

Era Brain: Evaluating a Permission-Aware Business Memory on Retrieval, Multi-hop Records and Access Control

Era Research

October 2026 · Era Technical Report · commit 3aba0bbfb · Preprint v2

Abstract

Era Brain is the memory layer of Era, a platform that runs a business's records and conversations. It keeps synced records (customers, orders, invoices, payments) and facts drawn from conversations in one PostgreSQL database, and answers a query with hybrid keyword and vector recall plus a question-guided path search over typed relations. No language model is called while reading; the only model on the read path is a local embedding model. We evaluate retrieval on two public conversational-memory benchmarks, LongMemEval-S [1] and LoCoMo [2], answer accuracy on LoCoMo with an LLM judge, three generated business datasets, and a permission test with planted secrets. At least one evidence session is in the top 10 for 98.3% of 470 answerable LongMemEval-S questions, and all of them for 90.2%. LoCoMo answer accuracy is 76.1%, against 87.3% for the same model reading the full conversation. On a held-out business set the answer reaches the context for 88.9% of 325 questions; on a 100,000-row graph, path search answers 100/100 multi-hop questions where plain vector top-10 answers 1/100. Across 12,560 answers to restricted members, no planted secret leaked. Median recall time is 29–35 ms in process, with no network hop. The business datasets are synthetic and self-built; Section 6 lists what this evaluation does not show.

Table 1: Summary of results. Six headline numbers with their conditions. Values marked pending come from runs that have not finished.

LoCoMo answer accuracy 76.1% LLM judge · full context, same model: 87.3%	LongMemEval-S retrieval 98.3% ≥1 evidence session in top 10 · all sessions: 90.2% · 470 q	Held-out business questions 88.9% answer in context · 325 q (245 TR, 80 EN)
Multi-hop, 100k graph rows 100/100 path search · plain vector top-10: 1/100	Permission leaks 0 12,560 member answers · 7 read paths + 8 no-person callers	Recall latency p50 29–35 ms LoCoMo / LongMemEval · in process, no network hop

1 Introduction

Memory layers for language-model agents are usually evaluated on long chat histories: one user talks to an assistant for weeks, and the system must recall what was said [1, 2]. A business asks different questions. The answer to “has the invoice for this customer's last order been paid?” sits in three synced records joined by relations, may have changed since yesterday, and must not reach a sales assistant who is not allowed to see invoices. The language is often Turkish, whose suffixes defeat plain keyword search.

Public benchmarks for this setting are scarce. Agent benchmarks over CRM data [3, 4] and a 2026 benchmark for access control in shared agent memory [5] exist, but none measures a memory layer's retrieval over synced business records under per-person permissions. We therefore report two kinds of evidence: standard retrieval and answer-accuracy numbers on public conversational benchmarks, so that Era Brain can be placed next to published systems, and self-built business and permission tests, whose construction and limits are described in Sections 3 and 6.

This report makes four contributions:

1. a description of Era Brain's write, store and read paths, with each claim traced to source code (Appendix A);

2. retrieval and answer-accuracy results on LongMemEval-S and LoCoMo, reported with the benchmarks' own question counts and metric names;
3. a business-memory evaluation in Turkish and English: question types, multi-hop record joins at 100,000 graph rows, latest-value questions and context size;
4. a permission test with 865 planted secrets run through every member-facing read path, and a capability comparison with six memory systems as documented on 22 October 2026.

2 System

Era Brain has a write path, a store and a read path (Figure 1). The write path turns synced records and conversations into facts, entities and typed relations. The store is one PostgreSQL database. The read path answers a query without calling a language model. Background jobs refine the graph; this evaluation did not exercise them.

2.1 Architecture

Records from connected apps are written without a language model. Each record is tagged with the module it came from (sales, finance, HR, ...), versioned from its own date, and written once even when a sync is retried: the memory, its links, edges, version and sync ledger share one locked transaction, and entities are upserted idempotently before it. Conversations go through LLM fact extraction. Extracted facts link to people, companies, products and topics, not to synced records; an order number mentioned in a chat is matched to the record only at query time, through text.

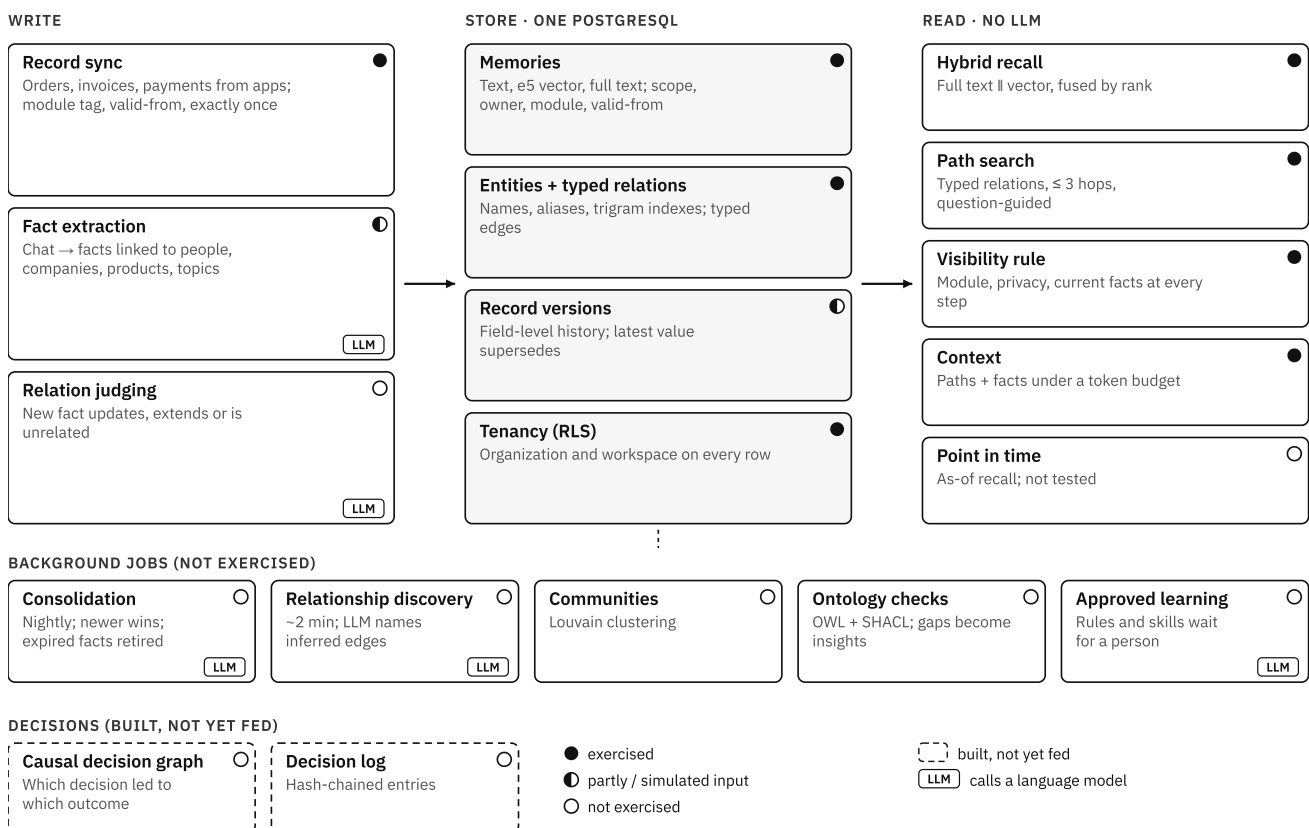


Figure 1: Era Brain components by layer. Filled marker: exercised by this evaluation; half: partly, or with simulated input; empty: not exercised. Dashed outline: built but not yet fed with production data. LLM: the component calls a language model.

Two components exist in code but receive no production data yet: the causal decision graph and its hash-chained log¹.

¹ Each log entry is chained to the previous one with SHA-256, so editing or removing an earlier entry breaks the chain. Removing the newest entries, or rewriting the whole chain, cannot be detected without an external anchor.

Nightly consolidation also has a known gap².

2.2 Read path

A query is answered in six steps (Figure 2), inside one read transaction:

- 1 Anchors.** Names in the question are matched exactly first, with Turkish case endings removed (“Ağaoğlunun” → Ağaoğlu), then by word containment and trigram similarity, then by meaning (entities behind the nearest memories). Kind words (“order”, “fatura”, “talebindeki” → talep, with consonant softening) set the kind of node to reach.
- 2 Hybrid recall.** A full-text query, with English and Turkish stop words removed, and a vector query (HNSW, cosine) run side by side; their results are fused by reciprocal rank.
- 3 Path search.** From the anchors, a walk over typed relations goes up to three hops. At each level the 48 most promising nodes walk on (at most 600 in total). A relation that fans out to more than 50 neighbours is not followed unless the question names it. The walk stops once the asked relations and kinds are covered.
- 4 Rank.** Graph facts are scored by vector distance to the question and by how many of its word stems they contain; period words (“geçen ay”, “last week”) favour facts from that period. When the same attribute is stated twice, the newest comes first and the older one is marked as earlier.
- 5 Visibility.** Every statement carries the asker's visibility rule (Section 2.3): a node the person may not see is neither returned nor walked through.
- 6 Compact.** Anchors, answer paths with their evidence, and the best remaining facts are written under a token budget (about 1,000 tokens in chat, estimated at three characters per token).

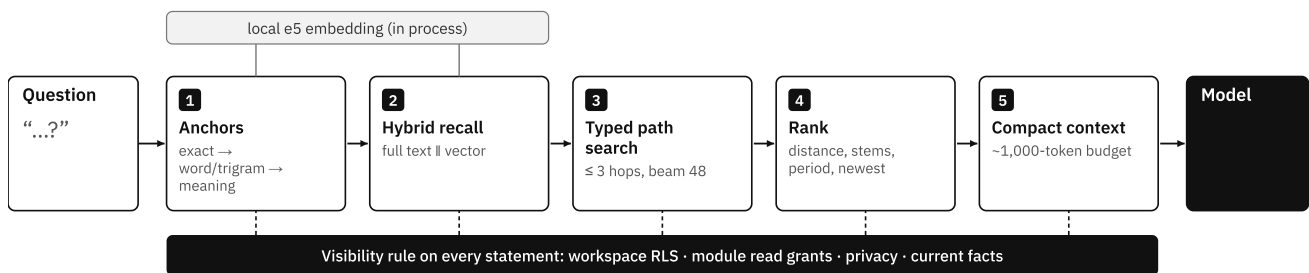


Figure 2: Read path for one query. The visibility rule is applied inside every step, not as a final filter. No language model is called; the only model is the local query embedding.

“Has the invoice for Kaya Tekstil's last order been paid?”

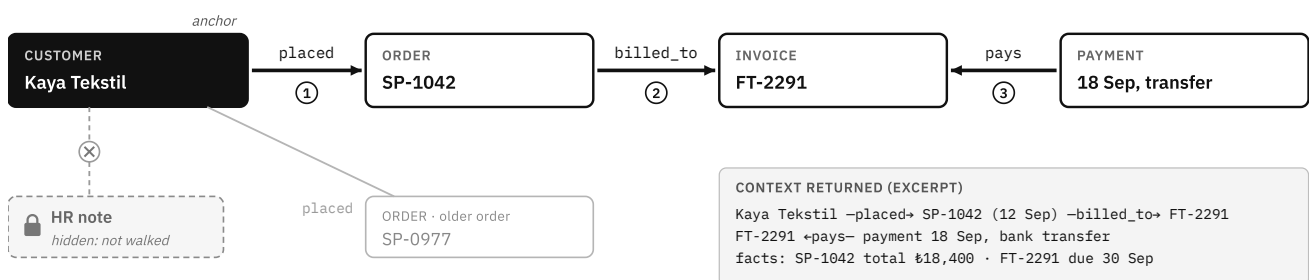


Figure 3: Worked example with illustrative data. The question names a customer and the kinds “order”, “invoice” and “paid”. The walk follows placed → billed_to → pays from the anchor and ignores the other branches; an HR note linked to the same customer is invisible to this member and is never traversed. Relation names follow the benchmark data; a live app's connector names them differently (Section 6).

2.3 Visibility rule

Access is enforced in two layers (Figure 4). PostgreSQL row-level security binds every connection to one organization and workspace.

² Consolidation runs at 03:00 UTC, and only for organizations with an active connector row; organizations that only sync records are not consolidated. Facts it derives are not yet tagged with module, workspace or privacy, so the permission result in Section 4.6 does not cover them.

On top of it, one shared visibility rule is added to every read query: facts from a module the member cannot read (resolved from the platform's role grants), other people's private or personal facts, and superseded facts are excluded. An entity named only by hidden facts is not a node, and an edge is followed only if its module is readable and both ends are visible. If the member's grants cannot be resolved, the rule falls back to reading no module. A product caller with no person, such as a workflow or a research job, reads only the workspace's shared facts. Module and privacy are query predicates, not row-level-security policies³.

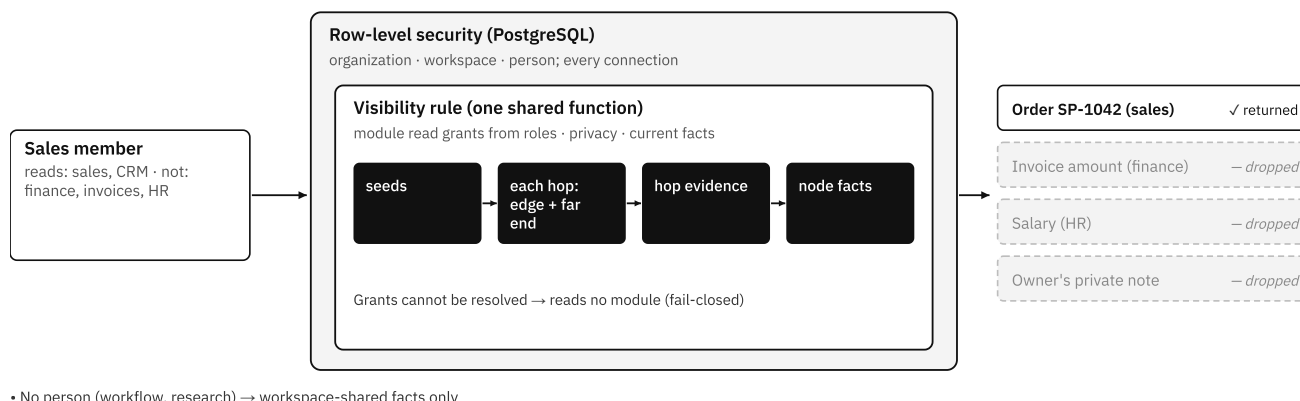


Figure 4: Two layers of access control. Row-level security scopes each connection to the organization and workspace; the visibility rule removes what this member may not read at every step of the read path, and fails closed.

2.4 What one search call touches

Figure 5 counts the services one search call touches. Era Brain runs the query embedding in process and sends about ten SQL statements to one PostgreSQL database in one transaction. Several current systems also avoid a language model on their default read path, among them Mem0 v3, Graphiti and Zep, and Suprememory [6, 7, 8, 9]. The difference in Figure 5 is therefore the number of services and network hops, not the absence of an LLM alone. In production the studio reaches the embedding model over one internal HTTP call; the benchmark ran it in process.

		services	LLM calls
Era Brain	App → PostgreSQL e5 embed in process · ~10 SQL · 1 transaction	1	0
Mem0 OSS v3	App → Embedder API unless local → Vector store + entity store	1–2	0
Mem0g (2025 paper)	App → LLM: entities from query → Embedder → Vector store → Neo4j	4	1
Graphiti (self-hosted)	App → Embedder API → Graph DB	2	0
Zep (hosted)	App → Zep API internals not published	1 + ?	?
Suprememory (hosted)	App → Suprememory API internals not documented	1 + ?	0
Letta	App → Agent LLM → Archival search	2	≥ 1

Figure 5: Services touched by one search call, from each system's documentation and code as of 22 October 2026. Services: separate systems the call reaches, including the database. Mem0g is the graph variant described in the 2025 paper [10]; it is no longer in Mem0's open-source package. Letta's agent decides each search with an LLM.

3 Evaluation setup

Table 2 lists the datasets. All runs used the code of commit 3aba0bbfb unless a footnote says otherwise.

³ The engine's library functions run unfiltered when called with no viewer at all; every product caller passes one (Appendix A).

Table 2: Datasets. The business sets were generated by us. H2 is a reworded variant (new seed, record texts and question wording) of H, a set the engine was tuned on; no engine code changed after H2 was generated.

Dataset	Language	Questions	Content	Origin
LongMemEval-S, cleaned [1]	EN	500 (470 scored)	~115k-token chat history per question	public · MIT
LoCoMo-10 [2]	EN	1,540	10 long conversations; adversarial category 5 excluded, as in [10]	public · CC BY-NC 4.0, evaluation only
Business set H2 (held out)	TR + EN (245 / 80)	325	customers, orders, invoices, payments, shipments, tickets, contracts, WhatsApp and e-mail threads	generated · product write paths
Business sets A and H	TR + EN	370 / 365	same family; used while tuning; permission sweeps with 865 and 790 secrets	generated
Scale set	EN	200 (100 multi-hop)	20,000 entities (12,000 people, 5,000 companies, products, projects, places), 40,000 memories, 40,000 typed relations	generated · bulk-loaded

3.1 Metrics

Evidence in top 10 (any / all). Whether at least one, or every, gold evidence item is among the 10 returned items: LongMemEval sessions or rounds, LoCoMo dialogue turns. LongMemEval's official retrieval metric is recall_all [1]; we report both. The 30 abstention questions have no evidence and are excluded, as in LongMemEval's own retrieval scoring. Four LoCoMo questions have no evidence label and count as misses.

Answer in context. On the business sets: the gold answer string appears in the returned context (case-insensitive, Turkish locale). It does not prove that the answer came from the right evidence; on the 100k three-hop questions the source facts for every hop are present in 5/40.

QA accuracy. A reader model answers each LoCoMo question from the returned context, and an LLM judge grades the answer against the gold answer, following [10]. Reader: claude-haiku-5-5; judge: claude-haiku-5-5.

Tokens. cl100k_base count of the context text as returned; memory lists are rendered one line per memory.

Latency. Median (p50) wall time of one recall call in the benchmark process, including the local query embedding and all queries to PostgreSQL in Docker on the same machine (Windows 11, 24 cores); one client, warm, first query excluded, no remote network.

3.2 What was not exercised

On LongMemEval and LoCoMo the conversations were stored as raw text without fact extraction, so only hybrid recall ran and the graph was not used. LLM extraction, relation judging, consolidation and relationship discovery were switched off to keep model cost at zero; their effect is not measured. In the business sets, records and relations were generated directly in the writer's input format rather than produced by a live app's connector mapping. Repeated runs differ by about one question per set, and latencies vary between runs⁴.

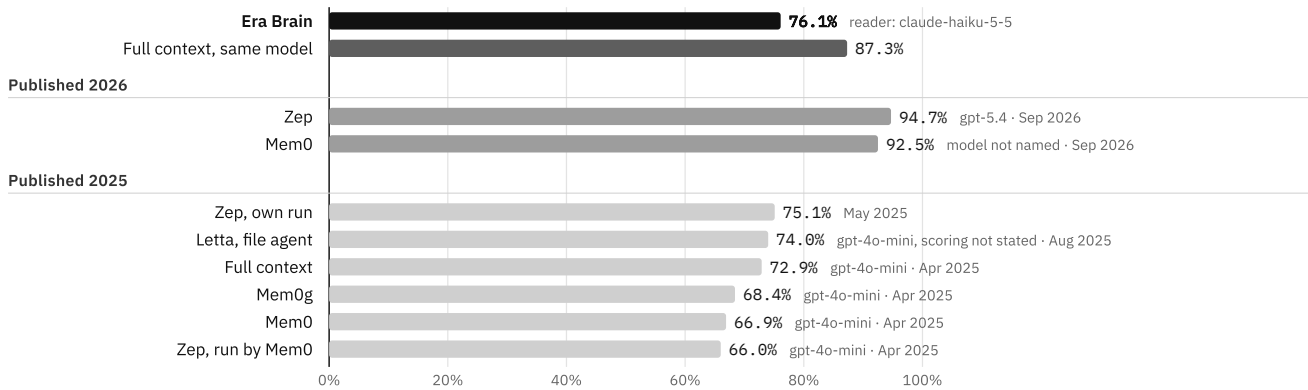
4 Results

⁴ LoCoMo and LongMemEval latency was measured before a final speed-only change to memory recall; LoCoMo p50 was 21 ms on an earlier code version without the visibility rule. The 100k memory-search latency (30 ms) comes from a separate run on the committed code. The chat-context path was swept for leaks with a 700-token graph budget; the committed budget of 1,000 tokens was leak-checked on the targeted permission questions (0 leaks in 490 answers).

4.1 Answer accuracy

Figure 6 sets Era Brain's LoCoMo answer accuracy next to a same-model baseline that reads the full conversation, and next to published figures. Published figures use different reader and judge models and were not re-run by us; they are grouped by date. Mem0's and Zep's 2026 results [11, 8] use newer readers than their 2025 papers [10, 12]. The 2025 Zep figure of 66.0% was produced by Mem0's team and is disputed by Zep, whose own run reports 75.1% [13]. LoCoMo's labels contain errors that affect every judge-based score⁵.

(a) LoCoMo, 1,540 questions



(b) LongMemEval-S

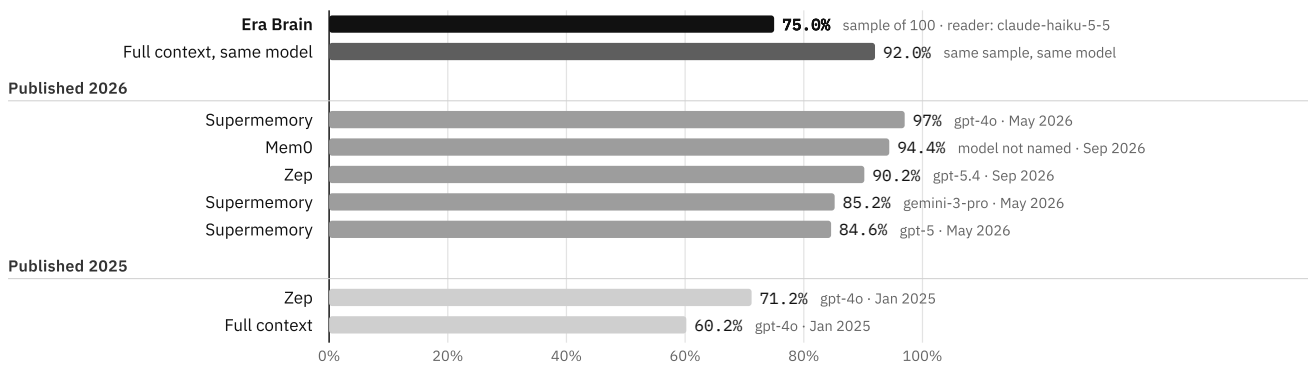


Figure 6: Answer accuracy with an LLM judge. Black: Era Brain; dark grey: same-model full-context baseline; grey: published by the paper or vendor named, with reader model and date. Values from different readers and judges are not directly comparable. Dashed: pending.

4.2 Retrieval

Figure 7 reports retrieval on the public benchmarks. On LongMemEval-S, hybrid recall puts at least one evidence session in the top 10 for 98.3% of questions, and every evidence session for 90.2%. A plain vector top-10 already reaches 97.9% on the “any” measure, so the benchmark leaves little room there; the “all” measures separate methods more. On LoCoMo, hybrid recall finds at least one evidence turn for 72.4% against 62.6% for plain vector, with about 596 tokens of context.

4.3 Business memory

Figure 8 breaks the held-out set down by question type. Direct lookups, two-hop questions (person → employer → tax number, IBAN or website; ticket → assigned staff phone; product → supplier phone), period questions and latest-value questions are answered almost completely. Two types remain weak: questions that join a record with a conversation (an order is named, the answer is in the customer's WhatsApp or e-mail) at 37/60, and the complaint → order → invoice → payment chain at 19/30. Turkish questions score 91.8%, English ones 80.0%.

Latest value wins. When a record is re-synced with a new value, or a conversation corrects an earlier statement, the newer fact must come first.

⁵ An independent audit reports 99 score-affecting label errors among the 1,540 questions, and a standard judge that accepted 62.8% of intentionally wrong answers [14].

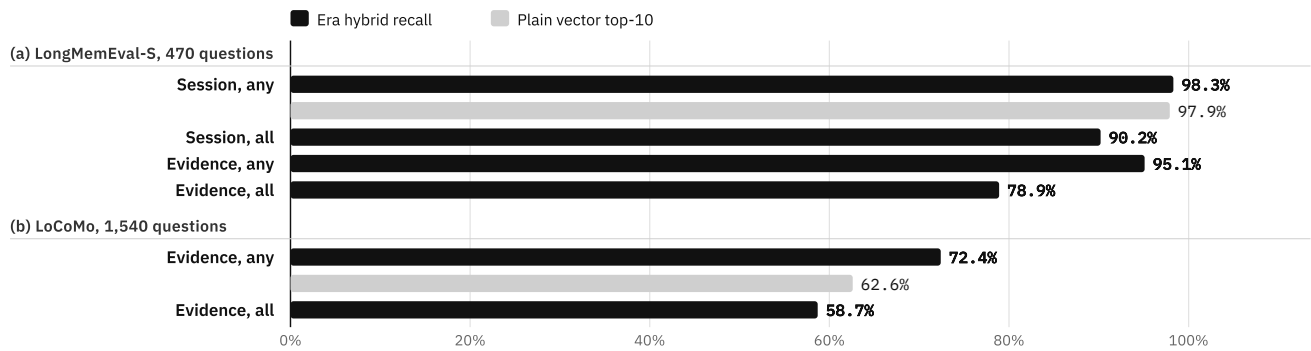


Figure 7: Retrieval on the public benchmarks, top 10 items. Any: at least one gold evidence item returned; all: every one. Plain-vector values come from the preceding run on the same data and machine.

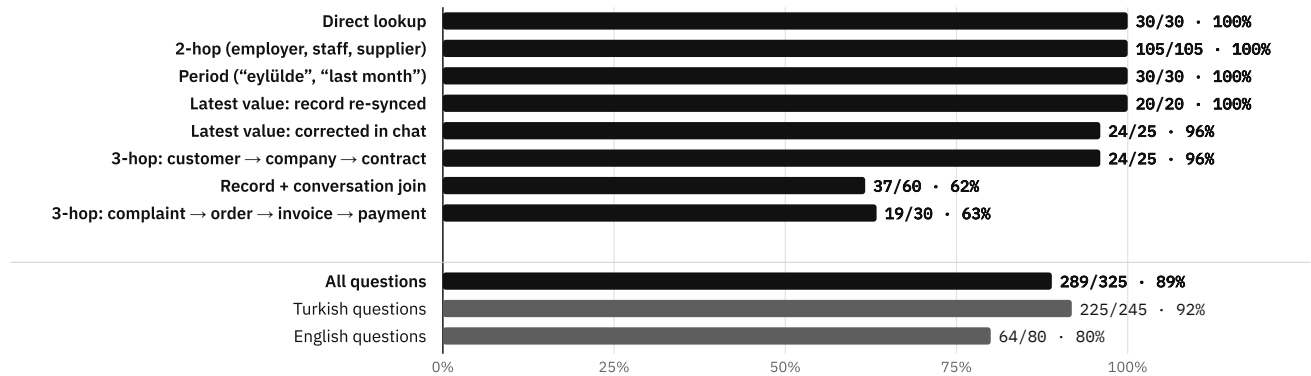


Figure 8: Held-out business set H2 (325 questions), answer in context by question type. Bars to scale; labels give correct / total.

This held for 20/20 record updates and 24/25 conversation updates.

Figure 9 isolates multi-hop questions on the 100,000-row scale set. Path search answers every direct, two-hop and three-hop question. A plain vector top-10 over the same memories answers 0/60 two-hop and 1/40 three-hop questions; its 35% overall score (70/200) comes almost entirely from direct lookups. The same pattern is reported for vector RAG on public multi-hop queries [15].

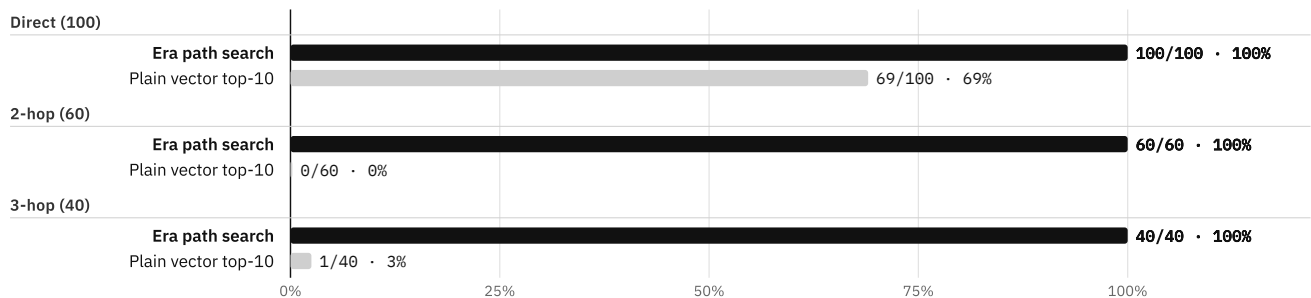


Figure 9: Scale set (100,000 graph rows, owner asking): path search vs plain vector top-10, answer in context.

4.4 Context size

Figure 10 compares context tokens per query. On LoCoMo, hybrid recall returns ten raw turns, about 596 tokens. On business questions, path search returns 1,067 tokens and an Era chat turn hands the model 1,245. The 2025 figures are what Mem0, Mem0g and Zep used for their QA scores; the 2026 systems use larger contexts at much higher accuracy. Fewer tokens at lower recall is not cheaper like-for-like, and tokenizers differ, so these numbers are indicative.

4.5 Latency

Figure 11 shows median search time. Era Brain's figures are measured in process without a network hop; vendor figures are measured through their hosted services or as stated in their publications, so the comparison indicates an order of magnitude, not a ranking. Against the fastest published hosted retrieval latency (Zep, 87 ms on LoCoMo), Era Brain's 29 ms is about 3× lower.

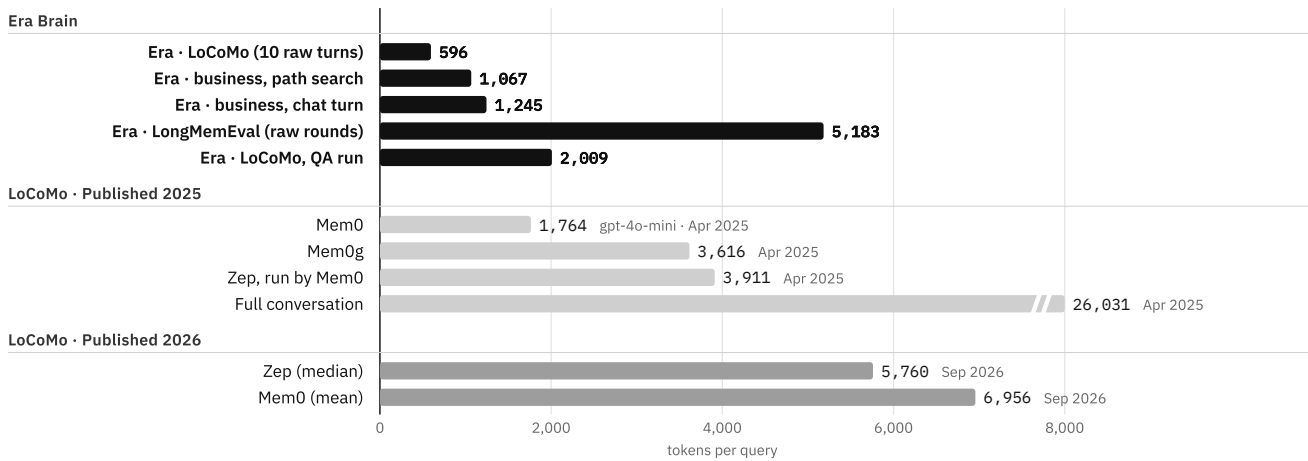


Figure 10: Context tokens per query (p50 unless noted). Black: Era Brain; grey: published LoCoMo figures with date. The full-context bar is cut; its value is printed.

Graph recall on business questions takes longer, 53 ms, because it walks up to three levels and applies the visibility rule at each⁶.

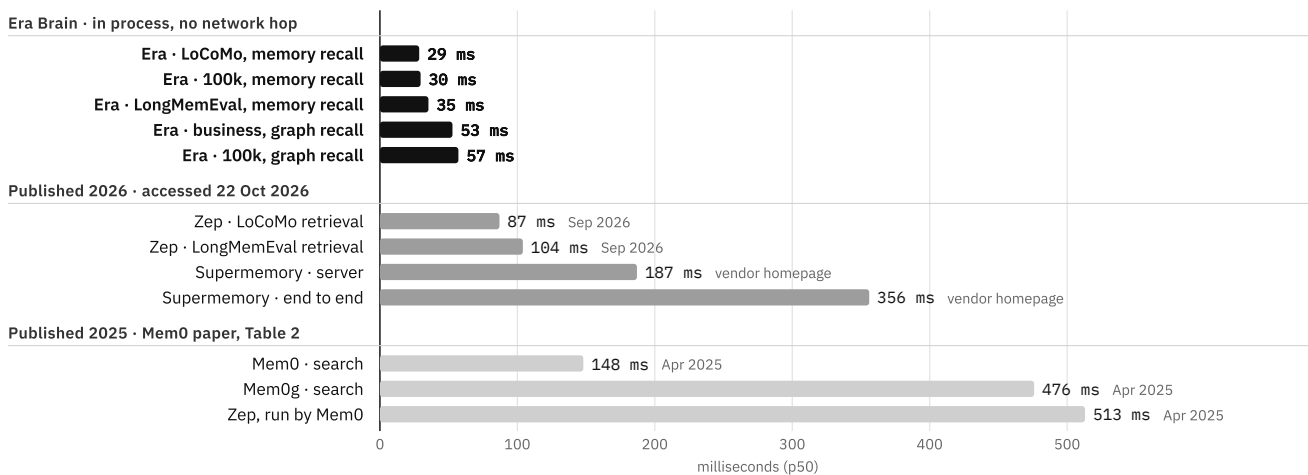


Figure 11: Median (p50) search latency in milliseconds, linear scale. Black: Era Brain, in process, local database, no network hop. Grey: as published (accessed 22 October 2026), measured through hosted services or as stated.

4.6 Permissions

In the permission test a sales member (no finance, invoices or HR) asks every question, and a finance member (no HR) asks every targeted permission question. 865 secrets are planted in set A (820 invoice amounts, 20 salaries, 15 credit limits, 10 private notes) and 790 in set H. A leak is any secret the asker may not see appearing in the returned context. Seven member read paths were swept: memory and graph recall in the agent service, recall in the studio, Era's brain-search and brain-graph tools, the Era chat-turn context and the builder copilot context. Eight callers with no person (research, two workflow blocks, internal recall and list, copilot and graph context, meetings brief) were checked separately. Figure 12 shows the result: no leak in any answer.

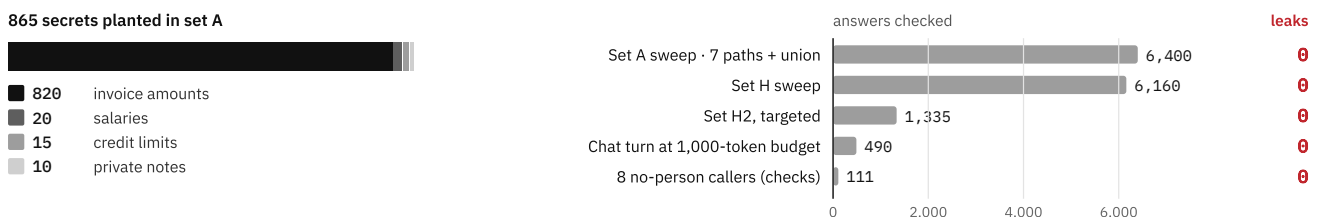


Figure 12: Left: composition of the 865 secrets planted in set A. Right: answers checked per test, with leaks found (red).

⁶ Mem0's 2026 README gives "latency p50 0.88 s" on LoCoMo without saying what it covers; it is not plotted. Supermemory's homepage figures (server and end to end) are plotted as published.

4.7 What changed the numbers

Table 3 lists the changes that moved the results most. Each row compares the engine before and after the named change on the same data and machine; most rows bundle several related edits, so a row does not isolate a single technique.

Table 3: Before and after comparisons. Before values come from earlier code versions on the same data.

Change	Before	After	Set, measure
Full-text search that uses its index, plus a batched access counter	4,031 ms	30 ms	100k · memory search p50
Typed path search with precise seeds and question-ranked facts, instead of a 1,000-node expansion	0/100 · 2,466 tok · 377 ms	100/100 · 1,060 tok · 55 ms	100k · multi-hop: answer, tokens, p50
Hybrid recall instead of plain vector top-10	62.6%	72.4%	LoCoMo · evidence, any
Path search instead of plain vector top-10	35%	100%	100k · all 200, answer in context
All Turkish matching and ranking work on the path engine (suffixes, consonant softening, end-node facts, record kinds, period words)	48.0%	88.9%	H2 · answer in context
Visibility rule on every step, with its row-level-security cost	34 ms	56 ms	first Turkish set (310 q) · graph p50

5 Comparison with related systems

Table 4 compares capabilities as documented by each project on 22 October 2026, from vendor documentation, papers and source code [6, 7, 16, 9, 17, 18, 19, 20]. We mark only what is documented; “?” means we searched and found nothing. Two findings cut against Era Brain. First, a read path without a language model is no longer distinctive: Mem0 v3, Graphiti, Zep and Supermemory skip the LLM by default. Second, fine-grained access exists elsewhere: Zep Enterprise offers policy- and attribute-based rules on ingestion metadata [16], Supermemory restricts members to namespaces, and Cognee grants permissions per dataset. What we did not find elsewhere is a per-fact rule derived from the business's own role and module grants and applied at every graph step, or a published permission-leak test.

Others lead in several areas: published end-to-end QA (Mem0, Zep, Supermemory); bi-temporal edges with invalidation (Zep and Graphiti; Era's edges carry no validity interval and graph walks read current facts only); managed extraction and custom ontologies (Zep, Mem0); hosted products, SDKs and integrations; documented compliance (Zep: SOC 2 Type II, HIPAA; Supermemory: SOC 2, HIPAA); and open evaluation harnesses (Mem0, Supermemory). On LoCoMo, a plain file-search agent (Letta, gpt-4o-mini) scored 74.0% [18], above Mem0's 2025 score, so a files-and-search memory such as a wiki is a strong baseline for chat recall. Our comparison concerns business records and permissions.

Table 4: Capabilities as documented on 22 October 2026. ● yes · ◐ partial, plan-dependent or limited · ○ no · ? not documented (searched). Era Brain's QA entry stays pending until the run in Section 4.1 completes.

	Era Brain	Mem0	Zep / Graphiti	Super-memory	Letta	Cognee	Wiki / Obsidian	Vector RAG
No LLM on the default read path	●	●	●	●	○	○	○	●
Business records written as typed, versioned data without an LLM	●	○	○	○	○	◐	○	○
Typed, question-guided multi-hop path search	●	○	◐	○	○	◐	◐	○
Validity time on graph edges (bi-temporal)	◐	◐	●	◐	○	◐	◐	○
Latest value wins at read time	●	◐	●	●	○	?	◐	○
Per-fact access rule from role/module grants, on every graph step	●	○	◐	◐	○	◐	◐	◐
Turkish morphology in keyword search	●	○	○	?	?	○	○	?
One datastore for vectors, text, graph and access	●	◐	○	?	●	○	●	●
Published end-to-end QA score	●	●	●	●	●	◐	○	○
Published permission-leak test	●	○	○	○	○	○	○	○
Published structured-business benchmark	●	○	○	○	○	○	○	○
Hosted product or public SDK	○	●	●	●	●	●	◐	?
Documented compliance (SOC 2, HIPAA)	?	?	●	●	?	?	○	?
Open evaluation harness	○	●	?	●	?	?	○	○
Memory editable as plain files	○	○	○	○	●	○	●	◐

Table 5: Published competitor figures used in this report, with reader model and date. All are end-to-end QA scores; none is retrieval recall.

System	Benchmark	Score	Reader / judge	Date	Ref.
Mem0	LoCoMo	66.9%	gpt-4o-mini	Apr 2025	[10]
Mem0g	LoCoMo	68.4%	gpt-4o-mini	Apr 2025	[10]
Zep, run by Mem0	LoCoMo	66.0%	gpt-4o-mini	Apr 2025	[10]
Zep, own run	LoCoMo	75.1%	—	May 2025	[13]
Full context	LoCoMo	72.9%	gpt-4o-mini	Apr 2025	[10]
Letta, file agent	LoCoMo	74.0%	gpt-4o-mini; scoring not stated	Aug 2025	[18]
Mem0	LoCoMo / LongMemEval	92.5% / 94.4%	model not named	Sep 2026	[11]
Zep	LoCoMo / LongMemEval	94.7% / 90.2%	gpt-5.4	Sep 2026	[8]
Zep	LongMemEval-S	71.2%	gpt-4o	Jan 2025	[12]
Full context	LongMemEval-S	60.2%	gpt-4o	Jan 2025	[12]
Supermemory	LongMemEval-S	97% / 84.6% / 85.2%	gpt-4o / gpt-5 / gemini-3-pro	May 2026	[21]

6 Limitations

- Answer accuracy is pending in this draft. Every other accuracy number is retrieval: whether the evidence reached the model, not whether a correct answer was produced.
- The business sets are synthetic and self-built. H2 is a reworded variant of a set used during tuning; validation on real customer data is still needed.
- Records and relations in the business sets were generated in the writer's input format. Today's connector mapping for a generic app emits generic relations (bought, paid by, related to), so a live graph may be sparser than the one walked here.

- “Answer in context” is a string match. On the 100k three-hop questions the source facts for every hop are present in 5/40; the answer arrives through the path line.
- The 100k set was bulk-loaded by SQL rather than written through the product, and only the owner asked.
- Latency is one in-process call with a single client on a local database; LoCoMo and LongMemEval latency predate a final speed-only change to recall.
- LLM layers (extraction, relation judging, consolidation, relationship discovery) were not run. Facts derived by consolidation are not yet tagged with module, workspace or privacy, so the zero-leak result does not cover them.
- As-of (point-in-time) recall was not tested, and graph edges carry no validity interval.
- Competitor figures are taken from their publications, not re-run; they use different models, tokenizers and setups.
- LoCoMo is licensed CC BY-NC 4.0; it was used for evaluation only and is not redistributed. The harness is not yet public.

7 Conclusion

Era Brain answers business questions from one PostgreSQL database with no language model on the read path. On the tests we built for that setting (multi-hop record joins, latest values, Turkish phrasing and per-member permissions) it found the answer for 88.9% of held-out questions and leaked no secret in 12,560 answers, with about 1,245 tokens of context per chat turn. On public conversational benchmarks its retrieval improves on a plain vector baseline, and its LoCoMo answer accuracy is 76.1%. Next steps: validation on real customer data, a public release of the harness, a run of a public access-control benchmark [5], and tagging consolidation-derived facts with module, workspace and privacy.

References

- [1] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, D. Yu. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. ICLR 2025. arXiv:2410.10813. Data: huggingface.co/datasets/xiaowu0162/longmemeval-cleaned (MIT). <https://arxiv.org/abs/2410.10813>
- [2] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, Y. Fang. Evaluating Very Long-Term Conversational Memory of LLM Agents. ACL 2024. arXiv:2402.17753. Data: github.com/snap-research/locomo (CC BY-NC 4.0). <https://aclanthology.org/2024.acl-long.747>
- [3] K.-H. Huang et al. CRMarena: Understanding the Capacity of LLM Agents to Perform Professional CRM Tasks in Realistic Environments. arXiv:2411.02305, 2024. <https://arxiv.org/abs/2411.02305>
- [4] K.-H. Huang et al. CRMarena-Pro. arXiv:2505.18878, 2025. <https://arxiv.org/abs/2505.18878>
- [5] GateMem: a benchmark for multi-principal shared-memory agents. arXiv:2606.18829, June 2026. <https://arxiv.org/abs/2606.18829>
- [6] Mem0 documentation: OSS v2 → v3 migration; graph memory. <https://docs.mem0.ai/platform/features/graph-memory> (accessed 22 Oct 2026)
- [7] Graphiti (getzep/graphiti), graphiti-core 0.31.2, source code. <https://github.com/getzep/graphiti> (accessed 22 Oct 2026)
- [8] Zep. Research results (page modified 29 Sep 2026). <https://www.getzep.com/research/> (accessed 22 Oct 2026)
- [9] Supermemory documentation: search defaults; container tags (namespaces). <https://supermemory.ai/docs/concepts/container-tags> (accessed 22 Oct 2026)
- [10] P. Chhikara, D. Khant, S. Aryan, T. Singh, D. Yadav. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. arXiv:2504.19413, 2025, Table 2 (gpt-4o-mini). <https://arxiv.org/abs/2504.19413>
- [11] Mem0. Research results and 'The token-efficient memory algorithm' (blog, Apr 2026, updated 28 Sep 2026). <https://mem0.ai/research> (accessed 22 Oct 2026)
- [12] P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, D. Chalef. Zep: A Temporal Knowledge Graph Architecture for Agent Memory. arXiv:2501.13956, 2025. Blog: getzep.com/blog/state-of-the-art-agent-memory. <https://arxiv.org/abs/2501.13956>
- [13] Zep. Lies, Damn Lies, Statistics: Is Mem0 Really SOTA in Agent Memory? Blog, 6 May 2025. <https://www.getzep.com/blog/lies-damn-lies-statistics-is-mem0-really-sota-in-agent-memory/> (accessed 22 Oct 2026)

- [14] Penfield Labs. Audit of LoCoMo labels and judge leniency. Blog post, 8 Apr 2026.
- [15] Y. Tang, Y. Yang. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. arXiv:2401.15391, 2024. <https://arxiv.org/abs/2401.15391>
- [16] Zep documentation: attribute-based and policy-based access control (Enterprise). <https://help.getzep.com/v4/attribute-based-access-control.md> (accessed 22 Oct 2026)
- [17] Cognee (topoteretes/cognee), commit 37f788bd, source code. <https://github.com/topoteretes/cognee> (accessed 22 Oct 2026)
- [18] Letta. Benchmarking AI Agent Memory: Is a Filesystem All You Need? Blog, 12 Aug 2025. <https://www.letta.com/blog/benchmarking-ai-agent-memory> (accessed 22 Oct 2026)
- [19] A. Karpathy. LLM wiki (gist). <https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f> (accessed 22 Oct 2026)
- [20] pgvector: open-source vector similarity search for PostgreSQL. README. <https://github.com/pgvector/pgvector> (accessed 22 Oct 2026)
- [21] Supermemory. LongMemEval-S evaluation report, May 2026. <https://supermemory.ai/research/longmembench/> (accessed 22 Oct 2026)

Appendix A Where each claim lives in the code

Repository commit 3aba0bbfb. Paths are relative to the repository root; brain-paths = packages/db/brain-paths.ts, brain-viewer = packages/db/brain-viewer.ts.

Table A1: Claims and their source locations.

Claim	Code
No LLM call while reading; the only model is the local query embedding	brain-paths:1187, :55-63 · apps/brain-agent/src/brain/recall.ts:8-16 · apps/studio/lib/brain/recall.ts:1-14 · apps/studio/lib/brain/embed-client.ts:6-14
Embeddings: multilingual-e5-small, q8, CPU, 384 dimensions; HNSW cosine index	apps/brain-agent/src/brain/local-embeddings.ts:17 · embeddings.ts:22-28 · packages/os-kernel/sql/450_brains.sql:146,150
Keyword and vector legs run side by side; EN + TR stop words dropped before the GIN probe	apps/brain-agent/src/brain/recall.ts:184-217 · sql/898_brain_speed.sql:37-51,189-265
Anchors: exact name probe first, Turkish case endings removed	brain-paths:296-308,325-327,341-375,462
Then word containment + trigram similarity, then by meaning	brain-paths:377-446 (similarity :417), :471-491, :1239-1252
Kind words with consonant softening ('talebindeki' → talep)	brain-paths:597-605,690-718
Up to 3 hops; 48 best nodes walk on per level; at most 600 nodes	brain-paths:78-79,88,1227,1295-1297,1363-1365
A relation fanning out to more than 50 neighbours is not followed unless named (then 8)	brain-paths:72-76,785-791,1277
Facts ranked by closeness + question stems; period words favour that period (graph facts)	brain-paths:224-237,925-944,963-970,984-1024
Same attribute twice: newest first, older marked earlier	brain-paths:1064-1142,1616
Token budget for the context (estimated at 3 characters per token)	brain-paths:1610,1620-1687 · apps/studio/lib/era/turn/right-now.ts:112
Visibility rule: workspace, modules read, privacy, current facts, on every read	brain-viewer:184-232 · packages/os-kernel/sql/900_brain_viewer_access.sql:19-41
Hidden nodes are neither shown nor walked through (seeds, hops, evidence, facts)	brain-paths:357-366,425,487,779-798,873,947,1157-1161
Viewer resolution fails closed (reads no module)	brain-viewer:76-108
Row-level security on brain tables: organization / workspace / person	sql/092_tenant_strict.sql:392-400 · sql/898_brain_speed.sql:354-386 · packages/db/index.ts:7-18,78-170
No person: workspace-shared facts only (product callers)	brain-viewer:110-178 · apps/studio/lib/brain/brain-context.ts:122-123 · workflow-scope.ts:84 · apps/brain-agent/src/research/run.ts:90
Record sync: module tag, valid-from, versions; one locked transaction	apps/brain-agent/src/brain/connector-ingest.ts:199-201,242-369 (lock :244, tag :259-264)
Fact extraction and relation judging (LLM); six entity types	apps/brain-agent/src/brain/extraction.ts:26-37,66,99,232-276 · ingestion.ts:899-907
Consolidation: 03:00 UTC, organizations with an active connector row, one LLM pass	apps/brain-agent/src/index.ts:426-439 · consolidation.ts:73-118,139-154,186-214,271-286
Relationship discovery: co-occurrence + embedding candidates; LLM names an inferred edge	apps/brain-agent/src/brain/relationships.ts:1-30,140,201-203 · extraction.ts:385-420
Community detection: Louvain (networkx, seed 42); ontology OWL + SHACL checks	services/semantica-sidecar/app.py:122-189 · ontology-core.ts:82-165 · ontology-tick.ts:130-145
Hash-chained decision log; no production writer yet	apps/brain-agent/src/decisions/provenance-core.ts:48-90 · provenance.ts:45-110
Facts editable from Era's screens (correct, forget, keep private, confirm)	sql/450_brains.sql:156-245