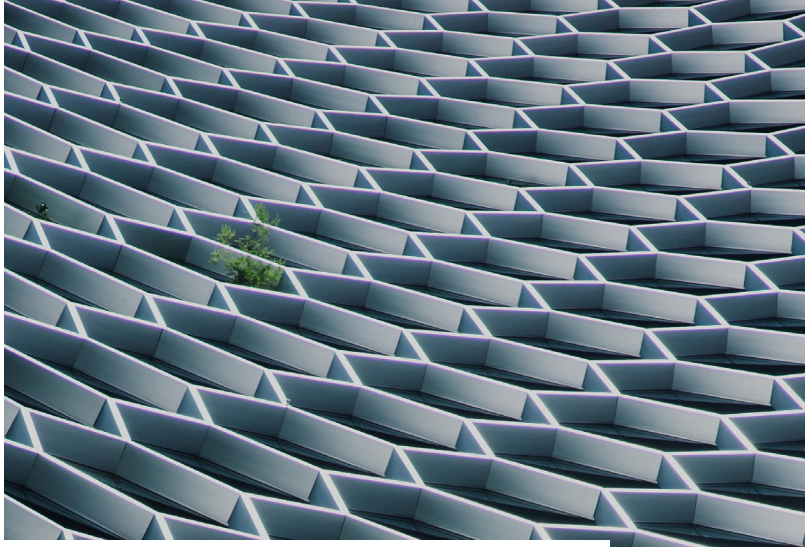
Horizons

S&P Global
Energy

451 Research Market
Insight Report Reprint



Coverage Initiation: Zep tackles agent memory limitations through its temporal context graph

April 10, 2026

by Melissa Incera

As enterprises advance along the agentic AI maturity curve, many are discovering how challenging it is to deliver reliably context-aware agents. Zep addresses these gaps by serving as the context layer for agents, unifying long-term memory, state and retrieval within a temporally aware graph.

Introduction

As enterprises advance along the agentic AI maturity curve, many are discovering how challenging it is to deliver reliably context-aware agents. Our Voice of the Enterprise data shows that memory is the top capability organizations expect from their agents (cited by 46.9% of respondents), highlighting how expectations are rising quickly, even as the underlying capabilities remain in early stages. Industry efforts have centered on traditional retrieval augmented generation and bolt-on memory. However, these approaches struggle with provenance, governance and temporality. Zep Software addresses these gaps by serving as the context layer for agents, unifying long-term memory, state and retrieval within a temporally aware graph. The result is fast context assembly with full preservation of decision provenance and access controls.

THE TAKE

Given that enabling stateful and context-aware agents has already become table stakes, it is striking how small the market around agentic memory remains. Only a handful of pure plays are tackling these challenges, while little by way of industry standards has emerged. We expect this subsector to accelerate rapidly, and Zep looks well positioned to capture that growth, provided it can scale accordingly. While the company is still very small today, we can easily see it becoming a de facto partner in this layer of the enterprise agent stack, if not a direct acquisition target, given the strong consolidation pressures across the ecosystem. Strategically, Zep's points of differentiation are its focus on the enterprise and its temporal understanding. On the latter, its graph integrates both raw (episodic) and derived (semantic) data, understanding both the temporal and emotional valence of it, and can therefore track how facts evolve over time. In benchmarks designed around longer multisession tasks, Zep's approach delivers meaningful improvements in accuracy and latency relative to full context baselines, while also reducing token consumption in long context scenarios.

Context

Founded in 2023 by Daniel Chalef, Zep (Zep AI) was born out of frustration with the stateless nature of large language models. While LLMs are powerful, their lack of memory can lead to repetitive or disconnected user experiences. The team started with an open-source project on GitHub to solve this problem, which has since evolved into a context engineering platform, designed to assemble relevant information from chat history and business data into a temporal knowledge graph. This allows AI assistants to recall past interactions and personal preferences, effectively serving as the memory layer for AI applications.

In 2024, Zep participated in Y Combinator and has since expanded its financial backing to $7.4 million. Key investors include Y Combinator, Engineering Capital and Step Function, alongside various angel investors. This funding has supported the development of its flagship open-source library, Graphiti, and the scaling of its infrastructure to handle the high-concurrency needs of enterprise customers. Zep reports to be nearing $1 million in annual recurring revenue, and reports that most of its customers are pre-IPO, although its enterprise customer base (Fortune 500 companies) is growing quickly.

Technology

Zep's base technology is Graphiti, an open-source, temporally aware knowledge-graph engine built specifically for agent memory. Zep structures this graph across three layers. The first is the episode subgraph, which captures raw messages, text and JSON as a complete, lossless source of truth. Above this sits a semantic entity subgraph, which extracts and organizes entities to create a richer, more navigable knowledge representation. The top layer is the community subgraph, which clusters related entities to provide broader contextual structure, and is responsible for chronological reasoning and for updating the truth-state when contradictions appear.

Crucially, Graphiti is bi-temporal in that it tracks both provenance (where information came from and when) and its validity over time, which ensures that outdated facts are deprecated as new information emerges. This enables agents to reason over a continuously evolving state, differentiating between what was true in the past and what is true now. The result is a reliable, auditable memory foundation that ensures that agentic decisions are grounded in the most current and contextually accurate data available. Purpose-built for context engineering, the engine integrates advanced capabilities like pattern detection, path analysis and temporal weighting alongside hybrid Vector and best matching (BM25) search, ensuring that agents can reason with both exact keywords and deep semantic relationships. The library supports both unstructured and structured data and multi-modal retrieval. It also enables a self-building ontology that autonomously identifies entities such as products, users or projects from incoming messages and maps their evolving relationships.

The company recently released an enterprise platform called Graphzilla, which serves as a graph engine supported by Zep's own vector database. A core differentiation from traditional, monolithic graph databases is the platform's "lakehouse" (or "context lake") concept architecture, which manages many medium-sized graphs (that could represent a specific user, team or project) rather than one single data store. To maintain fast query times, the system employs a hot graph memory management strategy, where only a percentage of active graphs are held in memory at any given time while the rest are continuously snapshotted and moved to cheap object storage.

Market strategy

Strategically, Zep has a heavy enterprise tilt and is marketing primarily to teams that are building agents that need governed, fast context (i.e., finance, support, operations, product teams). The platform is framework-agnostic and integrates with existing agent orchestration stacks. Zep's focus on enterprise readiness shows up in governance features like attribute-based access control (ABAC); ontology controls and flexible deployment options with Zep Cloud (fully managed); Zep Enterprise (bring your own cloud); and Graphiti, which is fully self-managed. It also targets sub-200-millisecond retrieval paths in production configurations while minimizing tokens. Independent benchmarks corroborate the company's story that the platform delivers accuracy and late accuracy, showing gains are strongest in multisession, temporal reasoning and preference categories.

In terms of pricing, Zep's approach has evolved into a credit-based model anchored on usage volume. This structure is designed to scale with the complexity and frequency of an agent's episodes, the company's unit of metric that translates to the ingestion and processing of a message or a piece of data uniting compute and token consumption under the hood. It offers a free plan with 1,000 credits per month and scales from there. For organizations requiring the abovementioned governance features, Zep offers custom Enterprise plans. Despite the fact that the company charges only for input data processing, the cost profile can grow expensive quickly, especially in more complex agentic use cases.

Competition

Competition for Zep is in a state of rapid evolution. Pure-plays are one piece of the puzzle, but approaches vary and many are still small. Mem0 is the highest-profile direct competitor and the best funded ($25 million raised in October 2025); however, it focuses more on the concept of portable memory and the personal assistant category rather than graphing complex enterprise relationships. Another player is Letta, formerly known as MemGPT. Developed at UC Berkeley's Sky Computing Lab, the project rebranded from MemGPT to Letta in 2024 to signal its transition from a research experiment into a production-grade platform for stateful AI agents. Today the company operates as a full agent platform with integrated memory where the LLM manages its context. Other specialist startups are cropping up with increasing speed; one example is Reload, which recently announced a pre-seed funding round as it builds out a shared memory layer for agents.

Zep's biggest competition at present will likely come from hyperscalers and agentic application providers that are memory primitives to make agents stateful. These would include Google Memory Bank in Vertex AI Agent Engine and Amazon Web Services AgentCore Memory. Model providers like OpenAI and Anthropic are also starting to make moves in this regard as they build out agentic solutions (Frontier and Cowork). Crucially, these are platform-bound managed services bound to their ecosystems, whereas Zep stresses neutrality and graph-native semantics. It does however make these platforms increasingly attractive as full-service agentic platforms.

While less of a direct competitive threat today, we are also starting to see some pivots from data management and enterprise search players seeking to position themselves at the agentic context layer. A good example here would be Glean, which is marketing its Enterprise Context layer as the definitive foundation for the agentic era. In recent months, Glean has been pivoting from enterprise search to positioning itself as the invisible middleware for AI, uniting context and retrieval fundamentals with agentic workflows.

SWOT Analysis

STRENGTHS

Zep AI's open-source foundation and research-driven approach provide a transparent, developer-first alternative to memory solutions offered by major cloud providers. The platform is uniquely engineered for performance-rich applications that demand real-time responsiveness. Its sophisticated temporal knowledge graphs offer a level of granular memory management that allows agents to accurately track how facts evolve over time, a critical differentiator for complex enterprise workflows.

WEAKNESSES

Given the expanse of what Zep is attempting to do, the cost of running its system at enterprise scale is expensive. Zep is looking to optimize here, but we see cost sensitivity rising as it relates to AI and agents generally, which could push enterprises to find more economical competitive options.

OPPORTUNITIES

Our data corroborates that Zep is targeting a very real problem (enabling stateful agentic systems) and as there are not many in the space, there is huge opportunity to position itself as the neutral, multi-LLM memory layer for enterprises wary of hyperscaler lock-in, providing a consistent context strategy across diverse model providers like OpenAI, Anthropic and Meta.

THREATS

Zep's biggest threat will likely come from the large vendors going to market with full agentic stacks — hyperscalers and large model providers. These incumbents can leverage existing enterprise credits and deep ecosystem integrations to offer good enough memory solutions as a free or bundled feature, potentially commodifying Zep's core value proposition.

CONTACTS

Americas: +1 800 597 1344
Asia-Pacific: +60 4 296 1125
Europe, Middle East, Africa: +44 (0) 203 367 0681

www.spglobal.com/energy
www.spglobal.com/en/enterprise/about/contact-us.html