

Semvec: Deutsches KI-Team löst das Gedächtnisproblem von KI-Agenten

17.04.2026, 11:21 | IT, New Media & Software

Pressemitteilung von: *Versino AG*



München/Altötting, 14.04.2026 – Während Plattformen wie Anthropic den Zugang für Drittanbieter-Agenten einschränken und Token-Kosten explodieren, präsentiert die Versino-AG-Tochter Semvec eine hochinnovative Memory-Architektur, die KI-Systeme grundlegend effizienter macht – und Europas digitale Souveränität stärkt. Das Tool steht ab sofort im Early Access zur Verfügung.

Die KI-Branche steht vor einem Wendepunkt. Steigende Energiekosten, knappe Rechenzentrumskapazitäten und explodierende Token-Kosten stellen zukünftig die wirtschaftliche Grundlage vieler KI-Anwendungen infrage. Gleichzeitig wächst mit dem Aufstieg multi-agentischer Systeme der Bedarf an effizienter Kontextverwaltung rasant – und die Abhängigkeit von einzelnen Anbietern wird zum Geschäftsrisiko. Mit Semvec präsentiert die Versino AG Gruppe jetzt eine Technologie, die diese Probleme gleichzeitig adressiert: die moderne O(1)-Memory-Architektur, die den Token-Verbrauch von KI-Systemen um über 60 % senkt.

Das Agenten-Zeitalter hat ein Gedächtnisproblem

Multi-agentische Systeme – also KI-Architekturen, bei denen mehrere spezialisierte Agenten autonom zusammenarbeiten – gelten als die nächste Evolutionsstufe künstlicher Intelligenz. Tools wie OpenClaw zeigen bereits heute, was möglich ist: Agenten, die eigenständig Dateien verwalten, Code schreiben und komplexe Aufgaben koordinieren. Doch die Praxis offenbart fundamentale Engpässe.

Erst Anfang April 2026 hat Anthropic den Zugang für Drittanbieter-Tools wie OpenClaw zu seinen Claude-Modellen drastisch eingeschränkt – mit der Begründung, dass diese Anwendungen eine „außerordentliche Belastung“ für die Systeme darstellten. Nutzer, die bislang mit einer Flatrate von wenigen Euros monatlich arbeiteten, sehen sich plötzlich mit potentiellen API-Kosten konfrontiert, die teils deutlich höher ausfallen. Der Grund: Agentische Systeme sind extrem kontexthungrig.

Genau hier setzt Semvec an: Statt immer größere Kontextfenster zu füllen, komprimiert Semvec den semantischen Kontext so, dass der Speicherbedarf mathematisch konstant bleibt – egal wie lange ein Agent arbeitet, egal wie komplex die Aufgabe wird.

O(1)-Memory: Eine grundlegend neue Architektur – kein RAG-Ersatz

Herkömmliche Ansätze zur Token-Reduktion wie Retrieval-Augmented Generation (RAG), Prompt Caching oder Kontextfenster-Management setzen an den Symptomen an: Sie versuchen, weniger Daten in das Sprachmodell zu laden oder bereits geladene Daten zwischenspeichern. Semvec verfolgt einen grundlegend anderen Ansatz. Die Technologie komprimiert nicht den Input, sondern das Gedächtnis selbst – auf semantischer Ebene und mit konstanter Komplexität $O(1)$.

RAG durchsucht externe Wissensbasen und fügt relevante Dokumente in den Kontext ein – das spart Tokens bei der initialen Anfrage, löst aber nicht das Problem wachsender Sitzungskontexte. Prompt Caching speichert wiederkehrende Prompt-Fragmente, reduziert aber weder die Kontextgröße noch den semantischen Informationsverlust über lange Sessions. Kontext-Kompression kürzt den Verlauf, verliert dabei aber unweigerlich Informationen und wächst dennoch mit jeder Interaktion. Semvec hingegen komprimiert semantischen Kontext auf eine konstante Größe – unabhängig von der Sitzungslänge. Das Ergebnis: Über 60 % weniger Token-Verbrauch bei gleichbleibender bis leicht verbesserter Antwortqualität und ohne relevantem Geschwindigkeitsunterschied.

Entscheidend: Semvec ersetzt bestehende Technologien nicht, sondern ergänzt sie. Die Memory-Architektur lässt sich mit RAG-Pipelines, Googles TurboQuant und Prompt-Caching und bestehenden Agenten-Frameworks kombinieren – und potenziert deren Effizienzgewinne.

Collective Intelligence: Wenn Agenten-Flotten voneinander lernen

Semvec Cortex adressiert ein zweites Kernproblem multi-agentischer Systeme: die Isolation. Heute arbeiten KI-Agenten in der Regel mit isolierten Kontexten - was ein Agent in einer Session lernt, bleibt für alle anderen unsichtbar. Cortex bricht mit diesem Prinzip über eine mehrstufige Architektur.

Auf unterster Ebene teilen Agenten ihr Wissen in Clustern - einem gemeinsamen semantischen Gedächtnis. Stellt ein Agent eine Frage, die ein anderer im Cluster bereits beantwortet hat, wird der LLM-Aufruf zusammengeschlossen. Regionen fassen mehrere Cluster zusammen und erkennen über Konsens-Protokolle korrelierte Muster - etwa wenn mehrere Agenten-Gruppen parallel in dieselbe thematische Richtung driften. Der Global Observer darüber meldet Anomalien, die auf keiner einzelnen Ebene auffallen. Ergänzend sichern semantische Anker Sessions gegen Themen-Drift und Prompt-Injections ab, und ein Export/Import-Mechanismus mit Prüfsumme ermöglicht die verlustfreie Übergabe des vollständigen semantischen Zustands an einen Nachfolge-Agenten.

Gerade im Unternehmenseinsatz, wo Dutzende bis Hunderte KI-Agenten parallel arbeiten, eröffnet das völlig neue Effizienz- und Qualitätsdimensionen: Was ein einzelner Agent aus einer Interaktion gewinnt, steht sofort der gesamten Flotte zur Verfügung.

Technologische Resilienz: Kein Zufall, sondern Designprinzip

Die jüngste Entscheidung von Anthropic, Drittanbieter-Tools vom Abo-Zugang abzuschneiden, zeigt exemplarisch die Risiken einseitiger Plattformabhängigkeiten. Semvec wird als Tochterunternehmen der Versino AG Gruppe mit rund Technologiekomponenten. Die Architektur funktioniert LLM-agnostisch: Semvec arbeitet also mit allen gängigen Sprachmodellen, unabhängig vom Anbieter.

„Die nächste große Effizienzrevolution in der KI passiert nicht bei den Modellen selbst – sie passiert bei der Infrastruktur. Unternehmen verbrennen riesige Summen an Token-Kosten, weil ihren KI-Systemen ein echtes Gedächtnis fehlt. Gleichzeitig zeigen Fälle wie die OpenClaw-Einschränkung, wie fragil eine eventuelle Abhängigkeit von spezifischen Plattformen sein kann. Semvec löst beides: eine Architektur, die mathematisch garantiert nicht mitwächst, und eine Technologie, die vollständig in Europa zu Hause ist“, sagt Michael Neuberger, COO der Versino AG.

Verfügbarkeit

Semvec steht ab sofort im Early Access zur Verfügung. Eine interaktive Demo unter demo.semvec.io ermöglicht es, den Token-Vergleich in Echtzeit zu erleben: Anhand von MT-Bench-Szenarien wird sichtbar, wie Semvec den Token-Verbrauch um über 60 % senkt – bei gleichbleibender Antwortqualität. Journalisten und Analysten erhalten auf Anfrage bevorzugte Zugänge.

Versino AG

Ebererstr. 7
84503 Altötting
Deutschland

www.semvec.io

Portrait

Über Semvec

Semvec ist eine KI-Speichertechnologie der nächsten Generation, die semantischen Kontext von LLM-Agenten dauerhaft komprimiert – ohne wachsenden Speicherverbrauch. Die geschützte O(1)-Architektur ermöglicht konstante Komplexität unabhängig von der Konversationslänge. Das Herzstück, Semvec Cortex, vernetzt KI-Agenten zu kollektiv intelligenten Flotten: Was ein Agent lernt, weiß sofort die gesamte Gruppe. Semvec ist vollständig in Europa entwickelt und funktioniert LLM-agnostisch mit allen gängigen Sprachmodellen. Weitere Informationen: semvec.io | Demo: demo.semvec.io

Über die Versino AG

Die Versino AG mit Hauptsitz in Altötting ist der führende SAP-Partner für mittelständische Unternehmen im deutschsprachigen Raum. Seit über 20 Jahren begleitet das Unternehmen mittelständische und internationale Kunden bei der digitalen Transformation – zunehmend mit businessbezogenen KI-Lösungen. Mit rund 200 Mitarbeitenden und Standorten in Deutschland, Österreich, der Schweiz und Tschechien präsentiert die Versino AG ihre Expertise international. Als Mitglied und Gründungsmitglied von LinkedWorld – eines der weltweit größten SAP-Partnernetzwerke für den Mittelstand – verfügt Versino über ein globales Expertennetzwerk. Weitere Informationen: versino.de

Pressekontakt

Deutschland

MichaelNeuberger (COO)

neuberger@versino.de

News-ID: 1309545 • Views: 417 (Stand: 05.10.2026)

Link zur Pressemitteilung:

<https://www.openpr.de/news/1309545/Semvec-Deutsches-KI-Team-loest-das-Gedaechtnisproblem-von-KI-Agenten.html>