slots and none of the graph's.
- **Consolidation is load-bearing** (§4.4). Sleep is what turns an append-only store into a sparse, clustered one: disabling its decay pass reproduces the entire effect of disabling sleep, and no other isolated pass moves modularity, clustering or the weight distribution.
- **The memory converges to a small-world, modular regime** (§4.5). Over twenty cycles the small-world index rises from 1.8 to 10.2, modularity reaches 0.79 ± 0.02, and edge density falls 85%. None of this is rewarded by the objective, and all of it is measured against degree-matched random nulls, so the organization emerges from the memory's own dynamics with no reward term encouraging it.
- **The same agent transfers across domains** (§4.6). Three missions, three seeds each, give 25 experiments in physics against 6 for a notebook at matched cycles, and throughput scales with backbone size from 3 to 11. On LongMemEval-S and LoCoMo-10 the memory matches dense retrieval and exceeds a flat store by 37 and 13 points at 96% of the context cost.

Where a benchmark exercises the generator more than the memory the arms converge, and the analysis isolates the reason (§4.6).

2 Related Work

Autonomous research agents. End-to-end discovery systems now span the research pipeline. The AI Scientist line generates ideas, runs experiments and writes papers, and its second iteration replaces the fixed pipeline with an agentic tree search over experiments (Lu et al., 2024; Yamada et al., 2025); Kosmos couples a literature-reading loop to a data-analysis loop and reports discoveries corroborated by independent scientists (Mitchener et al., 2025). Benchmarks have followed, scoring end-to-end systems on fixed task sets under matched budgets (Shanghai Artificial Intelligence Laboratory, 2026; Gao et al., 2026). These systems differ from ours in where their state lives: in the prompt, or in a run directory of files. Their focus is orchestration, which tool to call next; ours is what the agent's record of findings should be made of once it exceeds the context window, a choice we measure through probes on the store and through end-to-end behavior. They also share a dependence on frontier commercial models: the AI Scientist line runs on GPT-4o, Claude 3.5 Sonnet and o1 (Lu et al., 2024; Yamada et al., 2025), and AutoScientists on Claude Sonnet 4.6 through the Claude Code agent (Gao et al., 2026). AutoScientist runs every role on one locally hosted open-weights model (§4.1), and we report the agent that setting produces.

Graph memory and retrieval. Graph-structured memory has been applied at several granularities. GraphRAG builds an entity graph over a corpus and answers by community summary (Edge et al., 2024); HippoRAG links retrieval to a personalized-PageRank spreading process inspired by hippocampal indexing (Gutiérrez et al., 2024); A-MEM lets an agent write and re-link its own notes (Xu et al., 2025); Zep assembles a temporal knowledge graph from conversation (Rasmussen et al., 2025); Mem0 extracts and consolidates facts across sessions (Chhikara et al., 2025). Surveys catalogue the design space (Yang et al., 2026; Du, 2026). These systems store associations, temporal order or extracted facts; we store typed relations that include explicit contradiction, and we compare

against a graph-memory arm that shares our store and retrieves by personalized PageRank, so the comparison isolates the retrieval rule on identical content.

Connectome organization in artificial systems. Biological connectomes are sparse, clustered and near-critical, and these properties are argued to shape computation itself (Bassett & Sporns, 2017; Betzel et al., 2016). A separate line treats network topology as the object of study in machine-learned scientific systems (Buehler, 2025; Jin et al., 2026); we follow it, treating the agent's knowledge graph as an object to be measured and reporting clustering, path length, modularity and the small-world index against a degree-matched null.

Evaluation of long-horizon memory. LongMemEval probes five memory abilities over long histories (Wu et al., 2025); LoCoMo-10 provides long multi-session conversations with single-hop, multi-hop, temporal, open-domain and adversarial questions (Maharana et al., 2024); MemoryAgentBench unifies such suites behind incremental multi-turn protocols (Hu et al., 2026). Judged suites are sensitive to rubric wording: reported scores move by tens of points when the rubric changes. We therefore report under a strict rubric, repeat under a lenient one, re-score every row with a second judge family, and audit the strict judge by hand (Ravidesh et al., 2026; Panigrahi et al., 2026). Expert-validated benchmarks exist for the discovery setting (Majumder et al., 2024); we use one to test whether memory architecture separates on questions requiring re-analysis.

3 AUTOSCIENTIST

3.1 PROBLEM FORMULATION

The agent receives a scientific question q, a set of foundational concepts, a tool budget and no supervision, and it runs for T diurnal cycles. Each cycle it may read literature, generate hypotheses, execute code in a sandbox and revise its memory; at the end of each cycle it consolidates. We evaluate three things: whether the memory answers queries that compose items which disagree; whether the components of consolidation each do work, by disabling them one at a time; and whether memory and agent transfer to a different domain unmodified. Write the memory after cycle t as a directed labeled graph

$$G_t = (\mathcal{V}_t, \mathcal{E}_t, w_t), \tag{1}$$

where $\mathcal{V}_t$ is the node set, $\mathcal{E}_t$ the set of directed edges and $w_t : \mathcal{E}_t \to [0, 1]$ the edge-weight map. Fig. 1 shows the components and how one cycle moves through them. The annotations matter more than the topology; Appendix A states the model as operators and Appendix C gives the schema and every constant.

3.2 THE CONNECTOME STATE

A node $v \in \mathcal{V}$ is an atomic epistemic item carrying a natural-language statement; one of seven types (MISSION, QUESTION, CONCEPT, HYPOTHESIS, EMPIRICAL, SYNTHESIS, GAP); an epistemic status in {settled, uncertain, contradicted, hypothetical}; an importance weight; a source quality, higher for literature and sandbox results and lower for sleep-produced speculation; and a 384-dimensional embedding in a separate nearest-neighbor index.

A directed edge $e = (u, v)$ carries one of seven types: CAUSES, SUPPORTS, CONTRADICTS, STRUCTURALANALOGY, EMPIRICALLYTESTED, TOWARDMISSION, ASSOCIATED. SUPPORTS and CONTRADICTS are stored as distinct types; we do not fold them into opposite weights on one edge, so a claim and its refutation remain separately retrievable and neither overwrites the other. An edge also carries a weight $w_e \in [0, 1]$, a confidence and a provenance label. A CONTRADICTS edge at high confidence states that the agent is sure the two items disagree, which is information no scalar similarity can express.

Weights decay. A synapse not reinforced by new evidence is halved every τ days,

$$w_e(t + \Delta t) = \max\big(0.01,\ w_e(t)\, 2^{-\Delta t/\tau}\big), \tag{2}$$

with a single half-life $\tau = 1/\text{decay_rate} = 100$ at the released rate and Δt the time elapsed since the previous consolidation. Edges below 0.02 are severed; edges carrying a contradiction, a mission

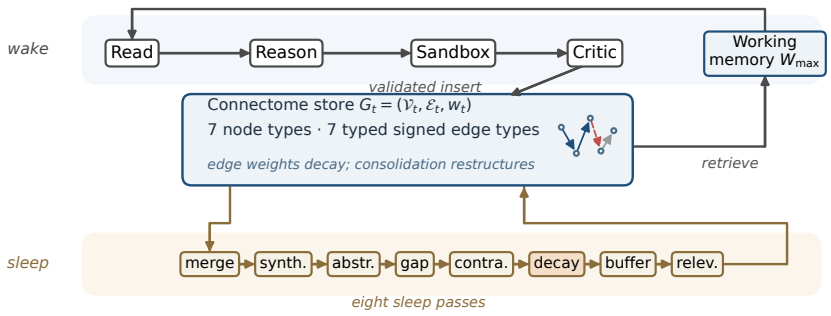


Figure 1: The AutoScientist architecture and the flow of one diurnal cycle. During wake the processes run in order, read, reason, test, gate, and every accepted item enters the connectome store through the critic's gate; a bounded working memory of size $W_{\max}$ retrieves from the store and hands the selected context to the next cycle. During sleep eight passes run in fixed order, entered from the store and folding back into it. Decay, shaded, is the pass that shapes aggregate structure at the horizons we measure (§4.4). Inset: the store's edges are typed and signed, with a contradicts edge drawn dashed.

link or a sandbox result above 0.7 confidence are exempt, so what decays is unconfirmed associative and supporting structure. The decay keeps the store finite, and §4.4 shows it does more structural work than the rest of consolidation combined; the generative passes earn their place by adding the content that keeps retrieval rich.

Three quantities recur below: the small-world index S, which compares observed clustering C and mean shortest-path length L against a degree-preserving random null, the Newman–Girvan modularity Q, and edge density ρ,

$$S = \frac{C/C_{\text{rand}}}{L/L_{\text{rand}}}, \qquad \rho = \frac{|\mathcal{E}|}{\binom{|\mathcal{V}|}{2}}, \tag{3}$$

$$Q = \frac{1}{2m} \sum_{i,j} \Big[A_{ij} - \frac{k_i k_j}{2m} \Big] \delta(c_i, c_j). \tag{4}$$

Here A is the weighted adjacency matrix, k_i the weighted degree of node i, $m = \frac{1}{2}\sum_{i,j} A_{ij}$, and $\delta(c_i, c_j)$ is one when i and j share a community. Modularity is computed by greedy maximization on the undirected weighted graph, and $S > 1$ means the graph is more clustered than the null without being much larger in diameter.

3.3 The diurnal cycle

During wake, five processes run in sequence. The reader projects extracted claims as Concept nodes with Causes or Supports edges, validating each new edge against the store. The thinker picks an open question by importance, selects a reasoning mode with an ϵ-greedy contextual bandit, and emits structured hypotheses. The sandbox writes and runs Python for each testable hypothesis and attaches an Empirical node with a verdict of supports, contradicts or inconclusive. The critic gates every insert, holding failures in a buffer for later evaluation, and a working memory of size $W_{\max}$ supplies the prompt context.

During sleep, eight passes run in fixed order, laid out in Fig. 1 and specified in Appendix A. Six restructure the graph: merge_duplicates merges near-duplicates and rewires their edges; synthesis_pass induces Synthesis hubs over dense cross-cluster groups; abstraction_pass promotes recurring patterns; gap_detection inserts Gap nodes into topological voids; contradiction_update surfaces Contradicts pairs and escalates unresolved ones; apply_decay applies the weight law and severs edges below a floor. The other two promote buffered insights and refresh mission relevance. No gradient flows into the graph from the mission objective: the store's structure is maintained by its own internal objectives alone.

Two scalars govern control. Dopamine $D \in [0, 1]$ scales the bandit learning rate, and frustration $F \in [0, 1]$ accumulates on failed or unresolved steps,

$$\eta_{\text{eff}} = \eta_0(1 + D), \qquad F_{t+1} = \min\big(1,\ F_t + 0.2\,\mathbf{1}[\text{failure}]\big). \tag{5}$$

Frustration rises by 0.2 on a failed experiment or an unresolved contradiction; when $F \geq 0.8$ the agent suspends the mission, drops D to 0.4 and incubates, wandering the graph for analogies until rest decays F to 0.4, at which point it resumes.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

All non-coding roles run on a locally hosted 32B instruct model (Qwen2.5-32B) served through Ollama, and the sandbox coder uses the same backbone; embeddings come from a local 384-dimensional sentence transformer. No role calls a commercial API: every role runs end to end on open weights, while the autonomous research agents of §2 depend on frontier commercial backbones (GPT-4o, Claude 3.5 Sonnet and o1 for the AI Scientist line, Claude Sonnet 4.6 for AutoScientists). Runs execute on two GPU servers, one cycle taking 10.9 ± 0.9 minutes on the long-horizon mission. The three missions, each run with seeds $\{101, 202, 303\}$, are: *physics of computation*, 20 cycles, on how critical phase transitions, edge-of-chaos dynamics and computational capacity relate; *systems biology*, 6 cycles, on which gene-regulatory motifs minimize variance under extrinsic noise while keeping switching fast; and *information theory*, 6 cycles, on whether non-binary LDPC codes hold channel capacity under burst noise without exponential decoding latency. Each is also run with a notebook baseline that uses the same components and appends findings to a flat file.

Probes. We evaluate memory separately from the agent. Representation probes are generated from each frozen brain: multi-hop items are chained through shared claims, and we measure whether a query returns the chain in correct order (*step continuity*), whether intermediate items survive (*interior recall*), whether the endpoint is returned (*target found*), and what fraction of returned slots hold planted unsupported content sharing vocabulary with the question (*decoy rate*). Competency probes test factual recall, recovery of the order two claims were recorded in, and resolution of a directly opposing pair. Every probe runs over the same corpus under three methods: the graph, iterative dense retrieval, and a flat store read by term overlap.

Judging. LongMemEval-S is stratified to 120 questions, 20 from each of its six abilities; LoCoMo-10 is stratified to 98 questions, 20 adversarial plus 19 or 20 from each of four categories. Both are answered by the same 32B backbone at equal context budgets. The primary judge is Gemma3-27B under a strict rubric that awards credit only for the gold answer, with token-F1 computed alongside. We repeat all 654 scored rows under a lenient rubric, re-score every row with a second judge family (Mixtral-8x7B) under the strict rubric, and audit the strict judge by hand on 50 stratified rows. Appendix D gives the agreement statistics.

4.2 WHAT THE MEMORY STRUCTURE MAKES RETRIEVABLE

Table 1 and Fig. 2(a) give both probe families on nine seeded memories, three per mission. The graph returns every chain with correct order, interior items and endpoint, and no decoy contamination, using 867 characters against 2,860 for dense retrieval and 3,268 for the flat store. Dense retrieval finds the endpoint almost as often, 0.945, but recovers the interior only 36% of the time, naming the right answer without the steps that justify it.

A decoy shares terms with the question and is supported by nothing, so dense retrieval scores it on similarity and it wins slots; the graph spreads activation along existing edges, and a decoy has none.

Factual recall is a control that every method passes; order and opposition discriminate. A flat store cannot express that one claim preceded another, and dense retrieval keeps order inside chunk text. The falsification probe asks which side of a recorded contradiction the memory holds; the answer is binary, so chance is 0.5. The graph addresses both endpoints from the paraphrase and reads the epistemic status each endpoint carries, resolving every case (1.000); permuting those statuses returns it to chance and deleting the contradiction edges leaves it intact, so the probe measures whether a

Table 1: Representation probes (top block) and competency probes (bottom block) on nine independently seeded memories, three seeds across three missions. Each cell is the mean ± standard deviation across brains, computed within a brain over all queries. The falsification row is a binary choice (chance 0.5) on the $n = 6$ memories holding a resolved contradiction; the flat store's value there is bimodal.

Probe	Connectome	Iterative dense	Flat store
Step continuity ↑	$\mathbf{1.000 \pm 0.000}$	0.321 ± 0.157	0.165 ± 0.094
Interior recall ↑	$\mathbf{1.000 \pm 0.000}$	0.361 ± 0.167	0.286 ± 0.154
Target found ↑	$\mathbf{1.000 \pm 0.000}$	0.945 ± 0.071	0.413 ± 0.136
Decoy rate ↓	$\mathbf{0.000 \pm 0.000}$	0.462 ± 0.132	0.553 ± 0.161
Context consumed (characters) ↓	$\mathbf{867 \pm 248}$	2860 ± 576	3268 ± 538
Factual recall ↑	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Recorded order ↑	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000
Opposing pair resolved ↑	$\mathbf{1.000 \pm 0.000}$	0.097 ± 0.153	0.778 ± 0.404

Table 2: Two standard long-memory suites, every question answered by the same locally hosted Qwen2.5-32B backbone at the same context budget, whereas the autonomous research agents of §2 answer through frontier commercial APIs. Absolute scores move by up to 27 points when the rubric or judge changes; the ordering does not, except on LoCoMo under the second judge. Intervals are 95% item bootstrap intervals for the strict rows. On the paired rows an exact McNemar test returns $p = 1.00$ (graph against dense) and $p < 0.001$ (graph against flat) on LongMemEval-S, and 0.79 and 0.06 on LoCoMo-10. The PageRank column is the propagation arm of Appendix F, run under the strict rubrics only.

Suite	Rubric or judge	Connectome	Iterative dense	PageRank	Flat store
LongMemEval-S ($n = 120$)	strict (Gemma3-27B)	0.550	0.558	0.533	0.183
	lenient	0.758	0.783	—	0.283
	strict (Mixtral-8x7B)	0.567	0.608	0.592	0.208
LoCoMo-10 ($n = 98$)	strict (Gemma3-27B)	0.378	0.398	0.337	0.245
	lenient	0.633	0.653	—	0.469
	strict (Mixtral-8x7B)	0.582	0.531	0.561	0.296

substrate can address and keep that field (Appendix B). Iterative dense retrieval resolves 0.097 and never exceeds chance on any memory probed, so it fails the probe on 100% of them; the flat store's 0.778 is bimodal across memories.

Paired on the query, the same gaps are decisive against dense retrieval, where exact McNemar returns $p < 10^{-4}$ on the binary probe rows and $p = 0.001$ on the opposing pair; Appendix B gives every interval and test.

4.3 Standard memory suites at a matched budget

Under the primary judge the graph and iterative dense retrieval are statistically indistinguishable, 0.550 against 0.558 on LongMemEval-S and 0.378 against 0.398 on LoCoMo-10, and both exceed the flat store by 37 and 13 points. Context is comparable across arms, 57.7k characters per question against 59.9k and 60.0k.

The fourth column shares the store but retrieves by personalized PageRank, the propagation rule behind HippoRAG-style retrieval. It lands in the same band, 0.533 on LongMemEval-S and 0.337 on LoCoMo-10, so the suites turn on the store, with the retrieval rule making little difference.

Paired on question, graph against dense gives $p = 1.00$ on LongMemEval-S and $p = 0.79$ on LoCoMo-10, so the two reach the same accuracy at this sample size, the graph doing it in 96% of the context; graph against flat gives $p < 0.001$ on LongMemEval-S under both judges and on LoCoMo-10 under the second judge, and $p = 0.06$ under the primary.

The per-ability breakdown adds two readings. The graph leads on LongMemEval-S multi-session questions, 0.300 against 0.200, and on LoCoMo-10 temporal questions, 0.500 against 0.400, both

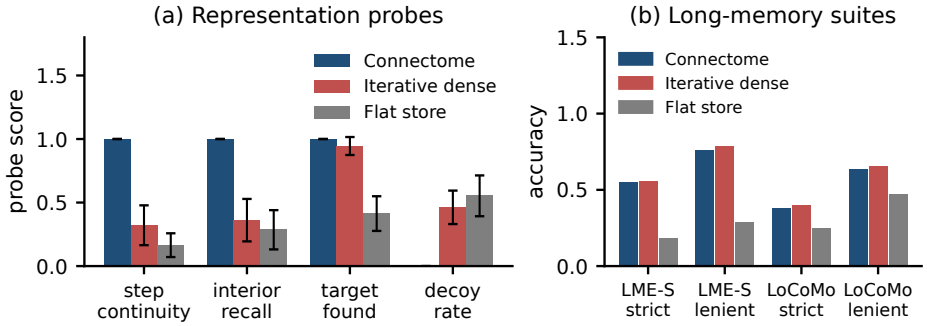


Figure 2: (a) The probes of Table 1, grouped by method; error bars are standard deviation across the nine brains. (b) The suites of Table 2 under the strict and lenient rubrics, with LongMemEval-S abbreviated LME-S. The flat store is last everywhere, and the graph and dense retrieval track closely.

Table 3: Consolidation ablation by controlled replay on the information-theory mission. Each row disables one pass and reports the paired difference from the full-sleep replay at the same seed and cycle offset, over three seeds and four offsets ($n = 12$ paired observations; the all-passes row pools every replay, $n = 96$). The structural effect concentrates in one pass: disabling decay is indistinguishable from disabling sleep entirely, and no other pass moves modularity by more than one thousandth of a point.

Disabled pass	Δ modularity	Δ clustering	Δ mean weight	Δ edges
apply_decay	$+0.0000 \pm 0.0000$	-0.0226 ± 0.0119	-0.0140 ± 0.0080	$+1.33 \pm 0.49$
merge_duplicates	-0.0007 ± 0.0005	0.0000 ± 0.0000	-0.0018 ± 0.0013	0.00 ± 0.00
contradiction_update	-0.0011 ± 0.0001	0.0000 ± 0.0000	-0.0027 ± 0.0003	0.00 ± 0.00
gap_detection	-0.0008 ± 0.0005	0.0000 ± 0.0000	-0.0021 ± 0.0013	0.00 ± 0.00
evaluate_insight_buffer	-0.0011 ± 0.0001	0.0000 ± 0.0000	-0.0027 ± 0.0003	0.00 ± 0.00
all eight passes	$+0.0000 \pm 0.0000$	-0.0226 ± 0.0115	-0.0140 ± 0.0077	$+1.33 \pm 0.47$
Reference (full sleep): $Q = 0.3990 \pm 0.0577$, $C = 0.5483 \pm 0.0693$, $\bar{w} = 0.6447 \pm 0.0219$				

of which compose items recorded at different times. The adversarial LoCoMo-10 category runs the other way: the flat store scores 0.600 against 0.200 for both competitors, because an adversarial gold answer states that the question cannot be determined and a store that retrieves nothing produces that answer by accident; excluding the category the suite reads 0.424, 0.449 and 0.154.

Agreement with the second judge family is 0.931 ($\kappa = 0.860$) on LongMemEval-S and 0.803 ($\kappa = 0.598$) on LoCoMo-10 over all 654 rows, and a hand audit of 50 stratified rows agrees with the strict judge on all but two. The LoCoMo gap sits in the adversarial category, where the second judge credits a partial refusal the primary judge does not, and where the graph overtakes dense retrieval, 0.582 against 0.531.

4.4 Consolidation and component ablations

Table 3 isolates the sleep passes against the full-sleep reference ($Q = 0.399$, $C = 0.548$, $\bar{w} = 0.645$). Disabling apply_decay reduces clustering by 0.023 and mean edge weight by 0.014 while retaining 1.3 edges per cycle that decay would have severed, and disabling all eight passes at once reproduces those numbers to three decimals. No other pass moves modularity by more than one thousandth. The attribution is insensitive to the decay timestep: Appendix H repeats it under the designed daily schedule and the wall-clock cadence of the live runs, and the differences agree to two decimals.

The logged records show why the effect concentrates in one pass: decay is the only map that removes structure, while synthesis and abstraction add nodes and edges without deleting any and no pair in these runs reaches the duplicate-merge threshold. The replay disables language-model calls, so the generative passes skip in every scenario including the reference, and the isolation covers the four non-generative passes exactly.

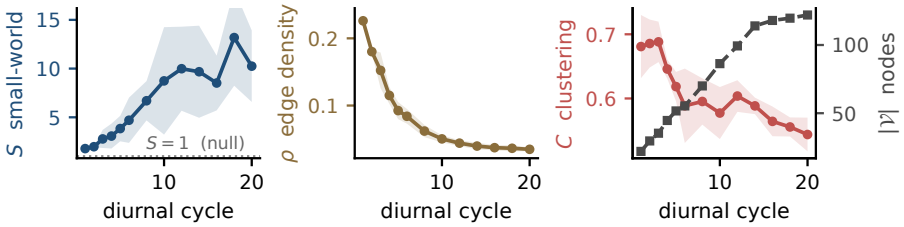


Figure 3: Emergent structure on the physics mission, three seeds, measured every cycle against a degree-matched random null. (a) Small-world index S; (b) edge density ρ; (c) clustering C (left axis) and node count $|\mathcal{V}|$ (right axis). Bands show standard deviation across seeds, and the small-world index rises monotonically in the mean. Density falls sixfold while the node count grows fivefold, so the store sparsifies faster than it grows, and no objective term rewards these quantities.

Table 4: Final state of the three missions, three seeds each: 20 cycles for physics and 6 for the other two. The last two columns are retention of the nodes present at cycle 1, the fraction still connected to the mission and the fraction still retrievable, with survival 1.000 in all three. Modularity Q is computed by greedy maximization on the undirected weighted graph and is not directly comparable to the per-cycle small-world index in Fig. 3.

Mission	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	Q	Communities	Connected	Retrievable
Physics	122 ± 7	253 ± 11	0.791 ± 0.020	31 ± 4	0.771 ± 0.056	0.955 ± 0.003
Biology	66 ± 7	185 ± 49	0.532 ± 0.059	8 ± 2	0.830 ± 0.095	0.975 ± 0.013
Information theory	36 ± 2	76 ± 13	0.429 ± 0.071	9 ± 2	0.801 ± 0.095	0.900 ± 0.068

4.5 Structure that emerges over twenty cycles

Fig. 3 follows the physics mission for twenty cycles. The store grows from 22 nodes to 122 while its edge density falls from 0.226 to 0.035, so each new node brings fewer connections than the average node already had. Clustering stays high, path length grows modestly to 3.48, and the small-world index rises monotonically from 1.79 to 10.25. The null makes this meaningful: $S > 1$ means observed clustering exceeds a degree-matched null, so the structure is not a by-product of having few edges. The biological comparison stays motivational, and we make no claim to a fitted match to measured connectome statistics. The same measurement holds across missions (Table 4), where the modularity ordering follows mission length and shows no dependence on domain.

Retention completes the picture: of the nodes present at cycle 1, 1.000 survive to the final cycle in all three missions, $77 \pm 6\%$, $83 \pm 10\%$ and $80 \pm 10\%$ remain connected to the mission, and 96%, $98 \pm 1\%$ and $90 \pm 7\%$ remain retrievable, with mean incident weight rising to 0.777 ± 0.057, 0.780 ± 0.076 and 0.710 ± 0.026, so consolidation removes edges without losing early knowledge.

4.6 Cross-domain missions and adaptive control

On the physics mission the connectome arm runs 24.7 experiments against 6.3 for the notebook at the same cycle count and falsifies 7.0 against 3.0; on biology it runs 12.0 against 5.3 with 3.0 supports against 0.7. Hypotheses come from the frontier of the store, so a traversable store exposes more testable claims.

Cycles are matched but compute is not: the notebook loop costs 0.9 minutes per cycle against 10.9, so per wall-clock hour it executes more experiments, while it stops supplying testable hypotheses by cycle 4 of 20 and the connectome arm keeps producing them (Appendix H reports both readings).

On information theory the connectome arm more than halves its error count, 3.0 to 1.3, on the same budget; the notebook produces more attempts there, 4.7 against 3.3, because runnable designs are scarce and a third of its attempts do not run.

Scaling, working memory, control. Holding the mission fixed at biology, executed experiments rise with backbone size: 3 with a 7B model, 8 with 14B, 11 with 32B, and the verdict mix shifts

Table 5: Sandbox output per mission at matched cycle counts, three seeds per cell. The connectome arm executes 3.9 times as many experiments as the notebook in physics and 2.3 times as many in biology; on the information-theory mission the notebook produces more attempts while the connectome arm cuts its errors by more than half.

Mission	Arm	Experiments	Supports	Falsifies	Inconclusive	Errors
Physics (20 cycles)	connectome	24.7 ± 2.3	3.0 ± 1.7	7.0 ± 2.0	8.3 ± 2.5	6.3 ± 3.1
	notebook	6.3 ± 0.6	1.0 ± 0.0	3.0 ± 1.0	2.0 ± 1.7	0.3 ± 0.6
Biology (6 cycles)	connectome	12.0 ± 1.0	3.0 ± 2.0	2.7 ± 2.1	4.3 ± 0.6	2.0 ± 1.7
	notebook	5.3 ± 0.6	0.7 ± 0.6	2.7 ± 0.6	1.3 ± 0.6	0.7 ± 0.6
Information theory (6 cycles)	connectome	3.3 ± 0.6	0.3 ± 0.6	0.0 ± 0.0	1.7 ± 0.6	1.3 ± 0.6
	notebook	4.7 ± 1.2	0.3 ± 0.6	0.0 ± 0.0	1.3 ± 0.6	3.0 ± 1.0

from one support and two falsifications at 7B to five supports and two at 32B, so the larger model writes more of the kind of hypothesis the sandbox can resolve, and even the 7B local model sustains a working discovery loop. Sweeping $W_{\max}$ on a fixed substrate exposes a clear optimum (Fig. 4(b), Appendix H): at $W_{\max} = 3$ the agent thrashes, evicting on every step, and holds 170 tokens at a coherence of 0.415, while from 7 upward coherence falls monotonically to 0.221 at 100. Efficiency, coherence per unit context, peaks at the smallest non-thrashing setting and decays from there, 0.238 at 7 against 0.024 at 100, so the default of 10 trades 0.070 of efficiency for headroom. Raising dopamine from 0.1 to 0.9 cuts episodes to convergence from 13.7 to 11.0 and cumulative regret from 7.63 to 5.75; frustration acts as a stagnation detector, raising action entropy from 0.804 to 2.05 bits when it fires.

Where the generator, not the memory, is the bottleneck. DiscoveryBench asks questions whose answers require re-analyzing a dataset, a setting that exercises the generator as much as the store. Across 546 judged rows the three arms converge (0.132 connectome, 0.143 fresh store, 0.104 notebook), and the analysis places the limit in the generator. Between 60% and 86% of answers recite a memorized filler sentence committing to a number that appears in no input file. Restricting to rows with a substantive attempt, quality is comparable (0.244, 0.286, 0.351) while the connectome arm produces 2.3 times as many substantive attempts, 86 against 37.

5 Limitations and Future Work

AutoScientist is designed to measure what a structured memory contributes to a long-horizon research agent. The consolidation ablation isolates the four non-generative passes over a two-cycle replay with LLM calls disabled, while the three generative passes are evaluated over the full trajectories; pass-disabled live runs are the direct next measurement. Results use a single backbone scale, so the generator is held fixed and the memory is isolated. The study spans three missions with three seeds each and 654 scored suite rows per judge setting. The suite comparison runs against flat and dense stores and a graph-memory arm that shares our store with propagation retrieval; other structured-memory systems are positioned in §2 and are not run end to end, and those systems draw on frontier commercial backbones, so a like-for-like compute comparison remains open. The agent-level comparison matches cycles, with the wall-clock reading in Appendix H. The reward covers only mission progress, and topology emerges without it; the component ablation of Appendix B places the probe abilities in addressing, the recorded status field and the typed support links, and a structure-forcing perturbation would sharpen that attribution into a causal statement.

6 Conclusion

We described an agent whose memory is one typed, signed, weighted graph revised by a sleep–wakecycle. The memory converges toward the sparse, clustered regime that biological connectomes occupy, and its typed links make chains and opposing claims retrievable where flat text and dense retrieval cannot, at 96% of the context cost, all on a single locally hosted 32B open-weight model, while the autonomous research agents reviewed in §2 rely on frontier commercial APIs.

AI USE STATEMENT

In this work, we used generative AI tools to draft and edit the manuscript text and to write the analysis, figure-generation and table-build scripts. We have not used generative AI tools to design the AutoScientist system, to execute the reported runs, or to produce any of the reported measurements, and the remaining disclosure categories are not applicable to this work. We also used these tools for copy-editing and for suggesting alternative phrasings. We have reviewed all AI-assisted work: AI-edited passages were checked against the archived run artifacts, and every generated script was executed and its output validated against the released results before inclusion. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies an autonomous agent that writes and executes code. Every program the agent produced ran in an isolated subprocess with a fixed timeout and no network access, and the reported runs use only public benchmarks and locally hosted models; no human subjects are involved and no personal data are collected or released. All runs and their artifacts are released for inspection, which is the safeguard we consider appropriate for work on autonomous research agents: the behavior can be audited after the fact. We report no conflicts of interest.

REPRODUCIBILITY STATEMENT

Appendix C lists every node and edge type, every wake and sleep module with its thresholds, and every control constant; Appendix D gives the judging protocol, the rubric wording and the agreement statistics; and Appendix G describes the released artifacts. Code, frozen brain states, per-cycle topology logs, every judged row under both rubrics and both judge families, the consolidation replay inputs and outputs, and the scripts that regenerate every table and figure in this paper from the archived runs are released together.

REFERENCES

Danielle S. Bassett and Olaf Sporns. Network neuroscience. *Nature Neuroscience*, 20(3):353–364, 2017.

Richard F. Betzel, Andrea Avena-Koenigsberger, Joaquín Goñi, Ye He, Marcel A. de Reus, Alessandra Griffa, Petra E. Vértes, Bratislav Mišić, Jean-Philippe Thiran, Patric Hagmann, Martijn van den Heuvel, Robert J. Zatorre, and Olaf Sporns. Generative models of the human connectome. *NeuroImage*, 124:1054–1064, 2016.

Markus J. Buehler. Agentic deep graph reasoning yields self-organizing knowledge networks. *arXiv preprint arXiv:2502.13025*, 2025.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Pengfei Du. Memory for autonomous LLM agents: Mechanisms, evaluation, and emerging frontiers. *arXiv preprint arXiv:2603.07670*, 2026.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A GraphRAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Shanghua Gao, Ada Fang, and Marinka Zitnik. AutoScientists: Self-organizing agent teams for long-running scientific experimentation. *arXiv preprint arXiv:2605.28655*, 2026.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. *arXiv preprint arXiv:2405.14831*, 2024.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, 2026.

Zehao Jin, Yaoye Zhu, Chen Zhang, and Yanan Sui. Whole-brain connectomic graph model enables whole-body locomotion control in fruit fly. *arXiv preprint arXiv:2602.17997*, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aaroh Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, et al. Kosmos: An AI scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025.

Siba Smarak Panigrahi, Jovana Vidaki, et al. HeurekaBench: A benchmarking framework for AI co-scientist. In *International Conference on Learning Representations*, 2026.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Vaibhava Lakshmi Ravidesh et al. Can AI evaluate AI scientists? a benchmarking study of autonomous research generation systems using automated multi-model review. *arXiv preprint arXiv:2607.28631*, 2026.

Shanghai Artificial Intelligence Laboratory. ResearchClawBench: A benchmark for end-to-end autonomous scientific research. *arXiv preprint arXiv:2606.07591*, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*, 2025.

Yutaro Yamada, Robert Tjarko Lange, Chris Lu, Shengran Hu, Cong Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Chang Yang, Chuang Zhou, Yilin Xiao, Su Dong, Luyao Zhuang, Yujing Zhang, Zhu Wang, Zijin Hong, Zheng Yuan, et al. Graph-based agent memory: Taxonomy, techniques, and applications. *arXiv preprint arXiv:2602.05665*, 2026.

A Formal model

This appendix states the model as a sequence of operators acting on one object, so that each claim in the main text can be read off a definition, and gives the definitions the controlled probes compute.

State. After cycle t the memory is the directed labeled graph $G_t = (\mathcal{V}_t, \mathcal{E}_t, w_t)$ of Eq. 1, whose nodes and edges carry annotations

$$v = \big(s_v,\ \tau_v,\ \sigma_v,\ I_v,\ q_v,\ z_v\big), \qquad v \in \mathcal{V}_t, \tag{6}$$

$$e = \big(u,\ v,\ \ell_e,\ w_e,\ \gamma_e,\ p_e,\ x_e\big), \qquad e = (u, v) \in \mathcal{E}_t, \tag{7}$$

with s_v the natural-language statement, τ_v the node type in {MISSION, QUESTION, CONCEPT, HYPOTHESIS, EMPIRICAL, SYNTHESIS, GAP} (the schema reserves one further type, ANSWER, that the released runs never produce), σ_v the epistemic status in {settled, uncertain, contradicted, hypothetical}, $I_v \in [0, 1]$ the importance, $q_v \in [0, 1]$ the source quality and z_v a unit-norm embedding in a separate nearest-neighbor index, so that cosine similarity is the inner product $z_u^\top z_v$. On an edge, ℓ_e is the type, w_e the weight, γ_e the confidence, p_e the provenance and $x_e \in \{0, 1\}$ the decay-exemption flag. The graph is directed and simple:at most one edge per ordered pair, and a relation observed again for an existing pair updates that edge in place, adding no parallel edge.

Seven edge types occur: SUPPORTS, CAUSES, CONTRADICTS, STRUCTURALANALOGY, ASSOCIATED, TOWARDMISSION and EMPIRICALLYTESTED. Writing k for a type, the typed adjacency matrices

$$A^{(k)}_{uv} = w_e\, \mathbf{1}[\ell_e = k], \qquad e = (u, v) \in \mathcal{E}_t, \tag{8}$$

make the epistemic distinctions first-class: support, contradiction and causation are distinct coordinate systems over the same node set, so a supported claim and its refutation are present at once and separately retrievable, and a typed path is a walk required to stay inside one or more of them.

Wake operators. A wake phase is the composition $\pi_t = (\mathsf{read}, \mathsf{reason}, \mathsf{test}, \mathsf{gate})$. Reading maps documents to candidate propositions with proposed typed edges, $\mathcal{R} : \mathcal{D} \to 2^{\mathcal{V} \times \mathcal{E}}$, and every candidate passes a gate,

$$G \leftarrow G \oplus \{(v, e) : g(v, e) = 1\}, \qquad g(v, e) = \mathbf{1}[\text{the critic accepts } (v, e)], \tag{9}$$

with rejected candidates held in the insight buffer, so none is discarded. Reasoning selects the question

$$v^\star = \operatorname*{arg\,max}_{v \in \mathcal{Q}_t \cup \mathcal{G}_t} I_v, \tag{10}$$

where $\mathcal{Q}_t$ is the open-question set on the agenda and $\mathcal{G}_t$ the GAP nodes, and chooses a mode a from {analogical, dialectical, reductive, experimental, integrative} in state $s = (\tau_{v^\star}, c_{v^\star})$ (node type and cluster) by ϵ-greedy exploration of a value table $Q(s, a)$, updated from the critic's verdict r ($+1$·confidence on acceptance, -1 on rejection, 0 when deferred) as

$$Q(s, a) \leftarrow Q(s, a) + \eta_0(1 + D_t)\big[r - Q(s, a)\big], \qquad \eta_0 = 0.1. \tag{11}$$

Testing maps a hypothesis to a program, runs it and returns a verdict $v \in$ {supports, contradicts, inconclusive}: an EMPIRICAL node is attached through an EMPIRICALLYTESTED edge with γ_e equal to the run's confidence and $x_e = \mathbf{1}[\gamma_e > 0.7]$, and a contradiction verdict above 0.6 confidence adds a CONTRADICTS edge. Two modulators carry the control signal, with the mission suspended at $F \geq 0.8$:

$$D_t \to 0.5 + \operatorname{sign}(D_t - 0.5)\max\big(0, |D_t - 0.5| - 0.3\Delta t\big), \qquad F_{t+1} = \min\big(1, F_t + 0.2\,\mathbf{1}[\text{failure}]\big) \tag{12}$$

with F reset to 0.4 and $D \leq 0.4$ on suspension, and the mission resumed once decay has brought $F \leq 0.4$.

Retrieval and working memory. Candidate generation scores a node against a query embedding z_q by

$$\operatorname{sim}(v) = z_v^\top z_q, \qquad \|z_v\|_2 = \|z_q\|_2 = 1, \tag{13}$$

and a graph query is a typed traversal: for a cross-cluster concept pair (a, b) the reference path is the shortest directed walk $a = v_1 \to \cdots \to v_m = b$ over edges of any type with at most five steps. Working memory is an ordered bounded list

$$\mathrm{WM}_t = \big(v_{(1)}, \ldots, v_{(W)}\big), \qquad W \leq W_{\max} = 10, \tag{14}$$

where focusing on a node moves it to the front, so the least recently focused node leaves first. The prompt context of a step is the list together with the neighbors those nodes are connected to, so its cost is bounded by $W_{\max}$ and the degree of those neighbors, independent of the size of the store. That bound is the mechanism behind the character counts in Table 1.

Consolidation. Sleep applies a fixed composition of eight maps,

$$\mathcal{C} = \mathsf{relevance} \circ \mathsf{buffer} \circ \mathsf{decay} \circ \mathsf{contradiction} \circ \mathsf{gap} \circ \mathsf{abstraction} \circ \mathsf{synthesis} \circ \mathsf{merge}, \quad (15)$$

applied to G_t in the order written: (1) merge contracts node pairs with $z_u^\top z_v \geq 0.88$ of equal type outside $\{$EMPIRICAL, MISSION$\}$, rewiring incident edges and deleting the weaker node; (2) synthesis rolls a sample of six recent nodes into a SYNTHESIS hub when the sample implies a mechanism none of them states; (3) abstraction promotes recurring patterns inside clusters of at least three nodes; (4) gap detection inserts a GAP node into a topological void between established facts (confidence 0.75, deduplicated at cosine 0.75); (5) contradiction update surfaces CONTRADICTS pairs and raises I_v by 0.05 at both endpoints; (6) decay multiplies non-exempt weights,

$$w_e \leftarrow \max\bigl(0.01,\ w_e\, 2^{-\Delta t/\tau}\bigr), \qquad \tau = 1/\mathsf{decay_rate} = 100, \quad (16)$$

with Δt the time elapsed since the previous consolidation, and severs edges whose weight has fallen below 0.02; (7) the insight buffer re-evaluates deferred candidates and promotes those that have since gained support; (8) mission relevance links nodes of importance above 0.6 to the mission when the assessed strength exceeds 0.4. The exemption set is explicit,

$$\begin{aligned}\{e : \ell_e = \text{CONTRADICTS}\} \cup \{e : \ell_e = \text{TOWARDMISSION}\} \\ \cup \{e : \ell_e = \text{EMPIRICALLYTESTED},\ \gamma_e > 0.7\},\end{aligned} \quad (17)$$

so a refutation is never lost to decay while unconfirmed associative and supporting structure is. Node quality decays only for items unverified for more than three days, $q_v \leftarrow \max(0.05, q_v - r\,\Delta t\, d_v)$ with $r = 0.02$ (0.04 for sleep-produced items) and d_v the number of stale days; no node in the released runs met that condition, so every measured effect of decay is carried by the edge law above. One cycle is the composition

$$G_{t+1} = \mathcal{C}\bigl(\pi(G_t)\bigr). \quad (18)$$

Probes. The controlled probes of §4.2 are defined on a reference path $\pi = (v_1, \ldots, v_m)$ and the returned set R:

$$\text{target found} = \mathbf{1}[v_m \in R], \quad (19)$$

$$\text{interior recall} = \frac{\#\bigl[(\pi \setminus \{v_1, v_m\}) \cap R\bigr]}{\#\bigl[\pi \setminus \{v_1, v_m\}\bigr]}, \quad (20)$$

$$\text{step continuity} = \frac{\#\bigl\{(v_i, v_{i+1}) : v_i, v_{i+1} \in R\bigr\}}{m-1}, \quad (21)$$

$$\text{decoy rate} = \frac{\#\bigl\{r \in R \setminus \pi : \text{no typed path } r \to v_m \text{ of length} \leq 4\bigr\}}{\#R}. \quad (22)$$

Step continuity counts the consecutive pairs of the reference chain that retrieval returns together, interior recall counts the intermediate items on their own, and the decoy set is exactly the vocabulary-matched but unsupported content that a similarity ranker cannot separate from a supported claim.

B WHAT CARRIES EACH PROBE

The probe table fixes the substrate and varies the retrieval rule. This appendix gives the paired tests behind its numbers and then fixes the rule and varies the structure underneath it, on the nine archived final brains of §4.2: the probe set, the embedding front-end and the answering rule are held constant, and one property of the stored graph is perturbed before scoring.

Paired tests on the probe table. Every substrate answers every query, so the comparison is paired on the query and on the brain (286 query pairs on the retrieval probes, 114 on the factual and ordering competency rows). Against dense retrieval the graph's chain advantage is $+0.084$ on target found (95% interval $[+0.052, +0.115]$, exact McNemar $p < 10^{-4}$), $+0.607$ on step continuity $[+0.558, +0.657]$, $+0.560$ on interior recall $[+0.510, +0.611]$, and -0.416 on the decoy rate $[-0.440, -0.393]$. Recorded order is a ceiling-versus-floor gap ($+1.000$, $p < 10^{-4}$) because an undated corpus cannot answer it at all, and factual recall is $+0.000$ with $p = 1.00$, the control that

every substrate passes. On the $n = 13$ resolved contradictions the graph is $+0.846$ over dense retrieval ($p = 0.001$) and $+0.308$ over the flat store, which does not reach significance at this sample size ($p = 0.125$) because that row is bimodal across memories. Synthesis support recall favours the graph by $+0.583$ over dense retrieval ($[+0.483, +0.667]$). Probes are built from the pristine brain, so the gold labels never see the perturbation, and every perturbation keeps the node set fixed.

The perturbations are: *labels permuted*, the edge type labels reallocated over the same edges (identical topology, weights and degree sequence); *labels uniform*, every edge relabelled ASSOCIATED, so no typing remains; *contradiction edges deleted*; *contradiction edges reversed*, so the signed direction is inverted; *statuses permuted*, the node status values reallocated over the same nodes (identical status histogram); and *degree-preserving rewiring*, double-edge swaps that keep every node's in- and out-degree while removing the typed neighbourhood structure.

Table 6: Component ablation of the graph under a fixed answering rule, pooled over the nine archived brains. Columns are the competency probes of Table 1; each cell is the fraction of probes resolved. The contradiction and synthesis columns exist only on the six and two brains that hold such structure, and their counts are printed with the column.

Graph perturbation	Factual ($n = 180$)	Order ($n = 180$)	Opposing pair ($n = 13$)	Synthesis ($n = 10$)
None (as stored)	1.000	1.000	1.000	1.000
Edge labels permuted	1.000	1.000	1.000	1.000
Edge labels uniform	1.000	1.000	1.000	0.800
Contradiction edges deleted	1.000	1.000	1.000	1.000
Contradiction edges reversed	1.000	1.000	1.000	1.000
Node statuses permuted	1.000	1.000	0.077	1.000
Degree-preserving rewiring	1.000	1.000	1.000	0.800

Three things follow. The first two are properties of the probes themselves, and the third speaks to the agent's memory. Factual recall and recorded order stay at 1.000 under every perturbation, including full rewiring: the first is answered by the embedding address that fronts the store, which returns the stored item itself, and the second by the "created_at" attribute on the endpoints, which no edge perturbation touches. Neither row therefore separates memory architectures by structure; they separate them by whether the substrate can address an item and keep a timestamp at all, which is why the flat store fails the ordering row outright.

The contradiction row is carried by the status field. Permuting the recorded statuses across nodes drops it to 0.077, one probe in thirteen, while deleting or reversing the contradiction edges leaves it at 1.000: the substrate is not re-deriving the verdict from the edge, it is reading the status each endpoint carries. That status is written when the node is admitted, and the sleep pass that revisits the pair does not set it: a sandbox verdict above 0.7 confidence settles its empirical node, an insight above 0.7 strength settles an uncertain endpoint it restructures, and everything else stays uncertain. The thirteen resolved pairs of these nine runs split eleven settled empirical nodes against uncertain hypotheses and two settled concepts against uncertain concepts. The probe is therefore a retention-and-addressing measurement, what a substrate can address from a paraphrase and keep, and it explains why contradiction_update, which elevates the importance of nodes on contradiction edges, moves no structural statistic in Table 3.

The synthesis row does respond to the structure it names: relabelling every edge uniform and rewiring the endpoints each cost 0.2 of the row, because both break the SUPPORTS links from evidence to the synthesis node that the probe expects. The two brains that hold enough synthesis nodes to build the probe make this the weakest cell in the table, and it is reported as such.

C SYSTEM COMPONENTS AND HYPERPARAMETERS

This appendix documents every component referenced in §3.2 and §3.3: the full node and edge schema, the wake and sleep modules with their thresholds, and the control constants. All values are as released in the code.

Node schema. Each node $v \in \mathcal{V}$ carries:

- *Statement*: the natural-language proposition.
- *Type* $\in$ {MISSION, QUESTION, CONCEPT, HYPOTHESIS, EMPIRICAL, SYNTHESIS, GAP}.
- *Epistemic status* $\in$ {settled, uncertain, contradicted, hypothetical}: nodes are created uncertain, sleep-produced structure is hypothetical, a sandbox verdict with confidence above 0.7 settles its empirical node, and an insight with strength above 0.7 settles an uncertain endpoint it restructures.
- *Importance* $I_v \in [0, 1]$, used for agenda selection and working-memory priority.
- *Source quality* $q_v \in [0, 1]$: 0.9 for literature, 0.8 for research output, 0.7 for conversation and sandbox results, 0.6 for consolidation and 0.3 for a speculation produced during sleep.
- *Embedding*: a 384-dimensional sentence-transformer vector, mirrored into a nearest-neighbor index for candidate generation.

Edge schema. Each directed edge $e = (u, v)$ carries one of seven types: CAUSES (directed mechanistic dependency), SUPPORTS (evidential reinforcement), CONTRADICTS (mutually exclusive claims or a falsification), STRUCTURALANALOGY (isomorphic structure across distant clusters), EMPIRICALLYTESTED (hypothesis to sandbox result), TOWARDMISSION (relevance projection to the objective), and ASSOCIATED (contextual co-occurrence). Those seven are the types the released runs produce; the schema also carries four analogue sub-types (surface analogy, structural analogy, deep isomorphism and a legacy analogous-to) of which only structural analogy is used. Every edge also carries a weight $w_e \in [0, 1]$, a confidence, and a provenance label in {LITERATURE, DREAM, THINKING, SANDBOX, CONSOLIDATION}.

Wake modules. The *reader/researcher* pair queries external sources, extracts propositions and causal statements, and projects them as CONCEPT nodes, validating each new edge against the existing store before insertion. The *thinker* selects an open question by importance, chooses among five reasoning modes (reductive, analogical, dialectical, experimental, integrative) with an ϵ-greedy contextual bandit ($\epsilon = 0.2$), and emits structured hypotheses. The *sandbox* decides whether a hypothesis is computationally testable, writes self-contained Python (ODE integration, Gillespie simulation, Monte Carlo), executes it in an isolated subprocess with a 30 s timeout, parses the numerical output, renders a figure, and assigns a verdict of *supports*, *contradicts* or *inconclusive*. The *critic* scrutinizes every insert for circular definitions, logical fallacies and unjustified extrapolation; insights that fail are held in an *insight buffer* and re-evaluated on later cycles, so a speculative idea is neither inserted nor lost. The *notebook* records morning intentions, experiment logs and the evening consolidation summary; the *observer* tracks mission progress.

Sleep passes. In execution order, with the constants used throughout: (1) merge_duplicates, merging node pairs above cosine similarity 0.88 with compatible topology and rewiring incident edges; (2) synthesis_pass, inducing a SYNTHESIS hub over a dense cross-cluster group; (3) abstraction_pass, promoting recurring patterns to higher-level nodes; (4) gap_detection, inserting a GAP node where a path between established facts has a topological void; (5) contradiction_update, surfacing CONTRADICTS pairs and escalating unresolved ones to the top of the agenda; (6) apply_decay, applying $w_e \leftarrow \max(0.01, w_e\, 2^{-\Delta t/\tau})$ with $\tau = 1/\text{decay_rate} = 100$ and severing edges below $w_e = 0.02$; edges carrying a contradiction, a mission link or a sandbox result above 0.7 confidence are decay-exempt (Eq. 17), while node quality decays only after three days unverified; (7) evaluate_insight_buffer, promoting buffered insights that have since gained support; (8) recompute_mission_relevance, refreshing TOWARDMISSION scores.

Control constants. Working memory $W_{\max} = 10$; bandit exploration $\epsilon = 0.2$; frustration increment $+0.2$ per failure; suspension threshold $F \geq 0.8$; post-suspension $D = 0.4$, $F = 0.4$; resumption when $F \leq 0.4$; sandbox timeout 30 s; embedding dimension 384; duplicate-merge threshold 0.88; gap confidence and gap deduplication 0.75; bandit learning rate 0.1; abstraction minimum cluster 3; synthesis sample 6; edge-weight floor 0.01 with severance below 0.02.

D JUDGING PROTOCOL AND AGREEMENT

The strict rubric awards one point only when the answer matches the gold answer, with adversarial items requiring an explicit statement that the question cannot be answered. The lenient rubric, taken from the published LoCoMo evaluation prompt, asks the judge to be generous about paraphrase. Every one of the 654 scored rows was judged under both rubrics by Gemma3-27B and under the strict rubric by Mixtral-8x7B. Table 7 reports the agreement.

Table 7: Agreement between the two judge families on identical model outputs under the strict rubric. Unparseable outputs: zero.

Suite	Rows	Raw agreement	Cohen's κ
LongMemEval-S	360	0.931	0.860
LoCoMo-10	294	0.803	0.598
Pooled	654	0.873	—

The hand audit covered 50 stratified rows chosen by a fixed seed, spanning both suites, all three arms and every question category, plus five rows where the two judges disagreed. The auditor agreed with the strict judge on 48 of 50 rows (0.960, Wilson interval $[0.865, 0.989]$, $\kappa = 0.92$). One disagreement withheld credit for an answer that stated a total in two components where the gold was a single summed figure; the other granted partial credit for two of three required items. Both corrections fall inside one standard error of the affected arm and act in opposite directions. On the five stress rows the human sided with the second judge three times, twice on adversarial items where the strict rubric penalized an answer that declined the question and then added unrequested context.

E PER-ABILITY BREAKDOWN

Tables 8 and 9 give the per-category breakdowns discussed in §4.3.

Table 8: LongMemEval-S by ability, 20 questions each, strict rubric.

LongMemEval-S ability	Connectome	Iterative dense	Flat store
Knowledge update	0.700	0.750	0.400
Multi-session	**0.300**	0.200	0.050
Single-session assistant	0.950	0.950	0.250
Single-session preference	0.250	0.300	0.100
Single-session user	0.850	0.800	0.250
Temporal reasoning	0.250	0.350	0.050

Table 9: LoCoMo-10 by category, strict rubric, 18 to 20 questions each. The adversarial row favors the flat store because an unanswerable question is answered correctly by an empty retrieval.

LoCoMo-10 category	Connectome	Iterative dense	Flat store
Adversarial (unanswerable)	0.200	0.200	0.600
Multi-hop	0.400	0.400	0.250
Open-domain	0.167	0.167	0.167
Single-hop	0.600	0.800	0.200
Temporal reasoning	**0.500**	0.400	0.000

F STRUCTURED-MEMORY BASELINE

The objection a reviewer would raise first is that the store is fine and the retrieval rule is doing the work, or the reverse. This appendix runs the arm that separates them: same item set, same graph as the store, same context window, same answer and judge models, but retrieval by personalized PageRank restarted on the embedding-nearest items over a similarity graph, the propagation rule at

the heart of HippoRAG-style retrieval (Gutiérrez et al., 2024). Everything is held fixed except the rule that decides what reaches the window.

The propagation arm sits inside the band the other arms occupy on every cell, so the PageRank column of Table 2 reproduces the picture without separating the designs: on LongMemEval-S it scores 0.533 and 0.592 against 0.550/0.567 for the graph and 0.558/0.608 for dense retrieval, and on LoCoMo-10 0.337 and 0.561 against 0.378/0.582 and 0.398/0.531. Paired on the question it differs from the graph by $+0.017$ (McNemar $p = 0.81$) and from dense retrieval by -0.025 ($p = 0.65$) on LongMemEval-S, and by $+0.041$ ($p = 0.39$) and -0.061 ($p = 0.24$) on LoCoMo-10, so all four intervals contain zero. It clears the flat store where the other arms clear it: $p < 10^{-8}$ on LongMemEval-S under both judges, and on LoCoMo-10 $p < 10^{-5}$ under the second judge against $p = 0.21$ under the primary, which is the flat store's adversarial category answering itself by retrieving nothing. Excluding that category it reads 0.397 against the graph's 0.423. Window cost is the same, 59.9k characters against the graph's 57.7k, so propagation over the store's own similarity graph reproduces the same picture: general long-memory scores do not separate these memory designs, while the probes of §4.2 do.

G Artifact release and regeneration

Seeds, run directories, frozen brain states, per-cycle topology logs, every judged row under both rubrics and both judge families (654 for the three main arms, plus the structured-memory baseline arm of Appendix F), the consolidation replay inputs and outputs, and the scripts that regenerate every table and figure in this paper from the archived run data are released together, each table listing the driver that produces it: `suite_significance.py` and `probe_significance.py` for the paired tests of Appendix B and §4.3, `typed_edge_ablation.py` for the component ablation, `effort_accounting.py` and `decay_timestep_sensitivity.py` for Appendix H.

H Effort accounting and decay timestep

Two readings of the same campaigns, so the agent-level comparison can be taken at matched cycles or at matched wall-clock time. Cycles are the designed unit of the diurnal loop; wall-clock is what a cheaper loop saves.

Table 10: Effort accounting for the two agent arms, means over three seeds, from the per-cycle timestamps and sandbox logs of the archived runs. Cycles are matched by construction; the notebook loop is an order of magnitude cheaper per cycle, so the last column reverses the throughput reading.

Mission	Arm	Cycles	Experiments	min/cycle	Wall-clock	Exp/cycle	Exp/hour
Physics (20)	connectome	20	24.7 ± 2.3	10.9	3.62 h	1.23	6.8
	notebook	20	6.3 ± 0.6	0.9	0.29 h	0.32	21.7
Biology (6)	connectome	6	12.0 ± 1.0	18.0	1.80 h	2.00	6.7
	notebook	6	5.3 ± 0.6	3.2	0.31 h	0.89	17.0
Information theory (6)	connectome	6	3.3 ± 0.6	10.6	1.06 h	0.56	3.2
	notebook	6	4.7 ± 1.2	2.7	0.27 h	0.78	17.3

Read per cycle, the connectome arm executes 3.9 times as many experiments as the notebook in physics and 2.3 times as many in biology; read per wall-clock hour, the cheaper loop executes more. Both readings are true, and the reports in the main text keep them apart. The mechanism behind the cycle-level difference is visible in the same logs: the notebook's experiment count stops at 6 to 7 in physics by cycle 4 and stays there for the remaining sixteen cycles while its scratchpad grows from 4.0k to 36.1k characters, and its counts on the other two missions plateau by cycle 4 as well, whereas the connectome arm keeps adding empirical nodes through cycle 15 to 16. The notebook does not stop working; it stops producing hypotheses that its sandbox can run, which is the effect a typed store is meant to have on the supply of testable claims.

Decay timestep. The decay pass reads its timestep from the interval since the previous consolidation, falling back to one day when no timestamp exists: the designed reading is one consolidation

per simulated day, while the live runs executed their cycles back to back. Table 11 repeats the decay contrast of Table 3 under both schedules. Clustering and edge-count differences agree to two decimals, and mean weight to 0.003, so the attribution of the structural effect to decay does not rest on the timestep. These arms replay offset 0 of three seeds; Table 3 aggregates four offsets.

Table 11: Decay timestep sensitivity: disabling apply_decay in the replay, under the designed one-cycle-per-day schedule and under the wall-clock cadence the live run actually had (median gap between cycles in brackets). Differences are reported against the full-sleep replay of the same snapshot.

Replay	Timing	Δ clustering	Δ mean weight	Δ edges
mission3_s101 (16.5 min)	designed	−0.0247	−0.0135	+1
	wall-clock	−0.0247	−0.0160	+1
mission3_s202 (6.8 min)	designed	0.0000	+0.0026	0
	wall-clock	0.0000	0.0000	0
mission3_s303 (9.3 min)	designed	−0.0186	−0.0128	+1
	wall-clock	−0.0186	−0.0153	+1
Mean of the three seeds	designed	−0.0144	−0.0079	+0.67
	wall-clock	−0.0144	−0.0104	+0.67

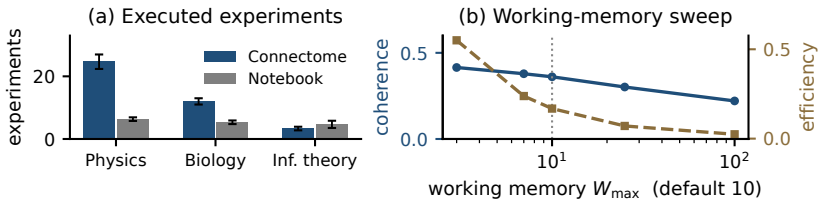


Figure 4: Executed experiments per mission (Table 5) and the working-memory sweep on a fixed substrate: thrashing disappears above $W_{\max} = 3$ and efficiency decays monotonically afterwards, so the default of 10 sits one step past the knee. The two series are on separate axes.