THE ACTIVANT *Opinion*

The Last AI Switching Cost?

Accumulated context was supposed to be the moat

August 2026

AUTHOR:

Scott Watson

Vice President, Research

About Activant

Activant is a research-led global investment firm that partners with high-growth companies. Since 2015, we have invested in category-defining businesses during their most critical phases of growth, partnering with founders who have won their initial battles and are ready for the next challenge.

Our approach pairs deep, proprietary research with patient, flexible capital and hands-on operational partnership. We work alongside founders and leadership teams to refine strategy, strengthen operations, and accelerate sustainable growth.

Activant Research is dedicated to uncovering the most exciting emerging technologies, sectors, and companies we believe will shape the future. Our research-driven perspective informs everything we do, helping us invest at meaningful inflection points and support founders in building enduring, category-leading businesses.

You can find out more about Activant and our research at https://activantcapital.com/.

The governance half of our April call has held. The context half has not. In Part II of our Cognitive BI series, we argued that platforms owning both semantic governance and operational context would control the analytics execution layer. This piece corrects the second half of that claim.

Every AI company pitches a moat, and since 2023 they have mostly pitched the same one. The model is rented and the interface gets copied inside a quarter, so the thing left to defend is accumulated context: the record of how one business actually works, built up inside a vendor's product until leaving becomes unthinkable. That claim underwrites most enterprise software funded in the years since, and it is the claim we now think is moving.

Until recently an AI application remembered things the way any software product does. Whatever the model learned about a customer's business, down to the correction someone made last Tuesday, accumulated in tables the vendor controlled, inside an account the customer could cancel but never carry away. The record and the software that wrote it were one asset, and that is what made the application hard to leave.

What is being built now pulls them apart. A memory layer is middleware between the model and the database: it intercepts each exchange, judges what is worth keeping, and writes it as structured rows into a database the customer already owns. The application still runs the workflow. It no longer holds the record.

Accumulated memory is, in our view, one of the last real switching costs in enterprise software. Records export cleanly and dashboards get rebuilt in a sprint. What does not move is the compounding record of how a business reads its own numbers and context. Move that record into infrastructure the customer owns and the most durable moat in software has changed address. Three consequences follow from the arrival of the bring-your-own-database (BYODB) memory layer.

01 The vector databases meant to own AI's data layer are being disintermediated by a free Postgres extension.

02 Semantic-layer vendors slide from mandatory front door to invoked calculator, and get priced accordingly.

03 Every application selling accumulated in-app context as its moat, including the platforms we backed in April, is watching that moat migrate one layer down the stack.

The third is the one we have most to answer for, and we come back to it. To understand why a separate layer is forming at all, start with the bill.

The Bill Has Moved to the Context Layer

A million tokens of GPT-3.5-class intelligence cost $20.00 in November 2022 and $0.07 by October 2024, a fall of more than 280x.[1] Enterprise AI bills climbed anyway. Jellyfish, measuring roughly 12,000 developers across 200 companies, found per-developer token consumption rising 18.6x in nine months.[2]

280×

fall in the price of GPT-3.5-class intelligence, Nov 2022 to Oct 2024

18.6×

rise in per-developer token consumption in nine months, across 200 companies

100×

token multiplier for reasoning models; 15× for multi-agent systems

Language models are stateless. Every agent re-sends its history with each request, rather like a lawyer who re-reads the entire case file before answering each phone call and bills for the re-reading. Our compute work put the multipliers at up to 100x the tokens for reasoning models and 15x for multi-agent systems, so the waste compounds precisely as agents get more capable.[3] The expense has moved off the model and onto the context wrapped around it, creating demand for a durable place to keep what the model needs to know. The first candidate for the job is already losing its position.

A Postgres Extension Is Eating the Data Layer

Through 2023 and 2024, the default for AI context was the standalone vector database, and Pinecone, Weaviate and Chroma won that position because no relational database could then retrieve embeddings at acceptable latency.[4] Then came pgvector, a free extension giving Postgres a native vector type and similarity search, so the database an enterprise already runs can hold an embedding in the same table as the row it describes. Memory no longer has to live somewhere the customer does not own, and on published benchmarks Postgres now matches a comparable Pinecone configuration at lower cost.[5,6]

Three frictions are pushing buyers off dedicated stores: read-heavy pricing that does not track stored volume, the engineering cost of keeping two stores in sync,

and re-embedding lock-in every time the model changes.[7,8] Postgres is comfortable below roughly 50 million vectors, contested to about 100 million, and strained above that.

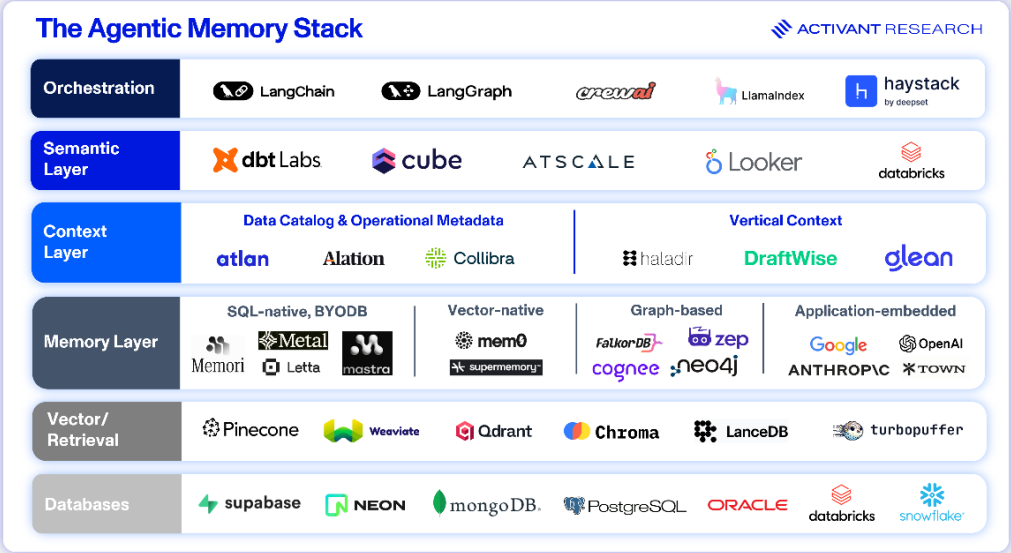
01 **Pricing**	One production account describes a retrieval workload that began at $50 a month and reached $2,847 by its third month on a dataset that never grew, then moved onto pgvector for roughly $200.[7]
02 **Consistency**	Syncing a separate store with the primary database costs engineering time that operators put beyond justification below roughly 50 million vectors, well above where most workloads sit.[7,8]
03 **Lock-in**	Embeddings from different models share no geometry, so changing model means re-embedding everything. The convenience that made a managed store attractive is what makes it the most rigid component in the stack.[8]

Notion, publicly praising Pinecone in January 2024, had moved its search to turbopuffer by that October, and Linear followed in 2025 reporting a 70% cost reduction.[9] Sacra estimates Pinecone's ARR fell roughly 47% through 2025, though other trackers show it closer to flat.[10] The M&A is harder to argue with: Databricks paid about $1B for Neon in May 2025 and Snowflake around $250M for Crunchy Data a month later.[11,12] When both warehouse vendors go out and buy Postgres, it tells you where AI's data layer will live. The layer going up in its place splits four ways.

The top row is where we spend our time. Bring-your-own-database middleware writes auditable memory into infrastructure the enterprise already runs, turning Supabase, Neon and the hyperscalers' Postgres fleets into the default substrate of AI memory without shipping a single AI feature. Buyers making the move cite HIPAA and SOC 2 coverage through controls they already operate, and data staying inside an approved VPC.[8] Memori Labs built the open-source engine that defined the pattern, distilling each exchange into structured rows inside whatever database the customer runs.[13] What it deliberately does not do is hold the record, which makes a low switching cost the pitch rather than a concession. Letta, from the Berkeley team behind MemGPT, runs the same conviction from the agent side.[14]

Which invites the obvious objection: if the layer we favor is open-source, deliberately portable and running on a database the customer already pays for, what is it selling? Not lock-in, and not the data. The write path is a position rather than a feature, because deciding what enters the record, and in what schema, sits with whoever runs the distillation pipeline. Operations is the second: the open-core-to-hosted path sells retention policy, lineage, access control and an audit surface over memory, which a regulated buyer cannot assemble from a repository. The third is the estate: spanning several models across several databases is a job neither a single lab nor a single database vendor can credibly do. The risk is worth naming plainly: Supabase, Neon or Databricks could absorb this as a bundled feature. That is why we underwrite the layer on compression economics and audit surface rather than on data gravity.

Mem0 has won the developer ground war, with 41,000 GitHub stars at its Series A and AWS naming it the exclusive memory provider for its Agent SDK.[15] Our reservation is with the managed product rather than the company: Mem0 is open-source and supports pluggable backends including pgvector, but the hosted vector store most buyers deploy is the component that must be rebuilt when the model changes. Zep models memory as a temporal knowledge graph, strongest where an agent must reason about what was true when, though the record still sits with the vendor. The distinction is less about features than address: whoever holds the write path into the customer's own database inherits the position the standalone vector store was built to own.

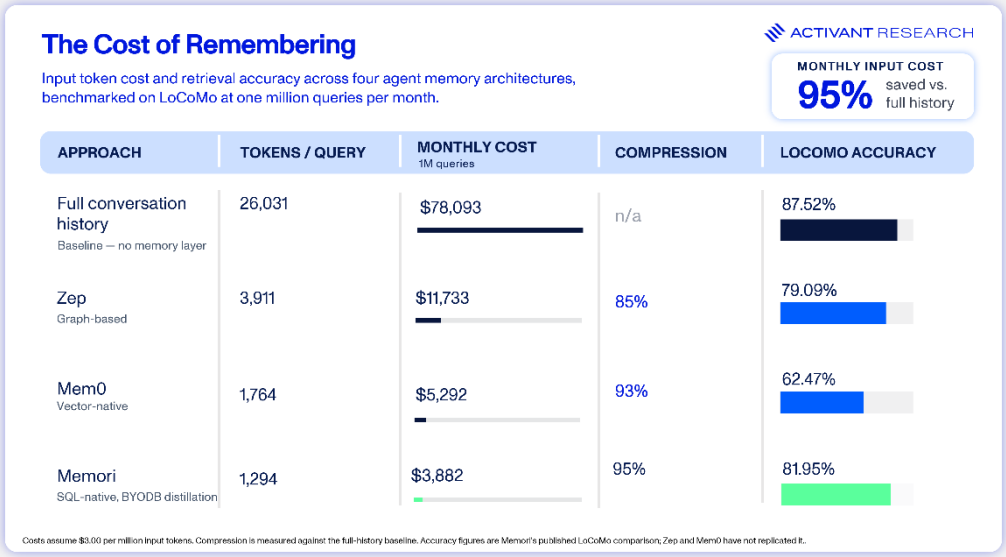


The agentic memory stack. The contest for the record is happening one layer above the databases that will end up holding it.

Sovereignty explains the preference. The token bill explains the timing.

The Number That Makes It a Purchase

On LoCoMo, an academic benchmark for recall across long conversations, passing the full conversation costs about 26,031 tokens per query.[16] At a million queries a month against $3.00 per million input tokens, the gap between re-reading and remembering runs to roughly $74,000 a month, or about $890,000 a year. A million queries a month is not a hyperscale number. It is one production copilot at enterprise scale, or a mid-sized support deployment.



Input-token cost of re-reading versus remembering, at 1M queries a month. Accuracy figures are Memori's own published comparison against Zep and Mem0. Neither competitor has replicated it.

That comparison is published by Memori, and Memori is the top-performing row in it. Zep and Mem0 have not replicated it and we have not verified it independently, so read every number as a ceiling.[17] The trade-off deserves naming. Recall accuracy falls from 87.52% on the full history to 81.95% with Memori's distillation, so the proposition is a 95% cut in input cost for about five and a half points of recall. Where a missed recall is a compliance event, that is disqualifying. For most agent workloads it is the right trade, and it is the trade the category is selling.

Halve the claimed compression and the workload still saves over $37,000 a month, more than enough to fund the middleware producing it. Prompt caching is the obvious objection, and Anthropic and Google both discount cached input by around 90%, though Anthropic charges 1.25 times base input to write a five-minute cache and twice base for an hour.[18] A cache discounts the act of re-reading. It does not carry memory across models, and it does not make the context smaller.

We would not underwrite this category on the arbitrage alone. That is where we part company with the more enthusiastic version of the argument. Token prices keep falling and the labs keep shipping better native memory, so the cost gap narrows from here. What does not narrow is who holds the record. Compression is what gets this layer bought in 2026. Ownership is what keeps it bought in 2030.

Compression is what gets this layer bought in 2026. Ownership is what keeps it bought in 2030.

What Gets Squeezed

The semantic-layer pure-plays get squeezed first, and the mechanism matters: dbt Labs and Cube do not lose the work. Every business needs governed metric definitions, and the semantic layer will still get called. Being called is worth considerably less than being consulted. Memory middleware intercepts unstructured reality upstream and decides what the model needs, while the Model Context Protocol standardized the transport and the Open Semantic Interchange initiative standardized the semantic model itself as a portable file any compliant tool can read.[19,20] An enterprise can swap dbt for Cube without disturbing anything downstream, and a component invoked but never consulted gets priced like a utility.

The second squeeze lands on the platforms our April piece championed, and on every application in the same position. Any AI application selling accumulated in-app context as its moat faces the same arithmetic. Once the record is written into the customer's database rather than the vendor's, the lock-in goes with it. What the application sees degrades as it stops being the only place that record gets written. Building a proprietary context engine instead means funding infrastructure R&D against middleware built for that one problem. Meanwhile the warehouses climb toward the same layer from below, with Snowflake shipping governed semantic views and Databricks making managed agent memory available on Lakebase.[21]

We should be precise about what we are retracting. We still expect those platforms to win deployments, and we are not calling the category wrong. What we no longer believe is that the context accumulating inside them is the durable asset, which was the half of the April thesis that justified the multiple.

The Case Against Us

An argument this convenient for a new category deserves its strongest opposition. Three counter-cases matter, and we concede two of them outright. The labs are internalizing memory, and that is not hypothetical: Anthropic has shipped a memory tool and context editing, and Google runs a managed Memory Bank inside Vertex.[22,23] Verticalized context beats horizontal wherever the schema is the product, and there the embedded engine should win. The third, that Postgres hits a ceiling above roughly 100 million vectors, we concede only in part.[24] The detail sits below.

01 The labs internalize memory

CONCEDED

This is not hypothetical. Anthropic has already shipped a memory tool and context editing, reporting that the pair improved agent performance 39% over baseline, while a separate 100-turn evaluation found context editing alone cut token consumption 84%.[22] Google runs a managed Memory Bank inside Vertex today, and OpenAI holds conversation state server-side.[23] That is the platform tier solving the same problem at marginal cost, bundled with the model, and it keeps every buyer who runs one model and answers to nobody about where the record sits. We concede that segment.

02 Verticalized context beats horizontal

CONCEDED

Where a workflow runs deep enough that the schema is the product, an embedded engine will out-understand any general-purpose layer, because memory and workflow are the same object. In legal, DraftWise mines institutional memory straight out of a firm's document system, the moat-owning pattern we described in our legal AI research. Those buyers purchase an outcome, not infrastructure, and we concede that ground too.

03 Postgres hits a ceiling

PARTLY

The viability of pgvector inflects around 100 million vectors, the top of the band where Postgres is comfortable, and analysts point to OpenAI moving key transactional workloads off Postgres at 800 million weekly users as the existence proof.[24] If agent write volumes outgrow what Postgres absorbs, part of the memory tier lands in object-storage engines such as turbopuffer, and the winner is a different incumbent inside the same layer. Note what that leaves untouched, which is where the record lives.

04 AND ONE THAT CUTS AGAINST US

Anthropic's memory tool is file-based and stores in the customer's own infrastructure, which is architecturally the pattern we are describing, shipped by a lab, for free, bundled with the model. What no lab has shipped is a cross-vendor export standard, and that gap is the thesis.[25]

What survives all three is the tier where enterprise infrastructure spending sits. The large enterprise of 2026 is multi-model by design, multi-database by history, and regulated by obligation, and managed memory fails that requirement set on portability. Server-side conversation state and hosted memory banks do not transfer to a rival's model. The lab exception noted above proves the rule rather than breaking it: even when a lab ships the right architecture, no lab has shipped a cross-vendor export standard, and an auditor cannot subpoena a cache.

That is a view on where value accrues, not a prediction about which companies win. Excellent businesses are being built in the shapes we have called structurally disadvantaged, and some will win their markets on execution and product quality regardless of where the record sits. We would rather be argued with than agreed with.

Where We Land

The revised view is straightforward. Platforms keep winning deployments while increasingly renting their memory, because the switching cost has moved from the application that answers the question to the substrate that remembers the business. We favor database-agnostic memory middleware and the Postgres estate beneath it as a durable category, and we are wary of paying a moat premium for accumulated in-app context.

Founders in the middleware tier inherit less than the application layer believed it owned. This middleware sees the schema and the write path but never a pooled corpus, so the compounding accrues to the enterprise. That is a thinner moat than the applications thought they had, and it is still a real business.

Two conditions would break this thesis. The first is a lab shipping fully portable memory, exportable and usable against a competitor’s model, which would collapse the sovereignty argument. The second is cache economics extending into durable cross-session persistence at cached-token prices, which would shrink the cost case to a feature. We watch both.

Step back and the whole piece describes a single migration. Cost opened the door, Postgres won the substrate, sovereignty keeps it open, and the switching cost that spent a decade locked inside the application layer is settling into the customer’s own database. Applications do not stop mattering. They stop being the thing you cannot leave. The premium paid for in-app context relocates to whoever occupies the write path and to the databases underneath. One consequence is worth flagging now and arguing later. If the record of how an enterprise thinks lives in that enterprise’s own database, the corpus required to teach a model to think like it accrues to the customer rather than to any vendor. That is the subject of our next piece.

If you are building here, we would like to hear from you.

---

1. Stanford HAI, AI Index Report 2025: Research and Development, 2025.
2. Jellyfish, Is Tokenmaxxing Cost Effective? New Data from Jellyfish, 2026. Primary source; also reported by TechCrunch.
3. Activant Research, AI Infrastructure: Compute (1/4), 2026.
4. pgvector, 0.5.0 release notes, 2023. HNSW index support, the point at which low-latency approximate search became viable in Postgres.
5. pgvector, 0.8.0 release notes, 2024.
6. Tiger Data, Pgvector vs. Pinecone: Vector Database Comparison, 2026.
7. Data Science Collective, Vector Databases Are Dying: Here's the Production Evidence, 2026. Practitioner-documented deployment costs, illustrative rather than representative.
8. Activant Research, vector-database expert-call review, 2026. Internal expert calls, not independently published.
9. Sacra, turbopuffer company profile, 2026.
10. Getlatka, Pinecone company profile, 2026.
11. Databricks, Databricks Agrees to Acquire Neon, 2025. Deal value from CNBC, 2025.
12. Snowflake, Crunchy Data acquisition announcement, 2025. Deal value from press reporting; Snowflake did not disclose terms.
13. GibsonAI, Memori open-source repository, 2025, and PRWeb, Memori Labs Launches Memori Cloud, 2026.
14. TechCrunch, Letta, one of UC Berkeley's most anticipated AI startups, has just come out of stealth, 2024.
15. Mem0, Mem0 raises $24M to build the memory layer for AI, 2025.
16. Maharana et al., Evaluating Very Long-Term Conversational Memory of LLM Agents, 2024.
17. Memori Labs, LoCoMo performance comparison, 2026. Published by Memori; not replicated by Zep or Mem0, and not independently verified by Activant.
18. Anthropic, Claude Developer Platform pricing, 2026, and Google, Gemini API pricing, 2026. OpenAI's cached-input discount is model-dependent and lower on older models.
19. Anthropic, Introducing the Model Context Protocol, 2024.
20. Snowflake, Open Semantic Interchange Initiative, 2025.
21. Databricks, Agent Bricks and Lakebase, 2026.
22. Anthropic, Managing Context on the Claude Developer Platform, 2025.
23. Google Cloud, Vertex AI Agent Engine Memory Bank documentation, 2026.

24. Guggenheim Securities, equity research note on database infrastructure, 2026, held in Activant's research base. Analyst commentary on OpenAI's migration of transactional workloads off Postgres.

25. Anthropic, Managing Context on the Claude Developer Platform, 2025.

Disclaimer: The information contained herein is provided for informational purposes only and should not be construed as investment advice. The opinions, views, forecasts, performance, estimates, etc. expressed herein are subject to change without notice. Certain statements contained herein reflect the subjective views and opinions of Activant. Past performance is not indicative of future results. No representation is made that any investment will or is likely to achieve its objectives. All investments involve risk and may result in loss. This newsletter does not constitute an offer to sell or a solicitation of an offer to buy any security. Activant does not provide tax or legal advice and you are encouraged to seek the advice of a tax or legal professional regarding your individual circumstances.

This content may not under any circumstances be relied upon when making a decision to invest in any fund or investment, including those managed by Activant. Certain information contained in here has been obtained from third-party sources, including from portfolio companies of funds managed by Activant. While taken from sources believed to be reliable, Activant has not independently verified such information and makes no representations about the current or enduring accuracy of the information or its appropriateness for a given situation.

Activant does not solicit or make its services available to the public. The content provided herein may include information regarding past and/or present portfolio companies or investments managed by Activant, its affiliates and/or personnel. References to specific companies are for illustrative purposes only and do not necessarily reflect Activant investments. It should not be assumed that investments made in the future will have similar characteristics. Please see "full list of investments" at activantcapital.com/companies/ for a full list of investments. Any portfolio companies discussed herein should not be assumed to have been profitable. Certain information herein constitutes "forward-looking statements." All forward-looking statements represent only the intent and belief of Activant as of the date such statements were made. None of Activant or any of its affiliates (i) assumes any responsibility for the accuracy and completeness of any forward-looking statements or (ii) undertakes any obligation to disseminate any updates or revisions to any forward-looking statement contained herein to reflect any change in their expectation with regard thereto or any change in events, conditions or circumstances on which any such statement is based. Due to various risks and uncertainties, actual events or results may differ materially from those reflected or contemplated in such forward-looking statements.