Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enterprise Knowledge Agents Using Vector Embeddings and Persistent Memory Architectures

Santosh Kumar Vangapelli

Independent Researcher, USA

ARTICLE INFO

ABSTRACT

Purpose. Enterprise knowledge access fails structurally because useful organizational information is fragmented across heterogeneous systems, expressed in inconsistent vocabulary, and restricted by real security boundaries. This article examines why conventional enterprise search and short-context language model assistants fail and proposes an integrated five-component reference architecture combining semantic vector retrieval, structured persistent memory, and policy-governed access control enforced before retrieval begins.

Design/methodology/approach. The article employs a structured synthesis of peer-reviewed and preprint literature spanning information retrieval, retrieval-augmented generation (RAG), large language model (LLM)-based autonomous agents, knowledge graphs, enterprise knowledge management, and dense passage retrieval to derive architectural requirements and the reference design presented.

Findings. Five structural failure modes of conventional approaches are identified. A ten-property comparative analysis confirms no existing architectural pattern jointly addresses semantic retrieval quality, cross-session continuity, and pre-retrieval access governance simultaneously. The proposed five-component architecture — retrieval, memory, authorization, orchestration, and response layers — addresses all ten properties.

Originality/value. The article advances the position that enterprise knowledge agents must be designed as long-lived, governed organizational systems rather than stateless retrieval wrappers. The integrated architecture and access control enforcement framework fill a gap in existing literature. The ten-property architecture comparison provides the first structured evidence that the integration gap is real across all three dimensions simultaneously.

Research limitations/implications. The architecture is derived from structured literature synthesis and has not been validated against a primary empirical enterprise deployment. Future empirical case study research should measure session continuity improvement, retrieval precision under role constraints, and reduction in repeated investigative effort.

Keywords: Enterprise Knowledge Management, Vector Embeddings, Retrieval-Augmented Generation, Persistent Memory, Access Control, Large Language Models, Knowledge Graphs

1. INTRODUCTION

Modern enterprises generate vast internal knowledge across documentation systems, engineering design reviews, incident postmortems, issue trackers, operational dashboards, policy repositories, and collaboration platforms. Alavi and Leidner (2001) defined this challenge precisely: knowledge management systems must support the creation, storage, retrieval, transfer, and application of organizational knowledge, and the central obstacles are architectural — how knowledge is indexed, who may access it, how it stays current, and how different systems are made to work together. These obstacles have not been resolved by a generation of enterprise search systems.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Large language models (LLMs) have brought renewed attention to the enterprise knowledge problem. Ouyang et al. (2022) demonstrated that instruction-tuned LLMs significantly outperform base models on knowledge-intensive generative tasks. Yet LLM parametric knowledge is static and bounded by training data. Kandpal et al. (2023) demonstrated this empirically: LLMs underperform specifically on long-tail knowledge — precisely the category most valuable in enterprise contexts. Service ownership conventions, incident-specific diagnostic paths, and team-level operational heuristics are exactly the long-tail knowledge that LLM parametric memory cannot reliably retain. External retrieval is therefore a structural requirement.

Two technical patterns address this requirement. Lewis et al. (2020) introduced retrieval-augmented generation (RAG), combining parametric LLM knowledge with non-parametric retrieval from external document stores. Izacard and Grave (2021) demonstrated that fusing multiple retrieved passages substantially improves answer quality over single-passage retrieval. Karpukhin et al. (2020) established the foundations of dense passage retrieval using learned dense representations, showing consistent improvements over sparse term-based retrieval. Klesel and Wittmann (2025) positioned RAG as the core enterprise AI design pattern, identifying metadata-aware reranking and source authority weighting as the dimensions separating enterprise-grade RAG from naive implementations. Wang et al. (2024) established that cross-session memory — along with planning and tool use — is the architectural component distinguishing capable enterprise agents from stateless chatbots.

This article makes three primary contributions. First, it identifies and characterises five structural failure modes of conventional enterprise knowledge access approaches. Second, it presents a semantic retrieval architecture combining vector embeddings with metadata filters, source authority weighting, freshness rules, and knowledge graph traversal. Third, it proposes a persistent memory architecture with four access-controlled namespace scopes, integrated within a five-component reference architecture in which access control is enforced pre-retrieval (Samtani et al., 2023), not post-hoc. Section 2 characterises structural failure modes; Section 3 analyses limitations of common architectures; Sections 4 and 5 present the retrieval and memory foundations; Section 6 argues for access control as a first-class architectural layer; Section 7 presents the integrated reference architecture; Section 8 addresses failure modes and recommendations; Section 9 concludes.

2. WHY ENTERPRISE KNOWLEDGE ACCESS FAILS STRUCTURALLY

The foundational challenge of enterprise knowledge is distribution. Organizational truth is not stored in one place or expressed in one form. Useful answers to practical questions — who owns this service, why is this threshold set this way, what did we do last time this symptom appeared — depend on joining fragments across systems that were not designed to work together. Alavi and Leidner (2001) identified cross-system integration and access governance as the structural challenges limiting enterprise knowledge management, and argued that knowledge management systems must model creation, storage, retrieval, transfer, and application as distinct concerns.

The vocabulary problem compounds the distribution problem. Croft et al. (2010) established the theoretical basis for keyword retrieval and made explicit its central assumption: relevance is a function of term overlap between query and document. This assumption fails systematically in enterprise settings, where different teams use distinct and partially non-overlapping vocabularies to refer to the same systems, processes, and incidents. Keyword retrieval cannot bridge these gaps without explicit synonym management that is itself costly to maintain.

Table I presents five structural failure modes of conventional enterprise knowledge access approaches, each grounded in the peer-reviewed literature. Authority flattening (Alavi and Leidner, 2001) is especially damaging when organisations use LLM-mediated summarisation, because the model draws on all retrieved content with approximately equal weight regardless of source authority. Access boundary violation (Samtani et al., 2023) is structurally dangerous because LLM-mediated summarisation can expose restricted content even when the original documents are not quoted directly: the model's output may be shaped by content it processed but did not cite.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table I. Structural failure modes of conventional enterprise knowledge access approaches

Approach	Failure Mode	Root Cause	Consequence
Keyword-based enterprise search	Lexical mismatch (Croft et al., 2010)	Relies on literal term overlap; ignores conceptual synonymy across teams and roles	Relevant content not retrieved when employee vocabulary differs from source vocabulary
Short-context conversational assistant	Continuity failure (Wang et al., 2024)	Treats every session as isolated; no structured retention of prior findings or validated context	Cumulative investigative work lost at session end; users repeat context-assembly effort each time
Retrieval-only agent (no memory)	Stateless rediscovery (Wang et al., 2024)	Fetches and summarises documents but retains nothing durable across interactions	No continuity improvement over time; diagnostic paths and validated findings re-established from scratch each session
Broad-retrieval, late-filter system	Access boundary violation (Samtani et al., 2023)	Retrieves from a wide index; security filters applied post-retrieval at model layer or not at all	Restricted content enters the retrieval candidate set before policy is enforced; leakage risk is structural
Authority-flattening index	Provenance loss (Alavi and Leidner, 2001)	Indexes all content without modelling source authority, freshness, or intended use	Policy documents, incident notes, and informal threads treated as equally authoritative; responses are semantically fluent but organisationally unreliable

3. LIMITATIONS OF COMMON ENTERPRISE ASSISTANT ARCHITECTURES

Three architectural patterns currently dominate enterprise assistant deployments, and each addresses only a subset of the five structural failure modes identified in Section 2.

The first pattern is keyword-based enterprise search with an LLM interface attached. The LLM provides natural-language interpretation and synthesis, but the underlying retrieval mechanism remains keyword-based. The pattern addresses neither the lexical mismatch failure mode nor the authority flattening failure mode. Gao et al. (2023) confirmed that naive RAG systems without semantic indexing and authority-weighted reranking fail specifically on semantically complex enterprise queries.

The second pattern is a short-context conversational assistant. These systems succeed at single-turn synthesis but fail at continuity — the defining requirement of enterprise investigative knowledge work. Wang et al. (2024) identified cross-session memory as the component most frequently absent from enterprise LLM deployments and the most consequential for practical knowledge work. Without cross-session memory, the diagnostic path established last month, the service alias that caused confusion last quarter — none of these survive the session boundary.

The third pattern is a retrieval-only agent that fetches documents, summarises them, and stops. This architecture addresses lexical mismatch through semantic indexing but does not address continuity failure, stateless rediscovery,

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

access boundary violation, or authority flattening. Klesel and Wittmann (2025) identified this pattern's central deficiency: retrieval quality in enterprise settings depends on how content is indexed, filtered, source-weighted, and provenance-tagged — not on embedding similarity alone.

4. VECTOR EMBEDDINGS AS A SEMANTIC RETRIEVAL FOUNDATION

Vector embeddings address the lexical mismatch failure mode by representing text and enterprise artefacts in a continuous semantic space rather than a discrete vocabulary space. Lewis et al. (2020) demonstrated that dense embedding-based retrieval consistently outperforms sparse term-based retrieval on knowledge-intensive tasks. Karpukhin et al. (2020) provided the empirical foundation with dense passage retrieval (DPR): using learned dense representations trained on question-passage relevance pairs, DPR substantially outperforms BM25 on open-domain question answering benchmarks and generalises effectively to heterogeneous document collections.

Multi-passage fusion extends single-passage retrieval to handle queries requiring cross-source synthesis. Izacard and Grave (2021) demonstrated that a Fusion-in-Decoder (FiD) architecture substantially outperforms single-passage retrieval on knowledge-intensive tasks, with performance scaling as additional passages are incorporated. For enterprise questions about service ownership, incident history, or architectural context, a single retrieved chunk is typically insufficient.

However, neither dense retrieval nor multi-passage fusion is sufficient by itself. Gao et al. (2023) identified the combination of dense retrieval with metadata-aware reranking as the distinguishing property of advanced RAG systems. For enterprise agents, this metadata layer must apply organisational context — source type, document ownership, recency, classification level, and permitted audiences — as filter predicates at query time. Source authority weighting calibrates relevance scores: formal policy documents and production system registries are weighted above engineering proposals; approved architecture decision records above informal discussion threads.

Knowledge graph integration extends enterprise retrieval capability to multi-hop organisational questions. Ji et al. (2022) established that knowledge graphs capture organisational entity relationships — service ownership, team membership, system dependency chains — that are absent from unstructured text and cannot be recovered through embedding similarity. Pan et al. (2024) showed that synergised architectures combining LLMs with knowledge graphs substantially outperform either component independently on multi-hop reasoning tasks. Lan et al. (2023) confirmed that complex knowledge base question answering requires explicit reasoning over entity relationships. Together, these findings establish that the enterprise retrieval layer must combine dense embedding retrieval, multi-passage fusion, metadata-aware reranking, and knowledge graph traversal to cover the full range of organisational question types.

5. PERSISTENT MEMORY ARCHITECTURES FOR ENTERPRISE AGENTS

Retrieval addresses the freshness and breadth dimensions of enterprise knowledge access. But retrieval alone does not address continuity — the requirement that useful context established in one session should be available in the next. Wang et al. (2024) established that cross-session long-term memory is the architectural component that most fundamentally distinguishes capable autonomous agents from stateless language models.

Simple transcript retention is a weak form of enterprise memory because it stores large amounts of conversational noise without extracting durable structure. More effective memory systems apply explicit extraction logic to identify what should be preserved: that a service has multiple informal names that cause confusion, that a particular symptom pattern has historically correlated with a specific dependency issue, that a prior investigation validated a diagnostic path that should be tried early in related future incidents. Wang et al. (2024) termed this structured memory extraction and identified it as the design principle separating useful agent memory from raw conversation logs. Table II formalises this principle as a four-scope memory taxonomy for enterprise agents, with each scope governed by distinct retention policies and access rules.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table II. Persistent memory type taxonomy for enterprise knowledge agents

Memory Type	Scope	What to Store	Retention Policy	Access Rule
Session memory	Single interaction	Transient context: current query thread, intermediate reasoning steps, and working references for this session	Cleared at session end; never automatically promoted to any long-term store	Scoped to authenticated user for that session only; not readable by any other party
User memory	Individual across sessions	Validated aliases, preferred source types, repeated investigative paths, and stable working context for this specific user	Persistent until explicitly invalidated; version-aware and provenance-linked	Read and write by authenticated user only; not accessible to other users or teams
Team memory	Shared within a security boundary	Shared validated findings, known entity mappings, team-specific operational conventions, and approved diagnostic paths	Persistent; requires explicit promotion by an authorised team member; expiry policy defined per namespace	Read by all team members within the security boundary; write requires promotion with a provenance record
Organisation memory	Stable authoritative anchors across the enterprise	Formal policy definitions, service ownership registry, architecture decisions, and authoritative reference data	Long-lived; changes require governance approval; prior version retained in full with audit trail	Read by all authorised organisation members; write restricted to designated owners subject to governance review

Memory governance is not optional in enterprise settings. Samtani et al. (2023) identified persistent memory as a source of security risk qualitatively distinct from retrieval risk: derived memory artefacts can carry the sensitivity of their source material across time and across the users who access the same memory namespace, even when the original documents are no longer surfaced in retrieval results. The four-scope partitioning in Table II mitigates this risk by ensuring that memory artefacts are stored in namespaces whose access rules are at least as restrictive as those of the source material. Kandpal et al. (2023) add a structural motivation for memory beyond continuity: because LLMs systematically fail on long-tail organisational knowledge not encoded in their parametric weights, persistent memory is the mechanism through which enterprise agents accumulate and retain organisational knowledge that LLM pre-training cannot reliably capture.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

6. ACCESS CONTROL AS A FIRST-CLASS ARCHITECTURAL LAYER

The access boundary violation failure mode is the most difficult to mitigate after the fact because it is structural rather than incidental. Samtani et al. (2023) characterised this as the central security challenge in LLM-mediated enterprise knowledge systems and established the necessary corrective principle: identity and authorisation must be resolved before semantic retrieval begins. When a user submits a query, the request carries access context — identity, group and role membership, tenant or business unit, and permitted tool actions — and this context must be resolved against the organisation's policy system before the retrieval layer selects any candidate content.

In vector-based retrieval systems, pre-retrieval enforcement has specific and important indexing implications. Pan et al. (2024) noted, in the context of enterprise knowledge graphs, that access governance must be enforced at the entity level to provide genuine security. The analogous requirement for vector retrieval is that index chunks carry security metadata — source document identity, ownership, tenant, classification level, permitted groups and roles — as structured attributes indexed alongside their embedding vectors. A chunk that the requesting user is not authorised to read must not enter the retrieval candidate set at all, regardless of how semantically relevant it might be. Table III maps five access control requirements to their correct enforcement points in the architecture.

Table III. Access control requirements for enterprise knowledge agents

Requirement	Enforcement Point	Why Post-hoc Enforcement Fails
Identity resolution before retrieval begins (Samtani et al., 2023)	Authorisation layer — pre-retrieval	If identity is resolved after retrieval, restricted documents enter the candidate set and may shape the model's response even when they are not quoted directly
Security metadata on index chunks (Pan et al., 2024)	Retrieval layer — index construction	Chunk-level attributes must be applied as filter predicates at query time, not as post-retrieval masks which still expose restricted content to the model
Memory namespace partitioning by security boundary (Samtani et al., 2023)	Memory layer — namespace design	A shared long-term memory store across users or teams creates structural cross-boundary leakage; derived summaries can expose restricted source content even without surfacing the original documents
Tool-level authorisation independent of document access (Wang et al., 2024)	Orchestration layer — tool invocation	A user authorised to read incident notes should not automatically be authorised to invoke production actions or write to shared knowledge stores; document and tool permissions must be governed as separate authorisation domains
Derived summary and memory artefact governance (Samtani et al., 2023)	Memory layer — promotion logic	A summary or memory artefact generated from restricted sources inherits the sensitivity of those sources; demotion to a lower-classification namespace cannot be assumed safe without explicit governance review

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Persistent memory introduces a second and distinct access challenge. A retrieval system that correctly enforces access boundaries at query time can still create leakage through memory: if a user synthesises a finding from restricted documents and that finding is stored in a broadly readable memory namespace, the restricted content has been effectively exported. Tool access requires a third, separate authorisation domain. Wang et al. (2024) established that tool use is one of the three core architectural components of capable LLM-based agents; document access and tool invocation are distinct authorisation domains that must be governed independently at the orchestration layer. Sequeda et al. (2025) demonstrated in an enterprise question answering deployment that knowledge graphs serve as a formal framework for evaluating the validity of LLM-generated queries against governed data — a pattern that extends naturally to the authorisation layer of the proposed architecture.

7. INTEGRATED REFERENCE ARCHITECTURE FOR ENTERPRISE KNOWLEDGE AGENTS

The five structural failure modes identified in Section 2 and the architectural requirements established in Sections 4 to 6 together define the design space for an effective enterprise knowledge agent. The proposed reference architecture addresses this design space through five components: a retrieval layer, a memory layer, an authorisation layer, an orchestration layer, and a response layer. Table IV compares this integrated architecture against the two weak patterns identified in Section 3 across ten properties.

Table IV. Architecture comparison across ten properties

Architectural Property	Keyword Search + LLM (Croft et al., 2010)	Retrieval-Only Agent (Lewis et al., 2020; Klesel and Wittmann, 2025)	Proposed Integrated Architecture
Semantic retrieval across heterogeneous sources (Lewis et al., 2020; Karpukhin et al., 2020)	No — lexical matching only	Yes — embedding-based similarity	Yes — embeddings + metadata filters + source authority weighting + freshness rules + knowledge graph traversal
Continuity across sessions (Wang et al., 2024)	No	No	Yes — structured persistent memory with four-scope namespace partitioning and provenance
Source authority and provenance (Alavi and Leidner, 2001)	Not modelled	Source logged but not weighted or graded	Source type, authority level, recency, and classification modelled explicitly; visible in every response
Pre-retrieval access control enforcement (Samtani et al., 2023)	No — post-hoc or none	No — post-hoc or none	Yes — identity resolution and policy enforcement before the retrieval layer is invoked
Chunk-level security metadata as filter	Not present	Not present	Yes — tenant, role, classification, and

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

predicates (Pan et al., 2024)			ownership attributes applied at query time
Memory namespace partitioning (Samtani et al., 2023)	None	None	Session / User / Team / Organisation boundaries enforced per Table II
Tool authorisation independent of document access (Wang et al., 2024)	Not applicable	Not applicable	Yes — tool-level authorisation enforced at orchestration layer separately from document retrieval
Derived content governance (Samtani et al., 2023)	None	None	Derived summaries and memory artefacts governed with the sensitivity classification of their source material
Knowledge graph integration for multi-hop reasoning (Pan et al., 2024; Lan et al., 2023)	No	No	Yes — knowledge graph traversal integrated with embedding retrieval for entity relationship and multi-hop organisational queries
LLM long-tail knowledge gap addressed (Kandpal et al., 2023)	No — relies on parametric memory alone	Partially — retrieval covers surface queries	Yes — retrieval + persistent memory covers long-tail and time-sensitive organisational knowledge not encoded in LLM weights

The retrieval layer indexes enterprise content using structure-aware embeddings (Karpukhin et al., 2020), with chunks carrying security metadata attributes applied as filter predicates at query time. Retrieval candidates are reranked using a combination of embedding similarity, source authority weights (Gao et al., 2023; Klesel and Wittmann, 2025), and document freshness scores. Knowledge graph traversal (Pan et al., 2024; Lan et al., 2023) is integrated for multi-hop organisational queries, and multi-passage fusion (Izacard and Grave, 2021) is applied in the response layer for queries requiring cross-source synthesis. The memory layer stores structured persistent knowledge across the four scopes defined in Table II, with explicit promotion logic requiring provenance records and namespace-appropriate authorisation for any durable write. The authorisation layer resolves identity and policy context before any retrieval or memory read is initiated (Samtani et al., 2023), and enforces tool-level authorisation at the orchestration layer independently of document retrieval permissions (Wang et al., 2024).

Two design principles are paramount across all five layers. The first is that retrieval and memory serve complementary purposes that cannot substitute for each other. Retrieval protects against staleness by consulting live or recently updated sources; memory protects against discontinuity by preserving validated context across sessions. Gao et al. (2023) identified this complementarity as the design rationale for advanced RAG architectures. The second principle is that provenance must remain visible throughout the pipeline, from retrieved chunk to memory entry to

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

final response. Ji et al. (2022) established that knowledge-based systems earn organisational trust through their ability to trace answers to identifiable, governed sources.

8. FAILURE MODES, RISKS, AND ORGANISATIONAL RECOMMENDATIONS

Even well-designed enterprise knowledge agents introduce risks that organisations must address proactively through governance. The first is semantic overreach: dense embedding retrieval can return conceptually similar but operationally irrelevant content. The mitigation is domain-specific metadata filtering layered over embedding similarity (Gao et al., 2023), not higher-quality embeddings alone. The second is stale memory: persistent memory that is not version-aware and provenance-linked will silently bias future answers as organisational reality changes. Memory must carry timestamps, version links to source documents, and explicit expiry or review schedules (Wang et al., 2024). The third is authority confusion: the failure mode reappears whenever provenance visibility is sacrificed for response fluency (Alavi and Leidner, 2001).

The fourth risk is security failure through three structural mechanisms identified by Samtani et al. (2023): broad indexing with late-stage filtering; shared long-term memory namespaces that allow derived findings from restricted sources to be read by unauthorised users; and coarse role-based authorisation. Pan et al. (2024) extended this analysis to knowledge graph contexts, noting that fine-grained, relationship-aware enforcement is required at the entity level. The fifth risk is LLM long-tail knowledge failure in systems that rely excessively on LLM parametric memory for organisational specifics. Kandpal et al. (2023) demonstrated that this failure is systematic and predictable.

Organisational recommendations follow directly from this analysis. Organisations should treat retrieval, memory, and authorisation as a single integrated architectural problem (Lewis et al., 2020; Wang et al., 2024; Samtani et al., 2023) — not three independent features assembled later. Practically, this means: investing in structured memory governance with explicit namespace partitioning before deploying persistent memory in a multi-user enterprise setting; layering metadata filters, authority weights, and freshness rules over embedding similarity before treating retrieval quality as a solved problem; enforcing identity resolution and chunk-level security predicates before retrieval rather than relying on model-layer filtering; integrating knowledge graph traversal for multi-hop organisational queries; and making provenance visible in every response so that users can assess source authority without relying on linguistic fluency as a proxy (Ji et al., 2022; Alavi and Leidner, 2001).

9. CONCLUSION

Enterprise knowledge access fails structurally because useful organisational information is fragmented across heterogeneous systems (Alavi and Leidner, 2001), expressed in inconsistent vocabulary (Croft et al., 2010), updated at different rates, governed by different authority levels, and restricted by real security boundaries (Samtani et al., 2023). These structural properties define five failure modes — lexical mismatch, continuity failure, stateless rediscovery, access boundary violation, and provenance loss — that conventional approaches address only partially and individually. The integrated five-component reference architecture proposed in this article addresses all five failure modes simultaneously through: vector embeddings with dense passage retrieval (Lewis et al., 2020; Karpukhin et al., 2020) and multi-passage fusion (Izacard and Grave, 2021) for semantic retrieval quality; metadata-aware reranking with source authority weighting (Gao et al., 2023) for provenance integrity; knowledge graph traversal (Pan et al., 2024; Lan et al., 2023) for multi-hop organisational reasoning; structured persistent memory across four access-controlled scopes (Wang et al., 2024) for session continuity; and pre-retrieval identity resolution with chunk-level security metadata (Samtani et al., 2023) for access boundary enforcement. The ten-property comparative analysis in Table IV confirms that no existing common architectural pattern addresses all properties simultaneously — this is the integration gap that the proposed architecture fills.

The central argument of this article is that enterprise knowledge agents should be designed as long-lived, governed organisational systems that combine retrieval, memory, and authorisation within a coordinated architecture — not as stateless retrieval wrappers with a language model interface attached. As enterprises accumulate increasingly complex internal knowledge environments, and as the empirical evidence of LLM parametric knowledge gaps

(Kandpal et al., 2023) makes external retrieval structurally necessary for reliable organisational knowledge access (Ouyang et al., 2022), this combination will become central to how knowledge work is performed at scale.

LIMITATIONS

This article has four limitations that future work should address. First, the five-component reference architecture is derived from a structured synthesis of peer-reviewed literature and has not been validated against a primary empirical enterprise deployment. Second, the failure mode catalogue in Table I is grounded in published literature rather than primary incident data; the relative frequency and severity of each failure mode in practice likely varies by organisational scale, sector, and knowledge management maturity. Third, the recommendations in Section 8 are most directly applicable to organisations with sufficient governance maturity and technical scale to invest in all five architectural components; smaller organisations may find that a subset provides adequate coverage for their current needs. Fourth, one cited source — Gao et al. (2023) — is an arXiv preprint rather than a peer-reviewed publication at the time of writing; it is retained because its findings on metadata-aware reranking are corroborated by the peer-reviewed sources cited alongside it, but readers should weigh this source accordingly.

REFERENCES

[1] Alavi, M. and Leidner, D.E. (2001), "Review: knowledge management and knowledge management systems: conceptual foundations and research issues", MIS Quarterly, Vol. 25 No. 1, pp. 107-136, doi: 10.2307/3250961.

[2] Croft, W.B., Metzler, D. and Strohman, T. (2010), Search Engines: Information Retrieval in Practice, Addison-Wesley, Reading, MA, available at: https://ciir.cs.umass.edu/irbook/ (accessed 1 January 2024).

[3] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M. and Wang, H. (2023), "Retrieval-augmented generation for large language models: a survey", preprint, available at: https://arxiv.org/abs/2312.10997 (accessed 1 January 2024).

[4] Izacard, G. and Grave, E. (2021), "Leveraging passage retrieval with generative models for open domain question answering", in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Association for Computational Linguistics, Online, pp. 874-880, doi: 10.18653/v1/2021.eacl-main.74.

[5] Ji, S., Pan, S., Cambria, E., Marttinen, P. and Yu, P.S. (2022), "A survey on knowledge graphs: representation, acquisition, and applications", IEEE Transactions on Neural Networks and Learning Systems, Vol. 33 No. 2, pp. 494-514, doi: 10.1109/TNNLS.2021.3070843.

[6] Kandpal, N., Deng, H., Roberts, A., Wallace, E. and Raffel, C. (2023), "Large language models struggle to learn long-tail knowledge", in Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR, pp. 15696-15707, available at: https://arxiv.org/abs/2211.08411.

[7] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W.-t. (2020), "Dense passage retrieval for open-domain question answering", in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp. 6769-6781, doi: 10.18653/v1/2020.emnlp-main.550.

[8] Klesel, M. and Wittmann, H.F. (2025), "Retrieval-augmented generation (RAG)", Business and Information Systems Engineering, Vol. 67 No. 4, pp. 551-561, doi: 10.1007/s12599-025-00945-3.

[9] Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W.X. and Wen, J.R. (2023), "Complex knowledge base question answering: a survey", IEEE Transactions on Knowledge and Data Engineering, Vol. 35 No. 11, pp. 11196-11215, doi: 10.1109/TKDE.2022.3223858.

[10] 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. (2020), "Retrieval-augmented generation for knowledge-intensive NLP tasks", Advances in Neural Information Processing Systems, Vol. 33, pp. 9459-9474, available at: https://arxiv.org/abs/2005.11401.

[11] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. and Lowe, R. (2022), "Training language models to follow instructions with human feedback", Advances in Neural Information Processing Systems, Vol. 35, pp. 27730-27744, available at: https://arxiv.org/abs/2203.02155.

[12] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J. and Wu, X. (2024), "Unifying large language models and knowledge graphs: a roadmap", IEEE Transactions on Knowledge and Data Engineering, Vol. 36 No. 7, pp. 3580-3599, doi: 10.1109/TKDE.2024.3352100.

[13] Samtani, S., Zhao, Z. and Krishnan, R. (2023), "Secure knowledge management and cybersecurity in the era of artificial intelligence", Information Systems Frontiers, Vol. 25 No. 2, pp. 425-429, doi: 10.1007/s10796-023-10372-y.

[14] Sequeda, J., Allemang, D. and Jacob, B. (2025), "Knowledge graphs as a source of trust for LLM-powered enterprise question answering", Journal of Web Semantics, Vol. 85, p. 100858, doi: 10.1016/j.websem.2024.100858.

[15] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z. and Wen, J.-R. (2024), "A survey on large language model based autonomous agents", Frontiers of Computer Science, Vol. 18, Art. 186345, doi: 10.1007/s11704-024-40231-1.