Oroi: a semantic network for conversational agents

Continuous memory for an infinite conversation

Igor Laburu
Gako AI
oroi@gako.ai

Preprint, July 2026
DOI: 10.5281/zenodo.21208930

Abstract

The usual memory of a chatbot (RAG, *retrieval-augmented generation*) stores the conversation as text and, for every question, fetches the fragments that most *resemble* it. This works well for querying documents, but a conversation is not a document: what matters often does not resemble the question, changes over time, and depends on what was just said. This document describes **Oroi**, a memory that stores the conversation as a directed, weighted *semantic network* — concepts joined by weighted associations — and equips it with a dynamics inspired by human memory: what is mentioned activates, activation spreads to what is associated, everything decays as turns pass, associations that recur are reinforced (Hebb's rule), and a background process consolidates what matters. Remembering combines *recognition* (what is still active from the recent conversation) with *evocation* (spreading from the cues of the question).

On a 224-scenario benchmark, **Oroi** ties a strong RAG baseline on direct questions — where the proposal does not aim to win — and pulls apart where resemblance confuses: with distractors (0.97 vs. 0), when a fact changes (0.25 vs. 0) and on chains of three associations (1.00). On the global retrieval score, 0.75 vs. 0.55: it wins 43 scenarios and loses none, a difference chance does not explain ($p < 10^{-4}$). The main limit today is the quality with which relations are extracted from each sentence.

1 Introduction

When an assistant with memory is asked "where do I work?", and weeks earlier it was mentioned that "my office is in Madrid", the answer exists, but it does *not resemble* the question: to reach "Madrid" one must pass through "office". The dominant memory in chatbots, RAG, cannot follow that path: it turns each text into a numeric vector that summarises its meaning (an *embedding*) and retrieves the fragments whose vector most resembles the question's. It is a good search engine but a poor recollection, for three reasons:

- **Time.** When a fact is corrected ("actually, I moved"), resemblance retrieves the stale value and the current one with no notion of which holds.
- **Association at a distance.** The answer often does not resemble the question, as in the example.
- **Anticipation.** Human memory pre-activates what is about to become relevant (the *priming* effect); a similarity index has no such notion.

We propose to treat conversational memory as a *dynamical system* over a network of concepts, rather than as a search engine over text. The network *finds*; the verbatim text of the

conversation, stored alongside, *speaks*. We do not claim to model the brain: we borrow from neuroscience a handful of operational ideas — activation, association, decay, consolidation — because they solve concrete problems of dialogue. Our falsifiable hypothesis is:

> *Activation-sensitive retrieval outperforms pure vector-similarity retrieval in long conversations with recurring references and topic shifts.*

It is accompanied by a second hypothesis this work does not yet measure: over ultra-long conversations, such a memory should moreover be more efficient, compact and bounded than archiving the whole transcript, because what grows is the consolidated knowledge, not the raw text; memory re-elaboration (§9) is the piece that will make it measurable.

2 The origin of the architecture

We combine established techniques and apply them to dialogue with a language model; none of the pieces is new, the bet is the combination. *Spreading activation* over semantic networks [1] models recall as activation flowing from the mentioned concepts to the associated ones. Computational psychology has spent decades simulating human memory with equations that reproduce its recall and forgetting times; its reference architecture (*ACT-R* [2]) distinguishes a slow strength (what is consolidated) from a fast activation (what is present), which we mirror with two timescales. Hebb's rule [5] — "what fires together wires together" — motivates the reinforcement of associations. The surprise signal follows the idea of *predictive coding* [6, 7], and we frame it as a contextual signal reflecting the relationship between Kahneman's System 1 (the network: fast, automatic) and System 2 (the language model: slow, deliberate) [8]: the network emits the warning of a break in conversational expectation; acting on it belongs to the orchestration layer. As the contrast reference we use RAG [9] with dense retrieval [10] and lexical BM25 [11], combined [12].

3 How it works

3.1 The network

Each conversation becomes a graph. A *node* is a mentioned concept ("office", "Madrid", "my boss"); an *edge* is an association between two concepts, with a *weight* that grows when the association recurs. The weight is what learns; the relation label ("is in", "works with") is only descriptive. Each node keeps two registers with two distinct rhythms: a fast **activation** — how present it is right now; it rises when mentioned and decays with the turns — and a slow **strength** — how consolidated it is; only consolidation (§3.3) touches it. The verbatim text of each turn is stored apart, as *episodes* linked to the concepts they mention. The separation mirrors the classical distinction between **semantic** memory (concepts and associations: the network) and **episodic** memory (the verbatim record: the text) [4]: the network finds, the text speaks.

3.2 Perceiving a conversation turn

An extractor (a small language model, one call per turn) turns what was said into a handful of concepts and relations: the phase neuroscience calls *encoding*. Each incoming concept either *resonates* with an existing node (**cosine similarity** above a threshold τ) or creates a new one. Recognising "the same concept" requires two practical remedies. Partial or misspelled proper names are rescued by a classical lexical search when the vector falls short of the threshold ("Guggen" must find "Guggenheim Museum"). And *values* — amounts, dates, quantities — are compared by exact identity, because their embeddings are nearly identical to one another:

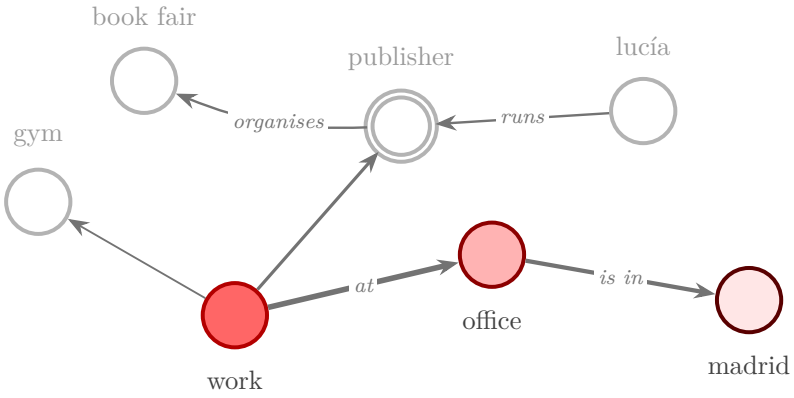

Figure 1: Detail of a semantic network at the moment of the question "where do I work?": the cue activates "work" (deep red) and activation spreads along the associations up to "madrid", decaying with distance (faint red). The width of each edge is its learned weight and the label, when shown, is only descriptive; the double border ("publisher") marks a consolidated concept; grey nodes are inactive.

"15,000 euros" and "60,000 euros" resemble each other more than the merge threshold itself, and would end up fused into one node.

Matching nodes receive a pulse of activation, larger if the relation matched as well. Activation then *spreads* to the neighbours for a few synchronous, damped rounds, by **gradient equalisation**: it flows only from the more active node to the less active one, in proportion to the difference and to the weight of the edge, and the source cedes what flows,

$$\text{flow}_{i \to j} \;=\; \lambda \frac{w_{ij}}{\sum_k w_{ik}} \max(a_i - a_j,\, 0), \tag{1}$$

where a_i is the activation of node i and w_{ij} the weight of the edge joining them. The difference $a_i - a_j$ is the gap between the two nodes: if the neighbour is as active as the source or more, nothing flows. Of that gap, each round transfers only a part, λ (in our case 0.5: half), shared among the neighbours according to each edge's weight; transferring little by little keeps the network from oscillating. And when the flow goes against the edge direction, it is further attenuated by a factor ϕ. Spreading only *redistributes*: what is directly perceived always stays above what merely receives spread (they equalise, never cross), and a network with no input fades out entirely, no exceptions (§7). At turn close all activation is multiplied by a decay factor δ (below one), counting time in conversation turns, never wall-clock; and a **homeostasis** applies a per-node ceiling, a global budget (lateral inhibition: if everything is active, nothing stands out) and a floor. Finally, **Hebbian reinforcement** updates the edges between co-active nodes — the weight grows with the product of their activations, at a learning rate η: $w_{ij} \mathrel{+}= \eta\, a_i a_j$ — and only those that already exist: a one-off coincidence is noise; if it recurs, sleep will promote it to an association.

Two more details. The extractor sees a sliding window of recent turns, solely to resolve "it", "that", "also": facts are extracted exclusively from the current turn. And there is a **speaker asymmetry**: the protagonist is the user; what the assistant asserts only enters if the user takes it up. Without it, the system would end up perceiving its own answers as facts and feeding on itself.

3.3 Memory consolidation

When the conversation goes idle, a background process **consolidates**, the way human memory consolidates during sleep [3]: it merges near-duplicate nodes (automatically when the resemblance is extreme; asking a language model in the doubtful band), promotes to edges the coincidences that recurred, transfers part of the activation to the slow strength, lets that strength

decay very slowly, and prunes what never took hold. The slow strength is the only thing that survives the passing of many turns; activation is *working memory*, and the clock does not touch it: a pause between sessions, however long, cools nothing. Reopening the application resets nothing: resuming a conversation is continuing it.

3.4 Memory retrieval

Retrieval has two routes, as in humans. **Recognition** is passive: whatever is still active from the recent conversation is available at no effort. **Evocation** is active: when recognition is not enough, the concepts of the question send waves through the network — a read that does not alter activation — and where several cues converge, the evoked accumulates; that is how the distant and the faded are reached. Being available, however, does not decide the order: a very recent but irrelevant fact must not win merely for being active. Candidates are ordered by

$$\text{score}(n) \;=\; \rho \cdot \text{relevance}(n) \;+\; \text{strength}(n), \tag{2}$$

the relevance arriving from the question's cues plus the node's consolidated importance, with a gain ρ that balances the two scales. The top K become a short text — the associations as prose, plus the most linked verbatim episodes — appended to the user's turn. And if nothing is lit, the memory stays *silent*: injecting nothing is the default, not a failure. Retrieving also leaves a mark: the memories that do get injected slightly reinforce their slow strength (the *testing effect* from the psychology of memory: remembering also consolidates).

3.5 Surprise

Through spreading, the network holds at all times an implicit expectation: what is active is what it "expects". When what arrives fits none of the currently active topics of the conversation — measured by affinity of meaning with what the network keeps lit — the memory emits a graded **surprise** signal. It is not a memory mechanism but one of *attention*: the fast, automatic part warns the slower, deliberate language model that the conversation has jumped to something unexpected and it is worth stopping to reason instead of staying on autopilot. The signal is implemented; its effective action on the conversation (a tag in the prompt of the model that answers, outside the memory: what in a human would call on consciousness) falls outside the scope of this work.

4 Alternative mechanisms

We mention other mechanisms that were considered, measured and discarded, each for the reason described:

- *Cut-off filters on evocation* (discarding nodes that receive little activation): they removed the distant multi-hop answer — the node three associations away is, by construction, the one that receives the least — and with it the paradigm's measurable advantage.
- *Clock-driven activation resets* (clearing working memory after a long wall-clock pause): this memory's time is conversational (turns), never wall-clock; reopening a session is continuing the conversation, not starting over.
- *Name canonicalisation in the extractor* (feeding it the list of known concepts so it reuses names): it fused distinct referents of similar surface.
- *The original spreading rule* (adding activation to the neighbour without subtracting it from the source): it let the network create activation out of nothing; its correction and measured effect are detailed in §7.

The common lesson: no mechanism is added, or kept, without measuring it under real-use conditions.

5 Implementation

The library is written in Python over SQLite, with the `sqlite-vec` extension for vector search; the embedding model is pinned in the database and never mixed. The core depends on no language model: the embedder, the extractor, the conversational model and the consolidation judge are external providers injected behind interfaces, so the same memory can serve any chatbot against any provider. Embeddings are computed in batches, extraction is a single call per turn, and the retrieved context is appended to the user's turn — never to the system message — to preserve prompt caching.

6 How we measure it

We evaluate on synthetic conversational scenarios, each a sequence of turns plus one or more *probes*: questions that declare which fragments the retrieved context must contain and which it must avoid. The metric, **Recall@context**, is objective (no LLM judge): a probe is correct only if everything required appears and nothing forbidden does. We compare, on identical scenarios, the conditions of Table 1; for each difference we estimate, with a standard contrast (the exact McNemar test), the probability that it arose by pure chance. The benchmark is meant as a *thermometer*, not a target: it deliberately includes the regimes where RAG should win, and a mechanism is adopted only if it generalises — never to lift a single probe.

Condition	Retrieval criterion
No memory	last k turns only
Hybrid RAG	by meaning (vectors) + by words (BM25), combined
RAG over facts	similarity over the *same* extracted nodes
Re-ranker	similarity re-ordered by activation
Oroi	activation over the same network

Table 1: Evaluated conditions. "RAG over facts" isolates the contribution of the dynamics: same extracted material and same serialisation; only the selection criterion changes.

7 Results

Evaluation on the generated corpus (224 scenarios, 32 per phenomenon), under the corrected dynamics of §3.2, with `gpt-4o-mini` as extractor and judge and `text-embedding-3-small` for the embeddings (Table 2).

7.1 The advantage, and where it lives

The global edge (0.75 vs. 0.55 for the best RAG) is no longer a sample-size question: **Oroi** wins 43 scenarios and loses none; the probability of such a difference arising by chance is below 10^{-4}.

The advantage concentrates in the cases that confuse similarity search. On *distractors* (facts with the same shape as the target, competing with it to surface: "Elena is a plumber", "Carlos is a lawyer"... and Carlos's job is what is asked), no RAG manages to bring the right fact without dragging in the neighbours (0.97 vs. 0). On *updating*, only **Oroi** sometimes lets the stale value decay out of the context (0.25 vs. 0). And on *three-hop* (the answer sits three associations away from the question) **Oroi** is perfect (1.00), though it should be said in full: with more sample

Phenomenon	RAG (hyb.)	RAG (facts)	Re-ranker	**Oroi**	n
Recurrence	1.00	1.00	0.00	**1.00**	32
Updating	0.00	0.00	0.00	**0.25**	32
Multi-session	1.00	1.00	0.00	**1.00**	32
Multi-hop	0.97	0.91	0.00	**1.00**	32
Multi-hop (3)	0.91	0.69	0.06	**1.00**	32
Convergence	0.00	0.00	0.06	0.00	32
Distractors	0.00	0.00	0.00	**0.97**	32
Global	0.55	0.51	0.02	**0.75**	224

Table 2: Recall@context by phenomenon and condition, under the corrected dynamics. Against hybrid RAG, **Oroi** wins 43 scenarios and loses none; against RAG-over-facts, 52/0 (McNemar test: $p < 10^{-4}$ in both cases).

the hybrid RAG closes in (0.91; RAG-over-facts stays at 0.69), because its word-based search finds the links of the chain when they share terms with the question.

On direct questions (recurrence, multi-session) all conditions tie, exactly as the benchmark is designed to make them. The re-ranker is rendered ineffective (0.02): it orders by amount of activation, and equalisation redistributes precisely that amount. Read cost is on par with hybrid RAG (≈ 1 s per query for both); at write time **Oroi** adds one extraction per turn that raw RAG does not need.

7.2 The spreading defect and its correction

Under the original spreading dynamics, three of the seven families scored 0 for *every* method, and **Oroi**'s global result (0.55–0.57) was indistinguishable from RAG's. The cause: the spreading rule added activation to the neighbour *without subtracting it from the source*. Two connected concepts could thus feed each other without limit, monopolise the inhibition budget and drown the relevant facts.

The fix applies the gradient equalisation of §3.2: activation flows only from the more active node to the less active one, in proportion to the difference, and the source loses what it cedes — when they equalise, the flow stops. We discarded two variants along the way. The first subtracted from the source everything it handed out, regardless of the gap: the source could end up below the receiver and the network would oscillate. The second suppressed the flow against the edge direction, and lost three-hop: the association from a category to its members travels that way.

To measure the effect of the fix in isolation, we compared the three spreading variants with one another over the same scenarios under identical conditions (an *ablation*; $n = 56$, **Oroi** only): equalisation lifted the global result from 0.55 to 0.73 with no family getting worse. The RAG systems in the comparison retrieve by similarity and make no use of activation spreading, so this fix does not change their results.

7.3 The balance between the relevant and the consolidated

In equation 2, the gain ρ sets the balance between the two criteria that order memories: relevance to the question and consolidated importance. We tested it with three values (1, 6 and 12, under the original dynamics) and the global result barely varied (0.55–0.57). The explanation: the imbalance ρ corrects hardly arises in short scenarios; it takes a long conversation for a concept to accumulate so much consolidated importance that it buries the concrete facts. In that regime the effect is unambiguous: with a high ρ, asking for a concrete figure returns that figure, and

not the dominant concept the conversation has been consolidating. That points to a gap in the evaluation corpus (it lacks long conversations), not in the mechanism.

7.4 From context to answer

Recall@context evaluates retrieval; what remains is whether, with that context, the full system *answers* the question well. To measure it, a small model (`gpt-4o-mini`) generates the answer using only the context retrieved by each condition, and a second model compares it with the expected answer and issues the verdict (the *LLM-as-judge* pattern).

At this level the ranking inverts: raw RAG 0.91, RAG-over-facts 0.83, **Oroi** 0.80. On *convergence* (the answer requires crossing two cues), **Oroi** answers better than any RAG (0.88 vs. 0.44 for raw), and *distractors* stop being a problem for every condition. But **Oroi** fails where its context was perfect: on the three-association chains (0.66, with the full evidence in front) and on *updating* (0.44 vs. 1.00 for raw).

The reason is the format in which the retrieved material is presented. Raw RAG hands over the literal turns, and a small model finds that text easy to use: the temporal reference travels inside the sentence ("I moved") and demands no deduction. **Oroi** hands over distilled facts, and that format is more brittle: one badly extracted relation label is enough to sow doubt, and the model, bound to the context, chooses to abstain.

The joint reading of the two levels is coherent: activation-based *retrieval* finds better, but the full chain performs only as well as its weakest link, which today is the presentation of what was retrieved. Memory re-elaboration (§9) targets precisely that link: memories as brief, precise prose instead of triples. With a production model generating the answers, instead of a small one, the gap should narrow; measuring it remains pending.

7.5 Scorecard of the three motivations

Of the three motivations of the paradigm (§1), the evaluation confirms the second: on association at a distance, **Oroi** is the only method that solves the three-association chains (hybrid RAG closes in thanks to its word-based search). As for the first motivation, time, it is today the weakest point: 0.25 vs. 0 in retrieval and 0.44 vs. 1.00 in answers. Although we marked each fact's recency in the serialisation, it was not enough for a small model: the temporal reference works better woven naturally into the text ("I moved") than as an external annotation, one more argument for textual re-elaboration. Finally, anticipation is implemented as the surprise signal (§3.5) and shows immediate practical utility: each turn emits a graded value that lets the memory notify the agent's orchestration layer of an unforeseen turn in the conversation, so it can re-plan, ask for clarification or leave autopilot. The mechanism is clear and adds no cost (it is computed over what was already perceived); what is missing is its evaluation protocol, which remains declared future work.

8 Limits

The proposal does not aim to beat RAG at flat similarity search; on that ground the scenarios are designed to end in a tie. The value of the paradigm lies where similarity does not reach: chains of associations, facts that change, topics that resume. Its current ceiling is extraction: an association can only be walked if the edge exists, and a small extractor often obtains the entities of an ambiguous sentence without the relation that binds them. The memory is worth what its edges are worth. A second limiting factor is the recognition of similarly named entities ("FedEx" and "FedEx Office"): sometimes they must merge into one concept and sometimes not, and deciding this reliably remains open. The third obstacle is computational cost: one extraction per turn at write time, which RAG does without. Finally, the current evaluation is

short and synthetic: it serves to validate mechanisms, but it does not replace an evaluation over long real conversations or a standard benchmark; both remain future work.

9 Conclusion and intended use

We have described a memory for conversational agents that remembers by activation (the recent, the associated and the consolidated) instead of only by similarity, and a protocol to check whether that helps where the hypothesis predicts. The results support the proposal on the structural case (association at a distance), on resistance to distractors and, with the corrected spreading, on the global result as well (0.75 vs. 0.55, $p < 10^{-4}$). The next advance does not lie in further refining the network dynamics, but in extracting relations reliably.

The intended use is very-long-range memory: relationships between a user and their assistant spanning months or years, across hundreds of sessions. In that regime, the facts the user mentions are not archived as inert text to be looked up: they live in a representation that evolves with use. Each mention reactivates them; each new context re-associates them; repetition consolidates them; emotional charge, or an explicit request to remember, fixes them instantly; and disuse fades them out. That selectivity is what brings the agent closer to human: it remembers what recurred and what mattered, not everything equally.

The architecture is moreover designed so that what grows over time is not the transcript but the stable knowledge: in RAG every turn takes the same space forever, the trivial as much as the important; here consolidation merges, prunes and distils in the background. The next planned piece is memory *re-elaboration*, akin to *reconsolidation* in human memory: during consolidation, old episodes whose facts are already assimilated will be rewritten into briefer, more precise and more relevant versions (the way human memory keeps the gist rather than the transcript), always keeping the original verbatim recoverable.

Since reopening a session is continuing, not starting over, the end goal is an assistant with whom the relationship accumulates: the consolidated is instantly available and the rest remains reachable by evocation. The code and the evaluation protocol are reproducible.

In one sentence: *continuous consolidation and recall of memories, for an infinite conversation.*

References

[1] A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975.

[2] J. R. Anderson et al. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004.

[3] S. Diekelmann and J. Born. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126, 2010.

[4] E. Tulving. Episodic and semantic memory. In *Organization of Memory*, Academic Press, 1972.

[5] D. O. Hebb. *The Organization of Behavior*. Wiley, 1949.

[6] R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex. *Nature Neuroscience*, 2(1):79–87, 1999.

[7] A. Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[8] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[9] P. Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*, 2020.

[10] V. Karpukhin et al. Dense passage retrieval for open-domain question answering. *EMNLP*, 2020.

[11] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in IR*, 3(4):333–389, 2009.

[12] G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. *SIGIR*, 2009.