

ENTERPRISE CONTEXT & MEMORY ARCHITECTURE · NOV 2026

Beyond Vector Search: Business Ontologies, Verified Retrieval, and Safe Agent Memory

A vendor-neutral engineering blueprint for CDOs, CIOs, and AI platform leads comparing semantic catalogs and four-scope memory isolation.

Key Finding: When retrieved passages state broad rules instead of exact entities, RAG models guess wrong half the time; fixing it requires governed business definitions and evidence-linked memory.

Ajay Pendey · Chief Knowledge Officer, ALLTIPLY

November 12, 2026 · www.alltiply.com

ALLTIPLY Labs Technical Report · Published November 12, 2026 · Author: Ajay Pendey, Chief Knowledge Officer

Core finding: Across 2026 enterprise deployments and fifteen November 2026 primary research and cloud-platform releases, the bottleneck in knowledge assistants and stateful agents has moved upstream of the vector database. When retrieval returns accurate category-level policies without exact entity definitions, language models stop abstaining and guess specific answers wrong roughly half the time. And when teams add naive persistent memory to fix missing context, unverified summaries degrade accuracy over time or risk leaking private facts across users. Production-grade enterprise AI requires two governed layers working together: a **permission-mirrored business ontology** and **four-scope, evidence-linked agent memory**.

Executive summary

- **Bigger vector indexes do not solve missing definitions.** In the November 10, 2026 CLADE preprint, researchers report that giving RAG models broad, factually true category passages caused abstention to drop by 13.8 to 17.4 percentage points while specific entity guesses were wrong roughly **50% of the time**. Adding a structured ontology raised end-to-end recovery from **10.0% to 26.8%**.
- **Unreconciled definitions cap complex analytical work.** In the November 7, 2026 DeepFinance benchmark of 21 $100M deal-diligence data rooms across 22 frontier configurations, the top setup passed only **57.6%** of practitioner rubric criteria, repeatedly performing exact arithmetic on conflicting metric definitions across files.
- **Every major data cloud shipped a semantic layer in the same 48-hour window, yet each stops at its own wall.** Between November 9 and November 11, Google Cloud made BigQuery measures generally available, Databricks made Genie Ontology and Unity Catalog Pages generally available, and Salesforce launched Salesforce Catalog. Each governs definitions inside its own platform; companies running ERP, CRM, ticketing, and document repositories across multiple clouds still need a thin, portable definition spine across systems.
- **Persistent agent memory fails without log pointers, privacy scopes, and promotion gates.** November 2026 studies from Anthropic (`13% -> 48%` wiki QA via offline log consolidation), WikiTrace (trajectory back-pointers), MixUA (`<= 0.02%` forbidden private-fact leakage across four scopes), Falsification-Gated Self-Evolution, and H-Mem show how to design organizational memory that improves pass rates without poisoning context or exposing confidential sessions.
- **Practical starting point:** Do not launch a multi-year enterprise taxonomy program. Certify the 25 to 40 business terms, formulas, permissions, and abstention rules for **one high-value workflow**, pair single-shot and agentic retrieval by query intent, and enforce four explicit memory scopes in code.

Chapter 1: Anatomy of the 50% RAG specificity collapse

For the past three years, enterprise teams followed a standard recipe for internal knowledge assistants: split PDFs, Confluence pages, and SharePoint files into 500-token chunks, embed them in a vector index, add keyword fusion and a cross-encoder reranker, and instruct the language model to answer only from retrieved text.

That pipeline handles straightforward lookup questions well when one self-contained paragraph states the exact answer. It breaks reliably in production on the questions executives, underwriters, support engineers, and operations leads care about most: questions where **a general policy lives in one document, the specific entity assignment lives in a database table, and the metric formula lives in a BI model.**

What CLADE proved about category-level retrieval

On November 10, 2026, S. M. Rayeed, C. Stewart, and J. Ma published CLADE, examining a failure mode that standard RAG faithfulness scores routinely miss. Suppose a user asks for a specific entity-level answer, and the retriever returns passages that are 100% accurate about the *parent category* without identifying the exact child entity.

According to the authors' experiments:

1. **Confidence rises right as precision disappears.** Compared with giving the model no retrieved context, supplying true category-level passages reduced model abstention by **13.8 to 17.4 percentage points**. Seeing familiar domain vocabulary convinced the generator it had enough evidence to answer.
2. **Half of specific guesses were wrong.** When models produced a specific entity answer from those coarse passages, **their guesses were wrong roughly 50% of the time**.
3. **Bigger models and stricter prompts did not fix the failure.** Scaling parameter counts and adding explicit prompt instructions to abstain when uncertain failed to eliminate over-specific guessing.
4. **Structured ontology links materially improved recovery, within clear limits.** Connecting retrieval to a structured domain ontology raised end-to-end recovery from **10.0% to 26.8%**. The authors also state an essential boundary condition: candidate coverage reached **98.5% only when the correct parent family was already known**, meaning upstream entity routing and disambiguation still determine the ceiling. (Note: CLADE is an arXiv preprint evaluated on biomedical hierarchies, not clinical guidance or an audited corporate benchmark.)



Exhibit 1: Reported specificity failure and ontology-assisted recovery in CLADE (Rayeed, Stewart, and Ma, arXiv:2611.14355, November 10, 2026 preprint).

Why the same failure caps financial and operational diligence at 57.6%

Three days earlier, on November 7, 2026, researchers published DeepFinance: Can AI Diligence a $100MM Deal?, testing 22 frontier model and agent configurations across 21 long-horizon private-equity and investment-banking diligence packages. Each task required an average of 27.1 hours for human analysts and was graded against practitioner rubrics averaging 84 criteria (ten runs per configuration).

Across all 22 configurations, **the top score met only 57.6% of rubric criteria**, and the four highest-scoring setups were statistically indistinguishable. Inspecting the error logs reveals the exact enterprise twin of the CLADE effect:

- **Unreconciled metric definitions across files.** A board deck reports "Annual Recurring Revenue" including contracted implementation fees, an ERP billing export excludes multi-year ramp schedules, and an audited statement books recognized subscription revenue. Without a governed ontology stating which source and formula governs each metric, agents pulled numbers from whichever document ranked highest in search and ran clean Python math on mismatched inputs.
- **Disconnected entity state.** A master services agreement states standard 45-day payment terms, while an unsigned email or signed amendment changes terms to 90 days for one subsidiary. Chunk-based RAG retrieves the clean master clause and answers with high confidence.

Field evidence from ALLTIPLY's enterprise RAG audits

In ALLTIPLY's own engineering work building a product help assistant with source citations and documenting seven silent retrieval defects in production, we observed the same structural pattern from inside an enterprise documentation stack:

- **Everyday user shorthand misses formal corpus vocabulary.** Staff and customers type internal product codes, abbreviations, and legacy terms that never appear verbatim in official manuals. Where an identifier

lacked an entry in the product-area mapping table, retrieval silently dropped area filtering and fell back to noisy global search.

- **Global fallback rules degrade exact lookups.** When a simple definition lookup had one exact authoritative passage, a naive rule requiring at least three passages above a rank-fusion threshold forced a broader search that injected tangential category text into the prompt - recreating the exact noise condition studied in CLADE.

Chapter 2: Vendor-neutral comparison of the November 2026 semantic layers



Certifying definitions, formulas, and access boundaries before retrieval turns unstructured text and database tables into verifiable context.

Between November 9 and November 11, 2026, Google Cloud, Databricks, Salesforce, Snowflake, and AWS all shipped features designed to sit between raw enterprise storage and AI agents. Because vendor marketing uses overlapping terms ("ontology," "measures," "catalog," "knowledge graph"), CIOs and CDOs need a clear, vendor-neutral view of what each tool actually does, its current release maturity, and where it stops.

What each platform shipped in November 2026

1. **Google Cloud `BigQuery measures` (GA, November 10):** Google Cloud announced general availability of `BigQuery measures`, embedding reusable metric definitions directly on BigQuery tables and pairing them with knowledge graphs. It supports composite metrics, multi-fact-table joins without double-counting (fan-out/chasm traps), semantic metadata discovery, and consumption from Looker, Conversational Analytics, and BI tools. *Caveat:* Google's post states GA without publishing independent accuracy benchmarks, and definitions govern structured data inside BigQuery rather than external SaaS workflows or unstructured SharePoint policies.
2. **Databricks `Genie Ontology`, `Unity Catalog Pages`, and `Genie Search` (GA/Preview mix, November 9-11):** In its November 2026 release notes and enterprise context blog, Databricks made three components

generally available: `Genie Ontology` (combining curated Unity Catalog metadata with usage-extracted business concepts), Unity Catalog Pages (domain-organized business definitions with `CERTIFIED` / `DEPRECATED` curator controls and a REST API), and `Genie Search` (permission-aware indexing across Google Drive, Jira, Confluence, and Workday), plus GA MCP connectors. *Caveat:* Unstructured document analysis inside Unity Catalog volumes is in **public preview**, external clean-room agent sharing is **beta**, and the release notes report no independent retrieval-quality evaluations.

3. **Salesforce `Catalog` and `Metadata Grounding MCP Server` (GA/Beta mix, November 10):** Salesforce detailed Salesforce Catalog, which indexes org-specific custom objects, fields, flows, Apex classes, validation rules, and Data 360 assets, updating automatically on deployment and enforcing signed-in user permissions. *Caveat:* Requires an edition supporting Data 360 plus admin configuration, and the **Metadata Grounding MCP Server is in beta**; its scope is Salesforce org metadata rather than warehouse financials or engineering docs.
4. **Targeted retrieval controls in `Snowflake` and `Amazon Bedrock` (November 9-10):** Two narrower releases solve important execution steps underneath an ontology. Snowflake added `FL_FILTER_PAGES` to `AI_EXTRACT` and `AI_PARSE_DOCUMENT` so SQL queries parse only relevant PDF page ranges instead of entire 200-page contracts. Meanwhile, AWS added single-shot and retrieval-only modes to `AgenticRetrieveStream` in Amazon Bedrock Managed Knowledge Base (controlled via `maxAgentIteration`), letting applications avoid multi-step agentic search latency on simple lookups (*note:* AWS's post provides architectural examples rather than measured benchmark figures).
5. **GraphRAG toolkits (`Neo4j`, `BG-RAG`, `Flexible GraphRAG`):** Open-source and graph vendors also advanced tooling this month, including Neo4j's `graph-explorer-kit`, Flexible GraphRAG v0.10.0 with Spanner Graph support, and concept-bridged BG-RAG. Neo4j's own developer write-up offers a candid engineering warning worth heeding: automated entity extraction over messy files produces noisy graph scans, and **showing related graph neighbors is not the same thing as proving a claim-to-evidence trace.**



Exhibit 2: Vendor-neutral comparison of November 2026 semantic and ontology layers across data scope, governance maturity, and cross-cloud portability.

Capability dimension	Google `BigQuery measures`	Databricks `Genie Ontology` + `UC Pages`	`Salesforce Catalog` + Grounding MCP	Custom / Private Ontology Spine
Primary strength	Reusable SQL metric formulas & multi-fact join safety	Unified lakehouse tables, curated domain pages & connector search	Live CRM custom objects, flows, Apex & user field permissions	Cross-cloud entity, metric & policy resolution across ERP, CRM & docs
Nov 2026 status	GA (Nov 10, 2026)	Ontology, Pages & Search GA; Volume doc analysis Preview	Catalog available with Data 360; Grounding MCP in Beta	Production-ready in customer VPC, on-prem, or hybrid cloud
Unstructured policy & PDF linkage	External to BigQuery measures	Public preview on UC Volumes; connector index via Genie Search	Limited to CRM metadata & Data 360 objects	Native clause-level citation + page filtering across repositories
Permission enforcement	BigQuery IAM & row/column ACLs	Unity Catalog grants + source ACL mirroring in Genie Search	Signed-in Salesforce user profile & field-level security	Unified principal token checked across SQL, vector & graph stores
Private / air-gapped portability	GCP-managed service	Cloud lakehouse service (AWS/Azure/GCP)	Salesforce multi-tenant cloud	Runs on customer hardware, private VPC (`Postgres` + `pgvector` /graph), or cloud

Architectural recommendation: federate platform catalogs behind one thin contract

Do not copy data out of BigQuery, Databricks, or Salesforce just to rebuild what their native catalogs already do well. Instead:

- Let `BigQuery measures` or `Unity Catalog` govern warehouse metric SQL where you already license them.
- Let `Salesforce Catalog` govern CRM object schemas and field-level permissions.
- Build a **thin, vendor-neutral ontology registry** (often 100 to 300 certified YAML/JSON entries backed by `Postgres`) that maps your company's canonical entities (`Customer`, `Contract`, `Product_SKU`, `Distributor_Tier`), business terms, aliases, authoritative system pointers, freshness rules, and abstention conditions across all of them.

Chapter 3: Why unchecked agent memory degrades accuracy and leaks private context

Once an organization connects an AI assistant to governed definitions and retrieval, the next demand from executives and power users arrives immediately: *persistent memory across sessions.* People want an assistant that remembers ongoing projects, prior exceptions, territory rules, and recurring workflows without starting cold every morning.

Naive implementations store conversation summaries in a vector table and retrieve the top five past notes into every new prompt. Within weeks, that design creates **memory poisoning**, **stale policy drift**, and **cross-user confidentiality breaches**.

Five primary studies published between November 6 and November 10, 2026 provide the strongest empirical blueprint to date for engineering enterprise-safe agent memory.

Five November 2026 breakthroughs in structured agent memory

1. **Offline log consolidation into linked wikis (`Anthropic`):** In Components of agent memory (November 6), Anthropic evaluated agents operating over workspaces up to 350,000 files and 420 million tokens. Rather than only appending live chat notes, an offline "dreaming" process periodically reviews interaction logs, fixes contradicted notes, and builds linked topic pages. In Anthropic's author-reported internal evaluations across 2 to 128 sessions, structured wiki memory raised graded document-question criteria from **13% to 48%**, and combining wiki memory with live file search raised a multi-document merger synthesis task from **73% to 87%** while saving roughly **20,000 tokens per question**. Anthropic explicitly notes two operational caveats: offline consolidation costs background compute, and consolidated notes can still become wrong when underlying source files change.
2. **Mandatory back-pointers to raw logs (`WikiTrace`):** In WikiTrace (November 10 preprint), Lee and colleagues directly target *memory-induced degradation* - cases where an agent performs worse *with* memory than without it because an earlier summary oversimplified or misread an outcome. `WikiTrace` links every distilled insight to intermediate trajectory summaries and raw execution logs so the agent verifies high-stakes claims against original evidence before acting, outperforming existing memory baselines across five benchmarks. A related November 9 preprint on Self-Organized Memory similarly reported gains on `LoCoMo`, `BEAM-100K`, and `LongMemEval-S` when agents revise linked topic documents and check original interactions, noting that file-based structured memory transfers cleanly when swapping underlying models.

3. **Four-scope privacy isolation (`MixUA`):** In MixUA (November 7 preprint), Wang and colleagues address multi-user enterprise environments where multiple employees interact with multiple specialized agents. By classifying distilled lessons into four strict sharing scopes - **user-agent pair**, **user**, **agent**, and **global** - `MixUA` improved task pass rates by up to **16 percentage points** over no-memory baselines while holding forbidden-scope private-fact detections to **0.02%** in a five-user, seven-agent study. (Treat the five-user scale as an early benchmark proof of concept that must be backed by hard database access controls in production.)
4. **Hypothesis testing before promotion (`Falsification-Gated Self-Evolution`):** In a November 10 preprint (arXiv:2611.14326), Tang and colleagues propose holding experience-derived lessons in temporary state as structured hypotheses and testing whether they cause regressions before committing, refining, or rejecting them.
5. **Quota-controlled retention and scheduled forgetting (`H-Mem`):** In an ACM publication on November 6, Wu and colleagues introduced `H-Mem`, using cluster-aware retention quotas to prune redundant notes under a strict **25% memory budget**, raising `LoCoMo` Qwen QA from **42.6 to 47.4** and `Evidence Survival@B` from **41.3 to 61.1** over a matched baseline.

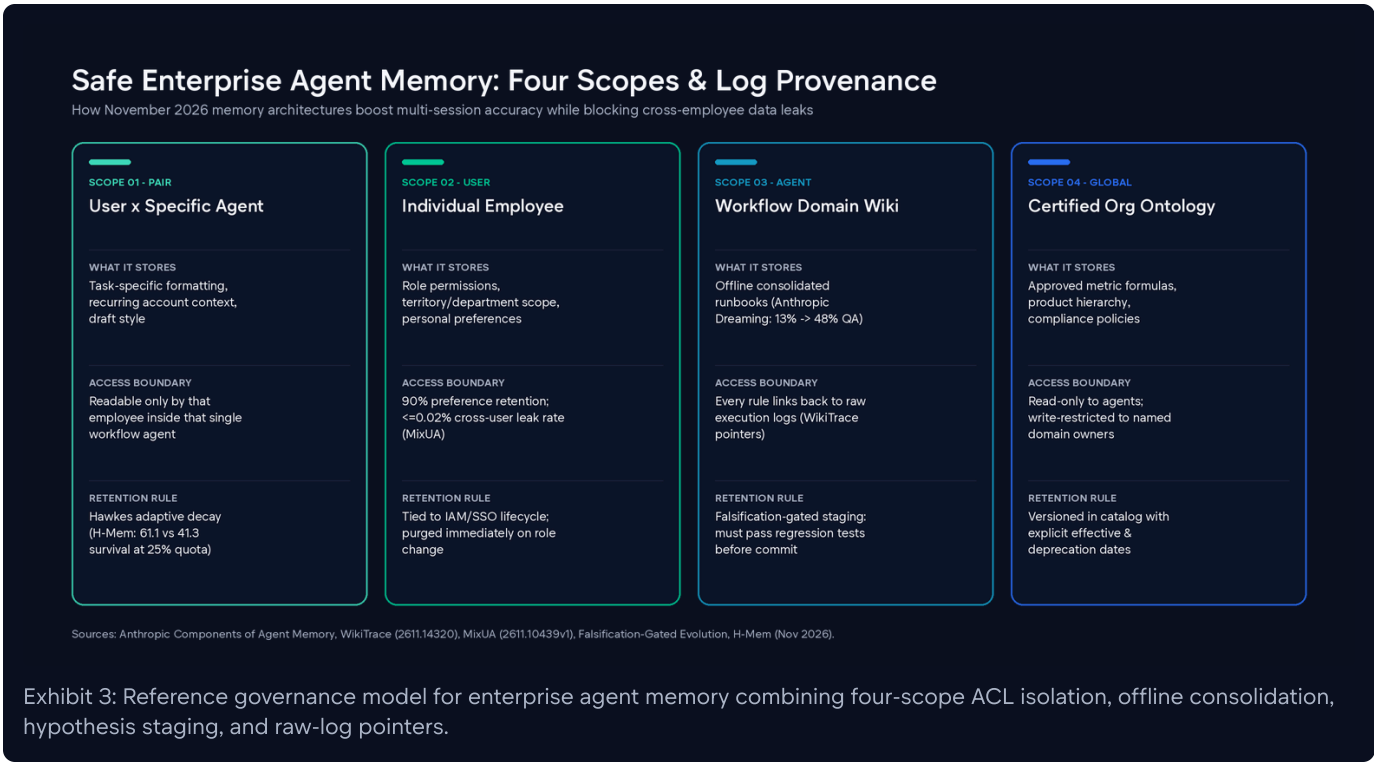


Exhibit 3: Reference governance model for enterprise agent memory combining four-scope ACL isolation, offline consolidation, hypothesis staging, and raw-log pointers.

Chapter 4: Reference blueprint - building a permission-mirrored ontology and memory spine

How do these pieces fit together in a practical engineering build that can ship in weeks rather than quarters, either in your cloud account or inside a private VPC?

ALLTIPLY's reference architecture separates **read-time query routing** from **write-time memory governance** across five deterministic stages.

Stage 1: Intent and entity pre-routing (before search runs)

Every incoming question or workflow trigger passes through a lightweight classifier that records three fields:

1. **Query intent** (`definition` , `reference_lookup` , `metric_calculation` , `how_to` , `troubleshooting` , or `multi_doc_synthesis`). Single-passage lookups run fast single-shot hybrid retrieval (`max_iterations = 1`) and skip unnecessary multi-hop planning; complex synthesis gets a larger token budget and multi-source checks.
2. **Canonical entity and term resolution.** User shorthand and acronyms are matched against your certified Ontology Registry. If a question asks for "net retention" or "preferred distributor credit," the router attaches the certified definition, authoritative table or policy ID, and required parameters (`account_id` , `effective_date`).
3. **Caller principal and scope token.** The user's identity, role groups, and tenant ID are bound to the request context outside the prompt so every downstream SQL view, vector query, and memory read filters at the database layer.

Stage 2: Permission-mirrored hybrid retrieval with page and freshness filters

- Structured metrics execute only against certified semantic views (`BigQuery measures` , `Unity Catalog` , or governed read-only SQL templates) rather than unconstrained model-generated SQL joins.
- Document search runs hybrid full-text + domain-tuned vector retrieval (`Postgres` FTS + `pgvector` /HNSW or your governed lakehouse index), filtering out `DEPRECATED` or expired documents and slicing targeted PDF page ranges where document structure is known.
- **Explicit specificity gate:** Before generating an answer, a check verifies whether the retrieved passages actually resolve the specific entity or tier asked about. When passages state only a category rule and a required entity attribute is missing, the system asks a targeted clarifying question or states the gap plainly instead of guessing.

Stage 3: Four-scope memory read and staging write

During execution, the assistant can read approved notes from four isolated partitions (`pair` , `user` , `agent` , `global`) matching the caller's principal token. When a session finishes:

- User-specific preferences write only to `pair` or `user` scope with a default TTL (for example, 90 days unless reconfirmed) and a user-visible delete button.
- Candidate workflow or domain lessons **never write straight to `agent` or `global` scope**. They land in a `memory_candidates` staging queue with PII scrubbed and `source_log_id` attached.

Stage 4: Offline consolidation, falsification check, and owner sign-off

On a scheduled nightly or weekly batch:

1. An offline worker groups staged candidates, checks whether underlying source documents or schemas have changed since `source_log_id` was recorded, and discards contradicted notes.
2. Remaining candidates run against your held-out regression suite (100+ domain questions, stratified by intent and product area). Any candidate memory that lowers retrieval or answer accuracy is rejected automatically.

3. Surviving domain updates surface in a lightweight review queue where the accountable data or process owner approves promotion to `agent` or `global` scope.

Stage 5: Separate retrieval, memory, and generation metrics in CI

An end-to-end "helpfulness" score hides decay until users lose trust. Track four numbers separately on every index, model, or memory change:

- **Retrieval Recall@K and MRR@K** by product area and intent.
- **Specificity & abstention calibration** on category-only trap questions (testing directly for the `CLADE` 50%-wrong-guess failure).
- **Cross-scope leak rate** across synthetic multi-user red-team sessions (target: `0.00%` unauthorized private-fact recall, enforced via SQL row-level security).
- **Evidence pointer validity** (% of active memory items whose underlying `source_log_id` or policy document remains current and non-deprecated).

Implementation roadmap: first 30 days

Week	Deliverable	Accountable owners	Pass criteria
Week 1	Pick one high-value workflow; document 25-40 canonical business terms, aliases, formulas, owners & abstention rules	Domain Lead + Data/AI Engineer	Zero unresolved collisions among top 30 workflow terms
Week 2	Build 100-question stratified evaluation set (including 20 category-only specificity traps & 20 permission-boundary tests)	Domain Reviewers + Engineering	Baseline Recall@K, specificity error rate & leak rate recorded
Week 3	Wire Ontology Registry + intent router + row-level permission filters into retrieval; connect existing cloud catalogs where present	AI Engineering + Security	Specificity guessing cut; 100% ACL parity on test principals
Week 4	Enable 4-scope memory table with `source_log_id` pointers, staging queue, TTL expiration & CI release gate	AI Engineering + Operations Owner	Zero cross-user leaks; automated CI block on eval regressions

---

Primary sources and evidence register (November 2026)

Every external benchmark and product capability cited in this report is attributed to its primary publisher. Preprints, vendor announcements, and preview features should be validated on your own corpus and workloads before production rollout.

1. Rayeed, S. M., Stewart, C., & Ma, J. (2026-11-10). *CLADE: Specificity collapse and ontology-assisted recovery in retrieval-augmented generation* (arXiv preprint). https://arxiv.org/abs/2611.14355
2. *DeepFinance: Can AI Diligence a $100MM Deal?* (2026-11-07, arXiv research publication). https://arxiv.org/html/2611.10926v1

3. Google Cloud (2026-11-10). *BigQuery measures GA: Unifying BI metrics and knowledge graphs*. https://cloud.google.com/blog/products/data-analytics/bigquery-measures-ga-unifying-bi-metrics-and-knowledge-graphs/
4. Databricks (2026-11-09 & 2026-11-11). *Genie Ontology, Genie Search, and Unity Catalog Pages GA - November 2026 Release Notes*. https://docs.databricks.com/aws/en/release-notes/product/2026/november
5. Databricks (2026-11-11). *Enterprise context in action: Genie Agents analyze more data sources*. https://www.databricks.com/blog/enterprise-context-action-genie-agents-now-analyze-more-data-sources-transform-your-work
6. Salesforce (2026-11-10). *Salesforce Catalog and Metadata Grounding MCP Server for AI Agents*. https://www.salesforce.com/blog/salesforce-catalog-ai-agents/
7. Snowflake (2026-11-10). *FL_FILTER_PAGES support in AI_EXTRACT and AI_PARSE_DOCUMENT*. https://docs.snowflake.com/en/release-notes/2026/other/2026-11-10-fl-filter-pages
8. Amazon Web Services (2026-11-09). *Single-shot and agentic retrieval with Amazon Bedrock Managed Knowledge Base*. https://aws.amazon.com/blogs/machine-learning/single-shot-and-agentic-retrieval-with-amazon-bedrock-managed-knowledge-base/
9. Neo4j (2026-11-09). *Agentic GraphRAG and graph-explorer-kit*. https://neo4j.com/blog/developer/agentic-graphrag-and-an-explorer-for-your-own-stack/
10. Anthropic (2026-11-06). *Components of agent memory*. https://www.anthropic.com/research/components-of-agent-memory
11. Lee et al. (2026-11-10). *WikiTrace: Evidence-linked agent memory with trajectory and raw-log verification* (arXiv preprint). https://arxiv.org/abs/2611.14320
12. *Self-Organized Memory for Long-Horizon Agents* (2026-11-09, arXiv preprint). https://arxiv.org/abs/2611.11929
13. Wang et al. (2026-11-07). *MixUA: Scoped lesson sharing across multi-user, multi-agent trajectories* (arXiv preprint). https://arxiv.org/abs/2611.10439v1
14. Tang et al. (2026-11-10). *Falsification-Gated Self-Evolution for Experience-Derived Agent Memory* (arXiv preprint). https://arxiv.org/abs/2611.14326
15. Wu et al. (2026-11-06). *H-Mem: Cluster-aware retention control under memory budgets* (ACM Digital Library). https://dl.acm.org/doi/10.1145/3799682.3840962

Related ALLTIPLY Labs research and engineering scope

- **Companion Insights:** Define your business terms before you add more documents to AI and What should an enterprise AI assistant remember?
- **Prior ALLTIPLY Labs Field Reports:** Intent-aware retrieval for enterprise RAG, Rebuilding vs tuning stalled AI systems, and Product help assistant case study
- **Engineering Scope:** Custom AI Development, Private AI Infrastructure, and Enterprise knowledge access boundaries

About ALLTIPLY Labs

ALLTIPLY is an applied AI lab and engineering company based in Atlanta. We design, build, and run custom knowledge assistants, voice AI, and cross-system automation for mid-size and large organizations, deployed in your cloud or on your own infrastructure when data residency requires it. To review a knowledge assistant, ontology, or agent memory architecture with our engineering team, book a call or send a short project brief.

Work with ALLTIPLY's Engineering Team: Schedule an architecture review at www.alltiply.com/company/schedule-your-call or submit a scoped brief at www.alltiply.com/company/project-planner.