Activation-Guided Graph Retrieval for Cognitive Memory Reconstruction in Language Model Systems

Marvin Danig
Achiral Research
marvin@achiral.ai

Architecture and analysis preprint
July 2026

Abstract

Retrieval-augmented language model systems generally select context by query-conditioned relevance. We argue that long-lived AI systems require a second operator: activation-conditioned memory availability. We specify an architecture in which request and goal cues seed an authorized memory graph, activation and fan attenuation allocate a bounded traversal frontier, retrieved entities recursively emit cues, and a typed evidence graph is reconstructed before generation. The specification yields five testable findings about candidate selection, traversal control, graph regularization, evidence reconstruction, and reinforcement. We derive implementation-level cost estimates and define ablations, failure conditions, and a reproducible evaluation protocol. This architecture-and-analysis preprint reports no benchmark measurements.

1 Introduction

Retrieval-augmented generation (RAG) gives language models access to information outside their parameters [3]. Dense, sparse, hybrid, hierarchical, and graph-based retrievers have substantially improved candidate selection and corpus-level synthesis [4, 5]. Persistent agent architectures additionally maintain records across interactions [6, 7, 8]. Yet persistence and retrieval do not by themselves define memory availability.

Let $M = \{m_1, \ldots, m_n\}$ be stored memory objects and q a request. A conventional retriever ranks objects using a query-conditioned score

$$s_q(m_i) = \mathrm{sim}(f(q), f(m_i)), \tag{1}$$

possibly fused with lexical, graph, metadata, or learned-reranking features. This score answers which objects are relevant to q. It does not fully specify whether an object should influence inference now: whether it is current or superseded, repeatedly useful or merely popular, direct evidence or model-generated reflection, aligned with the active goal, contradicted by later evidence, or inaccessible under policy.

We introduce *activation-guided graph retrieval* as a stateful control layer over candidate generation and graph traversal. Its target is

$$P(m_i \text{ influences inference now} \mid q, g, H_t, G, \pi, \sigma_i), \tag{2}$$

where g is goal state, H_t is interaction and use history, G is graph structure, π is authorization policy, and σ_i is the epistemic state of memory m_i.

Contributions. This paper (i) formalizes the activation-versus-querying distinction; (ii) specifies activation-guided, fan-aware graph traversal with recursive cue generation; (iii) defines evidence-graph reconstruction as the retrieval output; (iv) derives bounded-frontier, context, storage, and reinforcement estimates; and (v) gives testable hypotheses, ablations, and failure conditions.

2 Activation Versus Querying

Querying and activation are complementary operators over different conditionals. Query relevance can nominate candidates. Activation determines availability under system state.

Dimension	Query-conditioned retrieval	Activation-conditioned memory
Question	What matches or answers q?	What experience should influence inference now?
State	Commonly request-local	Conditioned on goal, history, graph, policy, and epistemic state
Selection unit	Passage, document, row, node	Entity, episode, procedure, edge, path, or subgraph
Time	Usually static between index updates	Strengthens, decays, and changes with outcomes
Structure	Items or neighborhoods	Budgeted traversal over evidence paths
Output	Ranked context items	Typed, provenance-bearing evidence graph

Table 1: Querying estimates request relevance; activation controls memory availability.

The distinction is especially relevant for causal, temporal, procedural, and decision-lineage questions. For "Why was Project X canceled?", lexical or dense retrieval may return the cancellation announcement and semantically similar status reports. A memory system may instead need to reconstruct

$$\text{account loss} \rightarrow \text{forecast revision} \rightarrow \text{budget reduction} \rightarrow \text{cancellation}. \tag{3}$$

The decisive evidence can be several graph steps away and need not resemble the query. The relevant retrieval unit is a path under a goal-conditioned policy, not necessarily a passage under a similarity metric.

3 Architecture

The memory graph $G = (V, E)$ may contain goals, domains, episodes, events, facts, chunks, entities, procedures, decisions, sources, and agent actions. Typed edges may encode temporal order, causality, containment, contradiction, update, provenance, derivation, similarity, assignment, or dependency.

At runtime the system:

1. derives cues from the request, active goal, workflow, and agent state;
2. resolves authorized seed nodes and generates external-search candidates;
3. computes activation for candidate memory entities;
4. allocates a bounded traversal frontier using activation and fan attenuation;
5. expands graph paths, allowing retrieved entities to emit new cues;
6. stops under budget, marginal-evidence, contradiction, or coverage criteria;
7. reconstructs a typed evidence graph with provenance and unresolved gaps;
8. supplies the evidence graph to generation and records qualified outcomes.

Authorization precedes content-bearing scoring and expansion. A node that is relevant but inaccessible must not leak content, identity, degree, or score through activation behavior.

3.1 Activation model

For candidate m_i, a minimal base term is

$$B_i(t) = \log(n_i + 1) - d_i \log(\Delta t_i + 1) + \beta S_i, \tag{4}$$

where n_i is a qualified successful-use count, d_i is a memory-specific decay rate, Δt_i is elapsed time since a qualifying event, and S_i is salience. Equation 4 is an engineering approximation inspired by ACT-R base-level activation [1, 2]; it is not a cognitive-equivalence claim.

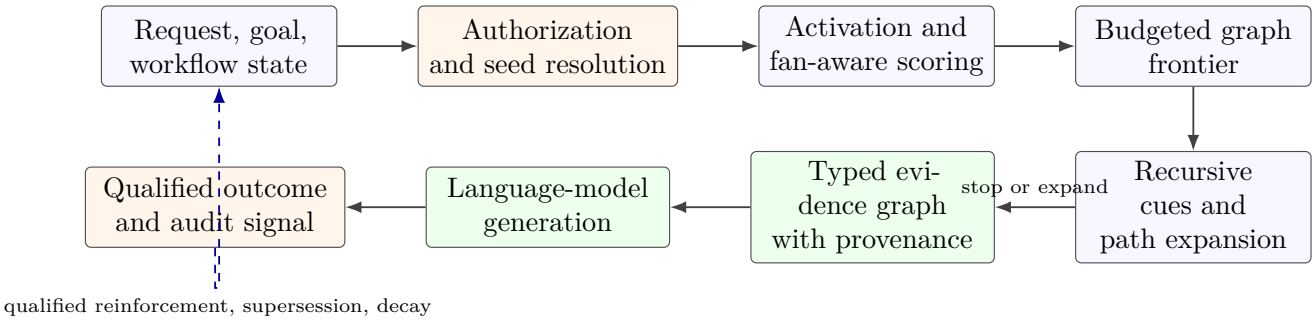

Figure 1: Activation-guided graph retrieval control loop. Authorization is evaluated before content-bearing activation or expansion; only qualified downstream outcomes update future availability.

A runtime score may combine query relevance, spreading activation, utility, verification, and penalties:

$$A_i = \eta s_q(m_i) + B_i(t) + \sum_j W_j S_{ji} + U_i + V_i - C_i + \epsilon. \tag{5}$$

Here $W_j S_{ji}$ is association from cue j, U_i is task utility, V_i is source or verification authority, and C_i penalizes contradiction, staleness, redundancy, or policy risk. Terms may be calibrated, learned, or replaced, but their contributions should remain auditable.

3.2 Fan-aware traversal

High-degree cues such as "customer," "roadmap," or "incident" are weak discriminators. We regularize their traversal influence with

$$\phi(\deg(j)) = \frac{1}{1 + \alpha(\deg(j) - 1)^2}, \tag{6}$$

and score the move from cue j to memory i as

$$F_{ij} = w_j \, a_{ji} \, \phi(\deg(j)) \, \psi(A_i), \tag{7}$$

where w_j is cue importance, a_{ji} is edge association, and ψ maps activation to traversal priority. A path score can be written

$$\text{Score}(p) = \sum_{(j,i) \in p} F_{ij} - \lambda \, \text{Cost}(p) - \gamma \, \text{Redundancy}(p). \tag{8}$$

The exact attenuation function is replaceable. The architectural requirement is that graph fan influences frontier allocation rather than being ignored.

3.3 Recursive cues and evidence reconstruction

Retrieved entities may introduce cues absent from the request. A cancellation record can activate a budget review; the review can activate a forecast; the forecast can activate a customer event. Recursive expansion continues while expected marginal evidence exceeds a threshold and the frontier remains within cost and authorization bounds.

The retrieval product is an evidence graph $G_e = (V_e, E_e)$, not an unordered list. G_e retains selected evidence nodes, typed relations, source provenance, retrieval paths, activation contributions, confidence, contradictory branches, and unresolved gaps. Generation therefore receives explicit relational support rather than being required to infer all causal and temporal structure from adjacent text fragments.

4 Exact Architectural Findings

The following findings are deductions from the specified architecture. They are not empirical performance claims.

F1: Similarity is a candidate signal, not a complete memory policy. Query relevance omits goal state, use history, epistemic status, and path-dependent availability unless those variables are explicitly added.

F2: Activation is most consequential as traversal control. Reranking a fixed top-k pool cannot recover evidence excluded by initial candidate generation. Activation-guided expansion can allocate retrieval budget to graph paths whose downstream evidence is not query-similar.

F3: Fan attenuation is graph regularization. Without degree-sensitive attenuation, generic high-degree cues can dominate expansion. Fan weighting reduces their frontier share while preserving them as routing signals.

F4: Evidence reconstruction changes the retrieval contract. Returning typed relations, provenance, and unresolved gaps moves causal and temporal structure from implicit generation-time inference into an auditable retrieval artifact.

F5: Reinforcement is a feedback-control problem. If retrieval increases future activation, false positives can become self-reinforcing. Updates must depend on qualified outcomes, and supersession must reduce authority without destroying lineage.

5 Analytic Estimates

All estimates in this section are implementation-level bounds under stated assumptions, not benchmark measurements.

5.1 Traversal

Naive breadth expansion with average branching factor b and depth h visits $O(b^h)$ nodes in the worst case. Suppose activation retains at most m frontier nodes per depth and expands at most b neighbors from each. Scoring examines at most hmb candidate transitions. Maintaining a size-m frontier with a heap yields

$$T_{\text{traverse}} = O(hmb \log m), \tag{9}$$

with $O(hmb)$ possible under linear-time bounded selection or bucketed scores. Memory for the active frontier is $O(mb)$ if each depth is processed incrementally. Activation does not eliminate graph-search cost; it supplies a stateful pruning policy.

5.2 Context and storage

Flat retrieval of k chunks averaging c tokens costs approximately kc context tokens. If the evidence graph retains r node summaries averaging s tokens, $|E_e|$ edge labels averaging e tokens, and p provenance tokens per node, then

$$T_{\text{context}} \approx rs + |E_e|e + rp. \tag{10}$$

This quantity is not necessarily smaller than kc. The proposed benefit is higher evidence density and explicit relational support; evaluation must report both support coverage and token cost.

For $N = |V|$ memory entities and $E = |E|$ relations, graph storage is $O(N + E)$. Activation metadata adds $O(N)$ state plus optional access-history storage. A compressed or windowed event history is necessary if raw histories grow without bound.

5.3 Reinforcement

For r retrieved entities, post-retrieval metadata updates require $O(r)$ writes and can be asynchronous. The computational cost is small relative to model inference. The statistical risk is not: updating on retrieval alone implements a popularity loop. Qualified reinforcement should use later citation, confirmation, successful action, or consistency evidence.

6 Evaluation and Falsification

We propose matched-model comparisons among (1) lexical retrieval, (2) dense or hybrid top-k, (3) graph retrieval without activation, (4) activation reranking over a fixed pool, and (5) activation-guided traversal with recursive cues and evidence reconstruction.

Tasks should require evidence structure: organizational why-questions, decision-history reconstruction, incident postmortems, temporal supersession, commitment recall, procedural next-action selection, and contradiction detection. Primary metrics are answer accuracy, evidence-node recall, causal-edge F1, temporal-order accuracy, contradiction recall, provenance completeness, stale-memory false-positive rate, and confidence calibration. Systems should also report latency, graph operations, context tokens, and human audit time.

Ablations remove base activation, goal state, fan attenuation, recursive cue generation, epistemic weighting, evidence reconstruction, reinforcement, and pre-retrieval authorization. The central claim is narrowed or falsified if a simpler graph reranker matches full traversal on causal and temporal recovery under equal model, evidence, and cost budgets.

Reproducibility protocol. Each comparison should publish the corpus snapshot or a synthetic generator, graph-construction rules, authorization policy, cue extractor, activation parameters, random seeds, model and prompt versions, traversal traces, retrieved evidence graphs, token and graph-operation counts, and per-task outputs. Statistical reporting should include paired confidence intervals, effect sizes, calibration error, and failure-case strata rather than aggregate accuracy alone. No benchmark dataset, implementation, or measured result is released with this architecture preprint; the protocol above defines the minimum artifact bundle for any subsequent empirical claim.

7 Failure Modes and Scope

Activation creates state and therefore path dependence. Principal risks include popularity lock-in, obsolete memories retaining authority, model-generated reflections being laundered into evidence, over-attenuation of useful broad cues, incorrect goals steering traversal, permission leakage through scores or graph degree, and evidence graphs too large to audit. Mitigations require source lineage, explicit supersession, qualified reinforcement, authorization before scoring, calibrated stopping rules, and exposure of score contributions and path histories.

The architecture is unlikely to justify its overhead for short-lived factual question answering over a small, stable corpus. Its intended regime is long-lived agents and organizations where experience, goals, evidence authority, and access policy change over time.

8 Related Work and Claim Boundary

ACT-R establishes activation, spreading activation, declarative chunks, and procedural cognition [1, 2]. RAG establishes retrieval from non-parametric stores for generation [3]. GraphRAG and RAPTOR establish graph- and hierarchy-structured retrieval [4, 5]. Generative Agents, MemGPT, and temporal knowledge-graph memory establish persistent agent-memory mechanisms [6, 7, 8].

We do not claim these components independently. The proposed contribution is their ordered runtime combination: goal- and history-conditioned activation; fan-aware traversal of a hierarchical organizational memory graph; recursive cue generation; evidence-graph reconstruction before language-model inference; and auditable reinforcement after qualified use.

9 Conclusion

Querying asks which stored objects match a request. Activation determines which prior experiences and evidence paths should be allowed to shape present behavior. The distinction turns retrieval from a stateless ranking step into a memory-control problem involving goals, history, graph structure, epistemic authority, authorization, and feedback. The architecture yields exact design consequences and tractable cost bounds; whether it improves model behavior is an empirical question with clear baselines, ablations, and falsification criteria.

Declarations

Author contributions. Marvin Danig conceived the architecture, developed the analytic specification, and wrote the manuscript.

Funding. This research received no external funding.

Competing interests. The author is affiliated with Achiral Research. Certain aspects of the described architecture may be covered by patent applications associated with Achiral.

Data, code, and materials availability. This paper reports an architecture and analytic results and therefore introduces no experimental dataset or benchmark code. A companion essay, revision record, and future artifact links are maintained at `achiral.ai/research/activation-guided-graph-retrieval`.

Ethics and limitations. The proposed system can amplify stale, biased, or incorrectly authorized memory if activation and reinforcement are poorly calibrated. Section 7 states the principal failure modes; deployment requires provenance, access-control testing, outcome-qualified updates, and auditable retrieval traces.

References

[1] Anderson, J. R. and Lebiere, C. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, 1998.

[2] Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004. doi:10.1037/0033-295X.111.4.1036.

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. arXiv:2005.11401.

[4] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., and Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. 2024. arXiv:2404.16130.

[5] Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., and Manning, C. D. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. *International Conference on Learning Representations*, 2024. arXiv:2401.18059.

[6] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi:10.1145/3586183.3606763.

[7] Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., and Gonzalez, J. E. MemGPT: Towards LLMs as Operating Systems. 2023. arXiv:2310.08560.

[8] Rasmussen, P., Paliychuk, P., Beauvais, T., Ryan, J., and Chalef, D. Zep: A Temporal Knowledge Graph Architecture for Agent Memory. 2025. arXiv:2501.13956.