

Article

Agentic Knowledge Curation Versus Full-Context Retrieval: An Empirical Study of Retrieval Failure Topology in Long-Context LLM Systems

Carlos A. Martín , Jesús M. Torres , Rosa M. Aguilar , Silvia Alayón *  and Manuel A. Bacallado 

Department of Computer and Systems Engineering, University of La Laguna,
38200 San Cristóbal de La Laguna, Spain; camartin@ull.edu.es (C.A.M.); jmtorres@ull.edu.es (J.M.T.);
raguilar@ull.edu.es (R.M.A.); mbacall@ull.edu.es (M.A.B.)

* Correspondence: salayon@ull.edu.es

Abstract

Karpathy's proposal to replace Retrieval-Augmented Generation (RAG) with plain-text knowledge bases maintained directly by language agents (Agentic Knowledge Curation) has gained traction in industrial applications yet lacks systematic empirical evaluation. To our knowledge, this study presents the first comparative evaluation of this paradigm using blind human assessment by two independent external reviewers. The corpus comprises the technical documentation of a production semantic search system for Spanish legal documents (11 files, ~19,300 tokens) alongside 100 questions verified against source code, distributed across direct retrieval, multi-document synthesis, and reasoning about absence. Claude 3.5 Sonnet was utilized as the reference model to isolate the retrieval architecture's effect. While overall accuracy was statistically indistinguishable between paradigms, agentic curation showed a significant advantage in direct retrieval over long documents. Conversely, error analysis revealed RAG false negatives associated with the lost-in-the-middle phenomenon, whereas agentic curation exhibited localized degradation in textual fidelity for queries requiring exact reproduction of formulas or sequences. These results characterize the differential error profiles of both paradigms, providing actionable design criteria for engineers managing technical documentation with language models in cloud environments.

Keywords: retrieval-augmented generation; agentic knowledge curation; agentic memory; blind human evaluation; positional bias; document knowledge management; agent-based intelligent systems



Academic Editors: Grigorios
Beligiannis and Georgios A.
Tsirogiannis

Received: 1 June 2026

Revised: 22 June 2026

Accepted: 30 June 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Retrieval-Augmented Generation (RAG) has become the dominant architectural paradigm for equipping Large Language Models (LLMs) with external, updatable, and auditable knowledge since its formalization by ref. [1]. Its adoption in production environments is widespread: when LLM-based systems must operate over proprietary knowledge (technical documentation, corporate repositories, regulatory corpora), the default architectural choice is a pipeline comprising document segmentation, embedding generation, a managed vector store, and similarity-based retrieval. This stack, extensively characterized by ref. [2], is now the de facto standard in both academia and industry.

Despite its success, RAG entails non-trivial operational costs and documented limitations. Managed vector stores introduce fixed monthly costs independent of usage,

embedding models incur token costs proportional to corpus size during reindexing, and ingestion pipelines require continuous development and maintenance. Two less visible but equally significant limitations also arise. First, knowledge opacity, where embeddings are not human-readable, complicating direct auditing, targeted correction, and verification without specialized tools. Second, the lost-in-the-middle effect identified by ref. [3], where model performance degrades when relevant information appears in intermediate positions within long contexts. This bias affects both direct long-context usage and, to a lesser extent, concatenated fragments in RAG pipelines.

Simultaneously, three recent developments challenge the presumed optimality of vector-based infrastructures. First, context windows in commercial models now exceed 100,000 tokens, with some reaching the million-token scale, making medium-sized corpora directly manageable. Second, the concept of the language agent has evolved beyond the ReAct framework [4] toward architectures where agents actively manage their own memory and knowledge base, e.g., MemGPT [5], A-MEM [6], Zep [7], as systematized by the survey presented in ref. [8]. Third, recent work on intra-RAG improvements, such as contextual retrieval [9]. RAPTOR [10] and GraphRAG [11] demonstrate that ingestion quality is as critical as retrieval mechanisms themselves.

1.1. Karpathy's Proposal and Its Product Instantiation

Within this context, Karpathy [12] proposes a radical simplification: replacing vector-based RAG infrastructure with a plain-text knowledge base, typically a Markdown wiki, maintained directly by the language agent. The core insight is that maintaining a technical knowledge base largely consists of structured administrative work (e.g., classification, summarization, updating, and organization), tasks at which modern LLMs excel. If an agent can read, synthesize, edit, and organize text files using basic filesystem tools (e.g., grep, find), then dedicated vector infrastructure becomes optional for medium-sized corpora.

This proposal is not merely theoretical. It is instantiated in practice in Claude Code, Anthropic's tool for software development assistance, where each repository includes a CLAUDE.md file that records design decisions, project conventions, and operational rules. This file serves as a shared, version-controlled, auditable memory accessible to both developers and agents. The model is similar to the one proposed by Karpathy: a structured, readable, and directly editable knowledge base, managed collaboratively by humans and agents.

We define agentic knowledge curation as the broader paradigm encompassing this approach, characterized by three principles:

1. Agent-mediated ingestion: the corpus is reorganized into self-contained, traceable text entries.
2. Deterministic retrieval: query-time access occurs via structured text navigation rather than vector similarity.
3. Localized maintenance: updates are performed through direct edits without embedding regeneration.

This paradigm aligns with agentic memory literature (MemGPT, A-MEM) in its core principle that agents are active constructors of knowledge structures rather than passive consumers of retrieved fragments. However, it differs in representation (structured plain text versus graphs or key-value stores) and the focus (a pre-existing documentary corpus versus accumulated experiential memory).

1.2. The Evaluation Gap

Despite the growing interest that the proposal has generated in industrial practice, and despite the momentum of the debate in technical forums and non-indexed publications, to

the best of our knowledge, peer-reviewed academic literature does not contain a systematic, empirical evaluation of the paradigm that uses real technical documentation with verifiable ground truth and external human review. Existing comparisons such as RAG vs. long-context approaches [13,14] rely on automated benchmarks and do not include explicit agentic curation conditions or human evaluation.

Similarly, agentic memory research (MemGPT, A-MEM, Zep) focuses on conversational and experiential tasks rather than technical document management. Intra-RAG improvements are evaluated against classical RAG but not against paradigms that eliminate vector infrastructure entirely [9,10].

This gap is particularly relevant for system engineers. Choosing between paradigms involves trade-offs across infrastructure cost, response quality, auditability, textual fidelity, and regulatory compliance. Without controlled empirical evidence, these decisions rely on intuition or non-transferable benchmarks.

This study does not compare agentic knowledge curation against a conventional embedding-based RAG pipeline. Instead, it contrasts agentic curation with a full-context retrieval condition in which the complete corpus is directly available to the model. This deliberately idealized baseline eliminates retrieval failures caused by chunking, embedding quality, and ranking, allowing the analysis to focus on the effects of knowledge organization itself.

1.3. Research Question and Contributions

This work addresses the following research question: What error profiles characterize agentic knowledge curation compared to full-context retrieval on technical documentation, and what design criteria can be derived for system engineers?

We operationalize this question through four testable hypotheses:

H1: *No detectable global performance difference in context-fitting corpora.*

H2: *Advantage in information access over long documents.*

H3: *Disadvantage in tasks requiring exact textual fidelity.*

H4: *Comparable absence detection, with fewer false negatives in agentic curation.*

The contributions of this work are as follows:

1. Conceptual formalization of agentic knowledge curation, distinguishing it from experiential agentic memory and intra-RAG mitigations, and situating it within the emerging taxonomy of agentic RAG.
2. First empirical evaluation using blind human assessment based on technical documentation of a system in production, with the ground truth verified against the source code.
3. Naturalistic observations consistent with the lost-in-the-middle effects in a full-context retrieval setting (identification of two retrieval failures confirmed by unanimous scoring).
4. Empirical observations suggesting a trade-off between semantic navigability and fine-grained textual fidelity.
5. Preliminary design considerations based on corpus characteristics, query types, and system constraints, including infrastructure costs, auditability, and regulatory compliance.

1.4. Article Organization

The remainder of the paper is structured as follows. Section 2 reviews related work on RAG architectures, positional biases, agentic memory, and evaluation methodologies.

Section 3 formalizes the agentic knowledge curation paradigm. Section 4 derives the research hypotheses. Section 5 describes the experimental design. Section 6 presents the results. Section 7 discusses implications and limitations. Section 8 concludes and presents future research lines.

2. Related Work

2.1. From Early RAG to Modular Architectures

The Retrieval-Augmented Generation (RAG) paradigm was formalized by ref. [2] as a two-stage architecture: a dense retriever selects relevant passages, and a sequence-to-sequence generator produces answers conditioned on the retrieved content. This approach builds on prior work in dense retrieval for open-domain question answering [15,16] and on sentence embedding methods based on Siamese networks [17]. Its key contribution was not retrieval per se, but demonstrating that coupling retrieval with generation enables language models to incorporate post-training knowledge without retraining.

Subsequent work has expanded this paradigm into increasingly sophisticated architectures. The authors of ref. [2] identify three main stages of evolution. Naïve RAG follows the original design with fixed-size chunking, single embeddings per chunk, cosine similarity retrieval, and simple prompt concatenation. Advanced RAG introduces optimizations such as query rewriting, routing, reranking, and noise filtering. Modular RAG further decouples components into interchangeable services, enabling more complex designs such as hierarchical retrieval (RAPTOR [10]), self-reflective retrieval (Self-RAG [18]), corrective pipelines (CRAG [19]), iterative retrieval-reasoning loops (IRCoT [20]), and adaptive routing strategies (Self-Route [13]).

Alongside these advances, the literature has documented structural limitations of classical RAG. Ref. [21] identifies recurring failure modes in production systems, including context fragmentation, retrieval of high-scoring but irrelevant passages, and difficulty recognizing missing information. Ref. [22], through the CRUD-RAG benchmark, shows that RAG systems perform significantly better on read operations than on create, update, or delete tasks, revealing a fundamental asymmetry in knowledge maintenance.

Operational costs further complicate the picture. Embedding models incur token costs proportional to corpus size during reindexing, vector stores introduce fixed infrastructure costs, and ingestion pipelines require ongoing maintenance. The opacity of embeddings adds an additional burden: since vector representations are not human-readable, auditing and correcting knowledge becomes indirect and tool-dependent. These limitations motivate reconsideration of whether vector-based infrastructures remain justified for medium-sized corpora under current model capabilities.

2.2. Positional Bias and the Lost-in-the-Middle Phenomenon

A key assumption underlying RAG is that the generator can effectively utilize all retrieved information. This assumption has been challenged by ref. [3], who systematically studied how language models process long contexts. Across two controlled tasks (multi-document question answering and key-value retrieval), they demonstrated that model performance depends strongly on the position of relevant information within the input. Their central finding is a U-shaped performance curve: models perform best when relevant information appears at the beginning or end of the context, and worst when it appears in intermediate positions. This effect, termed lost-in-the-middle, has proven robust across model families, including those explicitly designed for long-context processing.

This phenomenon has two important implications. First, long-context models do not eliminate retrieval challenges, they replace them with positional biases. Second, RAG pipelines are not immune: although retrieval reduces context length, concatenation of

multiple passages reproduces positional asymmetries at a smaller scale. Passages in intermediate positions receive less effective attention, even when individually relevant.

Subsequent work has refined these findings. Ref. [23] confirms degradation in reasoning tasks requiring distributed information across long contexts. Ref. [24] shows that the effect persists despite specialized long-context training, with only partial mitigation through attention reordering techniques. Ref. [25] further demonstrates that positional bias interacts with content features, with rare or salient terms partially offsetting degradation.

An underexplored implication is that lost-in-the-middle should manifest not only in synthetic benchmarks but also in naturalistic technical documentation, where relevant information often appears in non-prominent positions within long documents. If so, systems relying on attention-based access should exhibit systematic retrieval failures even when the relevant information is present within the available context. The present study examines this prediction by comparing a full-context retrieval condition with agentic knowledge curation, using external human evaluation to provide naturalistic observations that complement the synthetic findings reported by ref. [3].

Unlike conventional embedding-based RAG systems, the full-context retrieval condition used in this study eliminates retrieval-stage failures arising from document chunking, embedding quality, and ranking. Any observed failures therefore originate within the long-context processing stage itself, providing a conservative test of the effects of knowledge organization.

2.3. Intra-RAG Mitigations: Enriched Ingestion and Hybrid Retrieval

In response to the limitations of classical RAG, recent work has focused on improving performance within the paradigm itself. These approaches can be grouped into three categories: enriched ingestion, hybrid retrieval, and post-retrieval processing:

- Enriched ingestion addresses the loss of context caused by naïve chunking. Ref. [9] proposes contextual retrieval, where each chunk is augmented with a model-generated description of its role within the source document before embedding. This approach reduces retrieval failure rates significantly, especially when combined with lexical methods such as BM25. RAPTOR [10] takes a hierarchical approach, constructing recursive summaries that enable retrieval at multiple levels of abstraction. GraphRAG [11] extends this idea by representing documents as entity-relation graphs, enabling multi-hop reasoning over explicit connections.
- Hybrid retrieval combines dense semantic search with lexical retrieval methods, typically BM25 [26], using fusion techniques such as Reciprocal Rank Fusion [27]. The rationale is that dense retrieval captures semantic similarity, while lexical retrieval preserves exact matches critical for technical queries involving identifiers, versions, or code. Empirical evidence consistently supports this combination [9].
- Post-retrieval processing includes reranking, semantic deduplication, entailment-based filtering, and positional reordering to mitigate lost-in-the-middle effects. These techniques are largely complementary and can be combined. Anthropic reports that integrating contextual embeddings, hybrid retrieval, and reranking reduces failure rates by up to 67% compared to naïve RAG [9].

Despite their effectiveness, these approaches share a structural assumption: they seek to improve retrieval performance while preserving the underlying retrieval pipeline based on chunking, indexing, and retrieval mechanisms. The present study does not evaluate these architectures directly. Instead, it examines a different question: whether explicit knowledge structuring through agentic curation can alter information accessibility when compared with a full-context retrieval condition in which all source documents are directly

available to the model. This design isolates the effects of knowledge organization from those of retrieval quality and retrieval-stage optimization.

2.4. Agentic Memory and Knowledge Curation by Agents

Parallel to RAG development, another line of research has explored the role of language agents as active participants in knowledge management. The ReAct framework [4] established a foundational pattern in which models interleave reasoning steps with tool use. Subsequent work has extended this paradigm toward increasing autonomy. Self-RAG [18] introduces self-reflection to determine when retrieval is needed. FLARE [28] implements anticipatory retrieval triggered by low-confidence tokens. Search-o1 [29] integrates retrieval within extended reasoning chains. The authors of ref. [8] provide a comprehensive taxonomy of such systems, organized by agent structure, control flow, and knowledge representation.

More directly relevant are approaches that treat the agent as a builder and maintainer of its own knowledge base. MemGPT [5] introduces hierarchical memory inspired by operating systems, enabling agents to decide what to store and retrieve. A-MEM [6], inspired by the Zettelkasten method, allows memory entries to evolve, interlink, and generate their own contextual descriptions. Zep [7] models conversational memory as temporal knowledge graphs, while Mem0 [30] combines key-value storage with graph structures for persistent personalization.

Recent work has also explored the use of explicit intermediate structures to mitigate reasoning and memory limitations in LLMs. For example, GraphCogent employs coordinated multi-agent exploration over graph representations to support complex graph-understanding tasks [31]. Although its objective differs from the corpus-curation paradigm studied here, it reflects a broader trend toward external organization of information as a complement to direct context consumption. This perspective is consistent with the general hypothesis explored in the present work: retrieval performance depends not only on model capability and context size, but also on how knowledge is structured prior to access.

Agentic knowledge curation shares the core principle of these approaches (the agent as an active builder of knowledge structures, not as a passive consumer of retrieved fragments) but differs along three key dimensions:

1. Representation: it relies on structured plain text (e.g., Markdown) rather than graphs or specialized storage systems, enabling direct human readability and auditability.
2. Object of curation: it operates on pre-existing document corpora rather than accumulated experiential memory.
3. Genealogy: it originates from an industrial, production-oriented proposal [12], instantiated in tools such as Claude Code, rather than from academic memory architectures.

Although agentic knowledge curation shares certain characteristics with long-context retrieval, contextual retrieval, GraphRAG systems, and human-maintained documentation wikis, it differs from each in an important way. Long-context retrieval leaves the corpus unchanged and relies on attention during inference. Contextual retrieval enriches fragments while preserving the original corpus structure. GraphRAG reorganizes information into graph representations optimized for graph traversal. Documentation wikis provide structured knowledge but depend on manual maintenance. Agentic knowledge curation, by contrast, performs an autonomous ingestion-time transformation of an external corpus into a persistent, human-readable knowledge structure that subsequently becomes the primary retrieval substrate.

These distinctions support treating agentic knowledge curation as a separate paradigm rather than a variant of agentic memory systems. The present study does not evaluate whether agentic knowledge curation outperforms advanced retrieval architectures in gen-

eral. Rather, it examines whether explicit agent-mediated knowledge organization can improve information accessibility relative to a full-context retrieval condition in which the same source material is directly available to the model.

2.5. Evaluation of LLM-Based Retrieval Systems

The rapid expansion of RAG has led to a proliferation of evaluation frameworks. Automated methods based on LLM-as-a-judge, such as RAGAS [32], ARES [33], and TruLens, aim to quantify answer quality through metrics such as faithfulness, relevance, and coverage. While useful for rapid iteration, these approaches suffer from known limitations, including self-preference bias [34] and difficulty detecting subtle errors in technical domains.

Benchmark-based evaluations provide another approach. CRUD-RAG [22] evaluates systems across create, read, update, and delete operations on the corpus. LaRA [14] compares dense retrieval with long-context approaches, showing that optimal configurations depend on model size, task type, and corpus characteristics. The authors of ref. [13] demonstrate that long-context models can outperform RAG in quality under sufficient resources, while RAG remains more cost-efficient, motivating hybrid routing strategies. These findings suggest that retrieval quality is not the only determinant of system performance. Under some conditions, direct access to large contexts may constitute a viable alternative to retrieval-based architectures, although the mechanisms underlying such performance differences remain insufficiently understood.

However, these evaluations present three limitations relevant to this study. First, they rely predominantly on automated metrics. Second, they use generic benchmarks (e.g., Natural Questions, TriviaQA, HotpotQA) that do not reflect the heterogeneity of real technical documentation. Third, they do not examine whether explicit agent-mediated knowledge organization can alter information accessibility when compared with direct access to the same source corpus.

Human evaluation, particularly when blind and conducted by external reviewers, remains the gold standard for assessing open-ended tasks [35,36]. However, it is rarely applied in retrieval architecture comparisons due to cost and complexity, especially when combined with ground truth verification against source code or authoritative documentation.

An additional methodological gap concerns ecological validity. The authors of ref. [37] argue that findings obtained from generic corpora may not be transferable to naturalistic technical corpora with heterogeneous structures. This study addresses this limitation by using as its corpus the complete documentation of a real semantic search system in production, with the stylistic and structural heterogeneity characteristic of authentic technical corpora. Rather than evaluating a production retrieval pipeline, the study compares a full-context retrieval condition with agentic knowledge curation using the same underlying corpus. This design prioritizes ecological validity and isolates the effects of knowledge organization from retrieval-stage variables such as chunking, embedding quality, ranking, and retrieval failure.

2.6. Positioning of the Present Study

The preceding review highlights five gaps that this work addresses:

1. There is an ongoing debate regarding the conditions under which retrieval-based architectures remain preferable to direct long-context access [13,14], but the discussion has so far been framed largely as a binary comparison. Alternative approaches based on agent-mediated knowledge organization have received comparatively little empirical evaluation.

2. While lost-in-the-middle [3] has been extensively studied in synthetic settings, its manifestation in real-world technical documentation remains underexplored. This study provides naturalistic evidence through human evaluation.
3. Agentic memory research has grown rapidly over the past three years, but this research has focused on experiential and conversational tasks, rather than on the management of pre-existing technical document collection.
4. The documented retrieval-enhancement techniques, including contextual retrieval [9], RAPTOR [10] and GraphRAG [11], have primarily been evaluated within retrieval-based architectures. Comparatively little evidence exists regarding the extent to which explicit agent-mediated knowledge organization can influence information accessibility when retrieval-stage mechanisms are removed from experimental design.
5. Blind human evaluation with verified ground truth against source code is virtually absent in the literature, despite its importance for high-stakes technical applications. This methodological shortcoming is particularly acute in areas where the factual accuracy of responses has direct operational implications, such as technical documentation for systems in production, support for administrative decision-making, or document management in regulated environments.

This work contributes a focused empirical study addressing these gaps through a combination of (i) a naturalistic corpus (the complete technical documentation for a semantic search system in production), (ii) verified ground truth (questions based directly on the source code), (iii) controlled comparison between a full-context retrieval condition and agentic knowledge curation, (iv) a single reference model for isolating the effect of the generative architecture, and (v) blind evaluation by two independent reviewers.

This design does not constitute an evaluation of production-grade retrieval pipelines. Rather, it isolates the effects of knowledge organization by comparing agentic curation against a full-context retrieval baseline in which all source information is directly available to the model.

3. Conceptual Framework: The Agentic Knowledge Curation Paradigm

3.1. Definition

We define Agentic Knowledge Curation as an architectural paradigm for document-based knowledge management assisted by language models, characterized by four core properties:

1. Agent-mediated ingestion: the original document corpus is processed by a language agent during a deliberate ingestion phase. The agent reads the full corpus, identifies relevant semantic units, determines an organizational taxonomy, and produces a structured representation of the knowledge.
2. Structured plain-text representation: the output of ingestion is a collection of textual entries typically in Markdown format, organized hierarchically with cross-references. These entries are conceptually self-contained, explicitly traceable to their sources, and directly readable by humans without specialized infrastructure.
3. Deterministic textual navigation: query-time access relies on classical textual operations (path-based reading, index consultation, and pattern matching) rather than vector similarity. Retrieval is agentic in orchestration but deterministic in execution, ensuring traceability of each step.
4. Localized maintenance: knowledge updates are performed through direct editing of affected entries, without embedding regeneration or vector store reindexing. The knowledge lifecycle is thus decoupled from specialized ingestion infrastructure.

Taken together, these properties define a family of systems rather than a single system. Design choices such as entry granularity, taxonomy structure, navigation rules, and coverage policies remain open parameters. The common denominator is the shift in computational effort: semantically intensive processing is moved to ingestion, where the agent operates with full corpus access, while query-time operations become simpler, deterministic, and low-cost. This distinction is particularly relevant for the present study, which examines whether such ingestion-time knowledge organization can influence information accessibility when compared with direct full-context access to the original corpus.

The most direct instantiation of this paradigm is the architecture proposed by ref. [12] and implemented in Claude Code [38], where a CLAUDE.md file serves as shared project memory. Related systems, with varying degrees of alignment to the four defining properties, include agentic memory frameworks for personal assistants (e.g., Mem0 [30]) and project-oriented memory systems for coding assistants.

The paradigm proposed here should not be understood as a variant of GraphRAG or as a conventional documentation wiki. Unlike GraphRAG, it preserves knowledge primarily as human-readable text rather than graph structures. Unlike documentation wikis, its organization is produced through autonomous ingestion-time curation rather than manual editorial effort.

3.2. Components and Lifecycle

Agentic knowledge curation systems operate across three phases with asymmetric costs and requirements: ingestion, navigation, and maintenance.

The ingestion phase is the most resource-intensive in terms of both tokens and time. The agent processes the original corpus, extracts informational units (often guided by an explicit rule file encoding organizational conventions), designs a taxonomy ensuring full corpus coverage, generates the corresponding text entries with their traceability headers and sets up internal cross-references. A defining feature of this phase is active semantic compression and classification. The agent does not merely transform format; it reorganizes knowledge according to domain-relevant semantics, separating previously conflated information, making implicit conventions explicit, and preemptively resolving ambiguities.

The navigation phase is activated with each query. The agent typically begins from an index file (e.g., INDEX.md), traverses hierarchical structures, opens specific entries by path, and applies textual search (e.g., regular expressions) when direct navigation is insufficient. While the sequence of actions is agent-determined, each individual operation is deterministic and auditable. If no relevant path yields an answer, the system explicitly declares a knowledge gap.

The maintenance phase is activated when the corpus evolves or corrections are required. The agent identifies affected entries, edits them while preserving internal consistency, and records changes in a log (e.g., CHANGELOG.md). Unlike retrieval architectures that rely on embeddings and indexed vector stores, maintenance here consists of localized text edits, scaling with the number of affected entries rather than corpus size.

This shift toward ingestion-centered computation has three key implications. First, ingestion costs are substantial but amortized over subsequent queries, making the approach suitable for relatively stable corpora. Second, ingestion errors propagate to all downstream queries, as no correction occurs at inference time. Third, the paradigm explicitly trades retrieval-stage complexity for ingestion-stage organization. Consequently, performance depends not only on model capabilities but also on the quality of the curated knowledge structure produced during ingestion. This characteristic motivates the experimental design adopted in the present study. By comparing agentic curation with a full-context retrieval condition using the same

underlying corpus, it becomes possible to assess the extent to which ingestion-time knowledge organization influences downstream information accessibility.

3.3. Analytical Comparison with Related Paradigms

Agentic knowledge curation occupies a distinct position within a design space that includes retrieval-based RAG architectures, direct long-context access, retrieval-enhancement approaches, and agentic memory systems. Table 1 presents an analytical comparison based on five dimensions: locus of computational effort, knowledge representation, retrieval mechanism, update policy, and traceability.

Table 1. Analytical comparison of five knowledge management paradigms for LLM-assisted systems.

Dimension	Retrieval-Based RAG	Direct Long Context	RAG with Enriched Ingestion	Experiential Agentic Memory	Agentic Knowledge Curation
Focus of effort	Distributed: segmentation + inference	Focused on inference	Ingestion + inference	Continuous operation	Ingestion (agent-based semantic curation)
Representation	Vectorized fragments	Full corpus in prompt	Contextualized fragments + hybrid	Graph or key-value store	Hierarchical Markdown entries
Retrieval	Cosine similarity + top-k	Native attention	Dense hybrid + lexical with re-ranking	Agentic memory policies	Deterministic text navigation
Updating	Destructive reindexing	Context reloading	Partial reindexing	Agent-driven evolution	Localized editing
Traceability	By fragment, external metadata	Opaque, attention-dependent	By contextualized fragment	Varies by design	By input, readable explicit citation
Canonical references	[1,2]	[13]	[9,10]	[5,6]	[12,38]

Although Table 1 summarizes the principal architectural dimensions, two distinctions deserve particular emphasis. First, agentic knowledge curation differs from GraphRAG systems in that knowledge is reorganized into human-readable textual entries rather than graph entities and relations optimized for graph traversal. Second, it differs from software-engineering document-wiki approaches because the organizational structure is generated and maintained by the agent itself rather than manually authored by human contributors. In this sense, agentic knowledge curation can be interpreted as an intermediate paradigm between human-maintained documentation wikis and autonomous memory-management systems.

This comparison clarifies the paradigm's conceptual position. Relative to classic RAG, agentic knowledge curation replaces vector representations with structured text and shifts computation to ingestion. Relative to long-context approaches, it introduces explicit structural mediation that may reduce the impact of positional biases and information accessibility problems associated with large contexts. Relative to retrieval-enhancement approaches, it pursues a different design strategy by emphasizing knowledge reorganization during ingestion rather than optimization of the retrieval stage. Relative to agentic memory, it applies the principle of agent-driven knowledge construction to pre-existing document corpora.

The comparison presented in Table 1 is analytical rather than experimental. The empirical study reported in this article evaluates only one of these reference paradigms—full-context retrieval—against agentic knowledge curation. Consequently, the table should be interpreted as a conceptual positioning framework rather than as a summary of experimental comparisons.

3.4. Experimental Contrast Selection

Although five paradigms define the design space, this study adopts a deliberately constrained experimental comparison between two: full-context retrieval and agentic knowledge curation. This choice reflects the specific objective of isolating the effects of knowledge organization independently of retrieval-stage performance. This choice is motivated by three methodological considerations.

First, the contrast between two fundamentally different approaches to knowledge access. Full-context retrieval makes the entire source corpus directly available to the model and relies primarily on attention-based access mechanisms. Agentic knowledge curation reorganizes the same information into a curated knowledge structure designed during an explicit ingestion phase. Comparing these two conditions allows the study to isolate the contribution of knowledge organization while holding constant the underlying corpus and language model.

Second, the use of an idealized retrieval baseline. By providing the complete corpus directly to the model, the full-context retrieval condition eliminates retrieval failures arising from chunking strategies, embedding quality, ranking algorithms, and context-selection policies. Any observed performance differences therefore cannot be attributed to retrieval-stage errors, providing a conservative test of the value of agentic knowledge organization.

Third, analytical coverage of intermediate paradigms. Intermediate paradigms are well characterized in existing literature: long-context models [3,13], enriched RAG [9–11], and agentic memory [5,6]. Their positions within the design space are sufficiently characterized in the existing literature to serve as analytical reference points, even though they are not evaluated directly in the present experiment.

This design yields a dual interpretative perspective. At a narrow level, it provides empirical evidence regarding the relative performance of full-context retrieval and agentic knowledge curation under controlled conditions. At a broader level, it helps situate agentic curation within the wider landscape of knowledge-management architectures discussed in Section 2, although direct comparisons with intermediate paradigms remain a topic for future research.

3.5. Theoretical Predictions Derived from the Framework

The conceptual framework suggests three theoretical predictions that structure the hypotheses tested in Section 4:

1. Reversible compression (proposed explanatory construct): structured re-expression of a technical corpus by a competent agent preserves sufficient factual information to maintain comparable retrieval performance.
2. Structural visibility (proposed explanatory construct): by redistributing dispersed information into self-contained entries, agentic curation increases the accessibility of peripheral or non-prominent knowledge, potentially reducing the impact of positional biases such as lost-in-the-middle.
3. Deferred local fidelity (proposed explanatory construct): the same compression and synthesis processes that enhance visibility introduce localized degradation in tasks requiring exact reproduction of formulas, numerical parameters, or ordered procedures.

These predictions define an inherent trade-off: improved semantic navigability versus reduced fine-grained textual fidelity. Within the present study, these predictions are evaluated through comparison with a full-context retrieval condition using the same underlying corpus.

3.5.1. Unified Prediction: Retrieval Failure Topology

The three theoretical mechanisms proposed above can be interpreted as manifestations of a broader architectural prediction.

Rather than expecting one paradigm to be universally more accurate than the other, the framework predicts that agentic knowledge curation and full-context retrieval will exhibit different retrieval failure topologies. In particular, agentic knowledge curation is expected to reduce failures caused by positional invisibility, because ingestion reorganizes dispersed information into explicitly retrievable structures. At the same time, the semantic compression and synthesis inherent in ingestion are expected to increase localized fidelity degradation in tasks requiring exact reproduction of formulas, numerical parameters, or ordered procedures.

The central interpretative proposition of this study can therefore be stated as follows: we propose that Agentic Knowledge Curation may alter the topology of retrieval failures by exchanging positional invisibility for localized fidelity degradation.

Under this interpretative framework, the relevant question is not whether one paradigm is globally superior, but whether they fail differently, and whether those differences can be predicted from their underlying knowledge-organization mechanisms.

The four hypotheses tested in Section 4 should therefore be understood as operational consequences of this broader prediction. H1 evaluates whether the reorganization process preserves sufficient factual content to maintain overall performance. H2 tests the predicted reduction in positional retrieval failures. H3 tests the predicted increase in localized fidelity degradation. H4 examines whether the reduction in positional invisibility translates into fewer false-negative retrieval failures while preserving abstention behavior when information is genuinely absent.

3.5.2. Operational Definitions of the Analytical Constructs

The conceptual framework introduced above employs four analytical constructs that are central to the interpretation of the results: retrieval failure topology, structural visibility, reversible compression, and deferred local fidelity. These constructs are not proposed as formal mathematical variables, nor as established metrics in the retrieval literature. Rather, they serve as analytical dimensions intended to characterize systematic differences between knowledge-organization paradigms and to guide empirical comparison. They should therefore be interpreted as heuristic analytical constructs intended to organize empirical observations rather than as validated theoretical entities.

Retrieval failure topology refers to the distribution and qualitative structure of retrieval errors rather than their aggregate frequency. Two systems may exhibit similar average performance while failing on different classes of questions or under different representational conditions. In the present study, retrieval failure topology is operationalized through the pattern of scores across question categories, the occurrence of false negatives, the presence of fidelity-related errors, and the qualitative analysis of singular failure cases.

Structural visibility refers to the degree to which a piece of information is represented as an explicitly identifiable retrieval unit. Information embedded within long documents, tables, appendices, or non-prominent sections possesses lower structural visibility because successful retrieval depends on locating and attending to its position within a larger context. Information represented as a dedicated entry with an explicit title, location, and navigational path possesses higher structural visibility because it can be accessed directly through the organizational structure of the knowledge base. In this study, structural visibility is operationalized qualitatively through the contrast between information dispersed within the original corpus and information represented as self-contained entries in the curated knowledge base.

Reversible compression refers to the preservation of functional retrieval performance following semantic reorganization of the corpus. The concept does not imply perfect preservation of every textual detail. Instead, compression is considered reversible when the curated representation retains sufficient information to support retrieval performance comparable to that of the original corpus. In the present study, reversible compression is operationalized through the absence of a statistically detectable reduction in overall factual accuracy between paradigms.

Deferred local fidelity refers to localized degradation in the exact reproduction of information caused by ingestion-time synthesis and semantic compression. While semantic reorganization may improve navigability and retrieval accessibility, it can also reduce precision for content whose meaning depends on exact wording, symbolic notation, numerical parameters, formulas, configuration values, or ordered procedural sequences. Deferred local fidelity is operationalized through errors involving exact reconstruction of such content, particularly in the multi-document synthesis category.

Taken together, these constructs provide an interpretative framework for analyzing differences between knowledge-organization paradigms. The framework suggests that agentic knowledge curation alters retrieval failure topology by reducing failures associated with positional invisibility while increasing the risk of localized fidelity degradation. The hypotheses presented in the following section are motivated by this interpretative proposition and provide an empirical means of exploring its plausibility.

The purpose of these constructs is not to establish a new formal theory of retrieval systems, but to provide a vocabulary for describing and interpreting systematic differences observed between alternative knowledge-organization paradigms. Their formalization and validation remain topics for future research. Examples include measures of positional prominence, semantic redistribution, retrieval-path complexity, fidelity preservation, and retrieval-failure distributions across corpora.

4. Research Hypotheses

The hypotheses evaluated in this study are not independent claims. Rather, they constitute observable consequences of the broader theoretical prediction introduced in Section 3.5, namely that agentic knowledge curation alters retrieval failure topology by reducing positional invisibility while introducing localized fidelity degradation.

The four hypotheses operationalize different aspects of this interpretative framework and allow for its empirical evaluation through measurable retrieval outcomes:

H1: No detectable global performance difference in context-fitting corpora. When the document corpus fits entirely within the model's context window, the overall factual accuracy of agentic knowledge curation is statistically equivalent to that of a full-context retrieval condition implemented on the same model.

Rationale: this hypothesis directly operationalizes the reversible compression prediction (Section 3.5). If structured re-expression preserves a sufficiently large fraction of factual content, aggregate performance should not degrade. This expectation is consistent with findings by ref. [13], showing that long-context models can match or surpass RAG under sufficient resources, and with ref. [9], which identifies practical context thresholds beyond which vector infrastructure becomes optional. Under such conditions, agentic curation may offer comparable factual accuracy while relying on a fundamentally different mechanism for organizing and accessing knowledge.

H2: Advantage in information access over long documents. For queries requiring direct factual retrieval from long documents, agentic knowledge curation outperforms a full-context retrieval condition when relevant information is located in non-prominent or intermediate positions within the source documents.

Rationale: this hypothesis derives from the structural visibility prediction (Section 3.5). By reorganizing dispersed content into self-contained entries, agentic curation reduces dependence on original document position. This directly addresses the lost-in-the-middle phenomenon [3], whereby attention-based access degrades for information located in intermediate positions. Deterministic textual navigation over an explicitly structured representation may reduce the impact of positional bias, yielding a theoretical advantage in information accessibility. Among the four hypotheses, H2 is the most strongly grounded theoretically and is expected to provide the clearest empirical discrimination between the two experimental conditions.

H3: Disadvantage in tasks requiring exact textual fidelity. For queries requiring high-fidelity reproduction of formulas, numerical parameters, or exact procedural sequences, agentic knowledge curation performs as well as or worse than a full-context retrieval condition with direct access to the original corpus.

Rationale: this hypothesis operationalizes the deferred local fidelity prediction (Section 3.5). The semantic compression inherent in ingestion introduces localized information loss, particularly for expressions whose meaning depends on exact ordering or formatting. This trade-off is structurally linked to the advantage predicted in H2: the same transformation that enhances structural visibility may reduce verbatim precision in specific cases. H3 therefore represents the expected counterbalance to H2.

H4: Parity in absence detection with fewer false negatives. Both experimental conditions, when explicitly instructed to abstain in the absence of relevant information, achieve comparable performance in identifying queries whose answers are not present in the corpus. However, agentic knowledge curation exhibits a lower rate of false negatives (incorrectly declaring absence when relevant information is in fact present) particularly when that information appears in non-prominent positions.

Rationale: correct detection of absence is expected to depend primarily on instruction design, as shown in prior work on abstention policies (e.g., Self-RAG [18]) and should therefore be similar across paradigms. In contrast, false negatives depend on retrieval effectiveness and are expected to reflect the same positional biases underlying H2. The prediction is asymmetric: both systems should succeed in genuine absence detection, but agentic curation is expected to produce fewer false negatives when relevant information is present but difficult to access because of its position within large documents.

5. Materials and Methods

This section describes the reference system used as the source of the document corpus, characterizes the evaluation corpus, specifies the experimental design including model selection and the two experimental conditions (full-context retrieval and agentic knowledge curation), details the construction of the question set and ground truth, outlines the blind human evaluation protocol, and defines the statistical analysis plan. The full experimental protocol was designed to enable direct replication. The minimal artifacts required (question set, evaluation rubric, and configuration of both conditions) are described in the main text.

5.1. Reference System

The reference system for the experiment is a semantic search platform for Spanish legal documents, currently in use as a research tool for legal information retrieval, hereafter referred to as RJusticia. The choice of this system as corpus source is motivated by three methodological considerations. First, it is a real RAG system in active use, rather than an ad hoc experimental setup. Second, its technical documentation exhibits the stylistic heterogeneity typical of production systems (narrative documentation, configuration files, source code with docstrings, and README files) allowing for evaluation under realistic

conditions. Third, the experiment presents a methodological singularity that enhances ecological validity: the evaluation corpus is the documentation of a RAG system itself, situating the comparison between full-context retrieval and agentic knowledge curation in a context where knowledge-management design choices would directly impact system maintenance. The system architecture consists of five container-based services, as summarized in Table 2.

Table 2. Service architecture of the RJusticia reference system.

Service	Technology	Port	Function
postgres	PostgreSQL 17.6 + pgvector 0.8.1 + vchord_bm25 0.2.2	35432	Hybrid vector store and BM25
ollama	Ollama (nomic-embed-text 768d, embeddinggemma 768d)	11434	Local embedding generation
litellm-proxy	LiteLLM v1.77.5-stable	4000	Unified proxy for embedding models
backend	FastAPI/Python 3.13	8000	API REST with TEXT, SIMILARITY, BOTH
frontend	Next.js	3000	Web-based query interface

The search engine implements three retrieval modes:

- TEXT: BM25 with BERT-based tokenization over a lexical index.
- SIMILARITY: cosine similarity over L2 normalized vectors.
- BOTH (default): hybrid retrieval using Weighted Reciprocal Rank Fusion (WRRF) with a parameter of $k = 60$ and equal weights of 0.5 for each component. A methodologically relevant implementation detail is that the hybrid mode applies intersection rather than union of lexical and vector results: only documents appearing in both rankings are considered. This design increases precision at the cost of coverage.

The ETL (Extract, Transform, Load) pipeline exhibits a structural characteristic common to many embedding-based retrieval architectures: each execution drops and rebuilds the document table, regenerating all embeddings. Incremental updates are not supported. Consequently, any corpus modification requires full reprocessing, with cost proportional to corpus size. This characteristic illustrates one of the operational trade-offs associated with embedding-based retrieval architectures and provides practical context for the comparison conducted in this study.

5.2. Document Corpus

The evaluation corpus consists of 11 files from the RJusticia repository, selected to reflect the diversity of real-world technical documentation. It includes the following:

- One extended narrative user manual (~8000 words).
- Five README files for system components (etl, backend, frontend, ollama, despliegue general).
- Two deployment configuration files (docker-compose.yml, litellm_config.yaml).
- Three source code files with relevant docstrings (search_engine.py, pdf_utils.py, config.py).

The total corpus contains approximately 19,300 tokens, remaining comfortably within the context window of the model used in both experimental conditions.

The corpus has three key methodological properties. First, stylistic heterogeneity:

The corpus combines narrative prose, structured tables, source code, YAML configurations, and inline comments, avoiding the homogeneity bias of standard benchmarks. Second, uneven information distribution: the user manual accounts for approximately 40% of total corpus size, while the remaining files share the remaining 60%. This asymmetry is

methodologically relevant because it creates conditions under which information may be located in non-prominent regions of long documents, providing an appropriate setting for testing hypothesis H2. Third, authenticity: the corpus was not curated or simplified for the experiment. It was neither adversarially filtered nor stylistically normalized, thereby enhancing ecological validity in the sense discussed in Section 2.5.

Because the entire corpus fits within the model's context window, the experimental design permits direct comparison between full-context retrieval and agentic knowledge curation without introducing retrieval-stage constraints arising from corpus size.

5.3. Experimental Design

The experiment compares two conditions corresponding to the paradigms defined in Section 3.4 (Condition A: Full-Context Retrieval, Condition B: Agentic Knowledge Curation), both applied to the same corpus and executed using the same reference model.

The full-context retrieval condition does not implement a conventional retrieval pipeline based on chunking, embeddings, ranking, and context selection. Instead, the complete corpus is provided directly to the model. This design was chosen to isolate the effects of knowledge organization while eliminating retrieval-stage variability.

The design is paired: the same set of 100 questions is applied to both conditions under identical circumstances, enabling statistical comparison on matched samples.

5.3.1. Reference Model

Both conditions were executed using Claude 3.5 Sonnet as the reference model. A single-model design was adopted to isolate the effect of knowledge-access organization from generative capability. This choice ensures that observed differences can be attributed to knowledge organization and information-access strategy rather than model variation. Extending the study to multiple models is a natural direction for future work.

The selection of Claude 3.5 Sonnet is justified by three factors:

1. A 200,000-token context window, sufficient to accommodate the complete corpus in the full-context retrieval condition as well as the agentic workflow.
2. Strong documented capabilities in instruction following and tool use.
3. Stable commercial availability and public documentation, facilitating experiments replication.

5.3.2. Condition A: Full-Context Retrieval

In Condition A, the complete document corpus is provided directly to the model as context. No chunking, embedding generation, retrieval, reranking, or context-selection mechanisms are applied. The condition therefore represents full-context retrieval rather than a conventional Retrieval-Augmented Generation (RAG) pipeline.

This design was chosen to eliminate retrieval-stage variability and to isolate the effects of knowledge organization. Because all source information is available to the model, any retrieval failures observed in this condition originate from limitations in long-context processing rather than from document retrieval mechanisms.

The system prompt imposed three constraints:

- Responses must rely exclusively on the provided corpus.
- Absence must be explicitly indicated using [NO RESULT].
- Responses must follow a structured format (Answer, Source, Confidence, Time).

5.3.3. Condition B: Agentic Knowledge Curation

During the ingestion phase, an implementation-specific behavior was observed when processing large documents. In an initial ingestion attempt, the largest document in the

corpus (manual-usuario.md, approximately 8000 words or 10,000 tokens) was provided as an attached file. The resulting knowledge base exhibited incomplete coverage, with several sections of the source document not represented in the structured entries. The ingestion process was subsequently repeated in a second session in which the same document was embedded directly within the prompt. This produced a more complete representation, increasing the number of generated entries from 24 to 31. All experimental queries were executed exclusively on this second version.

Following ingestion, coverage was assessed through manual verification. The resulting wiki was compared against the source corpus to confirm that the major sections, subsystems, configuration files, procedures, and documented design decisions were represented in the curated knowledge base. This verification was qualitative rather than metric-based: no formal source-to-entry coverage matrix, automated diffing procedure, or quantitative completeness measure was employed.

This observation suggests that large attached documents may be subject to partial processing during ingestion, potentially due to internal truncation or sampling mechanisms. For replication purposes, documents exceeding approximately 8000 words ($\approx 10,000$ tokens) should preferably be provided as inline text, and coverage should be explicitly verified after ingestion.

More broadly, this observation suggests that ingestion robustness may constitute a critical dependency of agentic knowledge curation systems. Because knowledge organization is performed before query execution, incomplete ingestion can propagate systematically to all downstream retrieval tasks. The present experiment evaluates a single curated representation and does not quantify ingestion variability across prompts, formats, or models.

Furthermore, in Condition B, the knowledge base is the final wiki resulting from ingestion, containing 31 entries distributed across four thematic directories that emerged from agentic curation without prior instruction on a taxonomic schema: a core/ directory with four entries on fundamental concepts; a reference/ directory with 12 entries of comprehensive technical references; a procedures/ directory with 10 entries on operational procedures; and a context/ directory with five entries on design decisions and technical notes.

The system prompt specified the following:

- The agent role (knowledge management agent).
- The same source restrictions applied in Condition A, with the curated wiki serving as the sole admissible source of information.
- A deterministic navigation protocol with three levels:
 1. Index consultation (INDEX.md).
 2. Hierarchical navigation with links to specific tables of contents.
 3. Opening of specific entries, and, if the above routes prove insufficient, application of regex-based search (up to three query variants). If no information is found, the system returns [KNOWLEDGE GAP]. This mechanism serves the same abstention function as the [NO RESULT] response used in Condition A.

5.4. Question Set and Ground Truth

A set of 100 questions was developed, divided into three operational categories designed to assess complementary aspects of information access and knowledge utilization performance. Ground truth was verified a priori against source code and documentation, ensuring independence from model outputs.

5.4.1. Category 1: Direct Retrieval (n = 40)

Forty questions requiring exact factual retrieval from a single source (P001–P040). Notable examples include the PostgreSQL exposure port (35432), the LiteLLM proxy

version (main-v1.77.5-stable), the similarity SQL operators available in pgvector (`<=>`, `<->`, `<#>`, `<+>`), the WRRF *k* parameter value (60), and the embedding batch size (`EMBEDDING_BATCH_SIZE=32`). The category includes questions distributed across documents of varying length and prominence within the corpus, enabling assessment of potential positional effects on information accessibility.

This category evaluates factual information-access performance and provides the primary test bed for H2, particularly for questions whose answers are located in non-prominent regions of long source documents.

5.4.2. Category 2: Multi-Document Synthesis (n = 40)

Forty questions requiring integration across multiple sources (P041–P080). The complexity of this category lies not in the difficulty of the formulation but in the need to combine sources. Representative examples include the characterization of the destructive behavior of the ETL process and its operational implications, which combines information from `etl/README.md`, `docker-compose.yml`, and code migrations; the complete description of the WRRF mechanism with its exact scoring formula, which combines `user-manual.md` and `search_engine.py`; and the identification of startup dependencies between services determined by health checks, which combines `docker-compose.yml` and various README files. This category assesses the ability to integrate and reason across multiple sources. It also contains several questions requiring exact reproduction of formulas, parameters, or procedural details, making it particularly relevant for evaluating the deferred local fidelity prediction associated with H3.

5.4.3. Category 3: Absence Reasoning (n = 20)

Twenty questions whose correct answer is that the information is not present in the corpus (P081–P100). The questions were designed around aspects that are technologically plausible but not documented in the reference system: horizontal scalability limits, specific authentication mechanisms, high-availability configurations, multilingual support, concurrency thresholds, retention policies, and disaster recovery strategies. This category assesses abstention behavior and hallucination control, central to H4.

5.5. Blind Human Evaluation Protocol

The 200 generated responses (100 questions $\times$ 2 conditions) were evaluated by two independent external reviewers under blind conditions. Responses were anonymized and labeled only as System 1 or System 2, with no information on the correspondence between systems and experimental conditions. The mapping to experimental conditions was revealed only after scoring, eliminating expectation bias.

The two external evaluators participated voluntarily after being informed about the objectives of the study, the anonymous treatment of their assessments, and the intended use of the collected data for research and publication purposes.

Reviewers received the complete question set, the reference corpus, the ground-truth answer key, and written scoring instructions. Responses were anonymized and presented only as System 1 and System 2. Evaluators were instructed to verify factual correctness against the corpus rather than against the model-provided citations alone. Model-generated source references were visible but were not treated as evidence of correctness; evaluators independently checked the cited information against the corpus when necessary.

5.5.1. Evaluation Rubric

Responses were scored on a five-point ordinal scale, designed with explicit operational criteria to minimize interpretive variability among evaluators. Because the scale is ordinal, the primary statistical analyses reported in Section 6 use non-parametric paired tests

(Wilcoxon signed-rank), supplemented by descriptive statistics to facilitate interpretation. Table 3 summarizes the rubric. The full reviewer instructions and scoring rubric are provided in the repository, accessible at “<https://zenodo.org/records/20752159> (accessed on 22 June 2026)”.

Table 3. Five-point ordinal scale used in the double-blind human evaluation.

Score	Label	Operational Criterion
1	Incorrect	The response is factually incorrect; it reports [NO RESULT] or [KNOWLEDGE GAP] when information is actually present in the corpus (false negative), or it returns content irrelevant to Category 3 instead of acknowledging the knowledge gap.
2	Poor	The answer is partially correct but omits key elements or contains minor errors that affect the practical interpretation of the content.
3	Acceptable	The answer is correct but incomplete or not precise enough. An expert user could reconstruct the full information from it with some additional effort.
4	Good	The answer is correct and substantively complete, with a reference to the source and an appropriate level of technical accuracy.
5	Excellent	The answer is correct, complete, accurate, and includes explicit attribution to the source. For Category 3, it correctly states the absence of information and identifies the type of information not included in the corpus.

For each response, the final score was computed as the arithmetic mean of the two evaluators’ ratings. Means and standard deviations are reported as descriptive summaries to facilitate interpretation and comparison with prior LLM evaluation studies. Because the underlying scale is ordinal, inferential analyses are based on paired non-parametric tests (Wilcoxon signed-rank), with parametric results reported only as a robustness check.

5.5.2. Evaluators’ Qualification and Procedure

The two evaluators were researchers who were not involved in the development of the reference system or the design of the experiment, and who had sufficient technical expertise to verify responses concerning software architecture, retrieval systems, and technical documentation against the source code and the repository documentation. They were provided with written instructions and performed evaluations independently without communication. Average evaluation time ranged from 3 to 7 min per response.

Disagreements were not adjudicated. Each evaluator scored responses independently, and final scores were computed as the arithmetic mean of the two ratings. The purpose of the agreement analysis reported in Section 6.1 was to characterize the consistency of independent evaluations rather than to produce a consensus judgment.

5.6. Statistical Analysis

The analysis comprises three components: characterization of the inter-evaluator agreement, testing of hypotheses regarding differences in the two experimental conditions, and qualitative analysis of specific cases.

5.6.1. Inter-Evaluator Agreement

The agreement between the two raters is characterized by four complementary statistics, recognizing that no single statistic fully captures agreement in multi-level ordinal rating scales:

- Cohen’s κ : provides a random-error-corrected measure of agreement using the interpretive thresholds established by ref. [39]: values between 0.21 and 0.40 are interpreted as fair agreement, between 0.41 and 0.60 as moderate, between 0.61 and 0.80 as substantial, and above 0.80 as almost perfect.

- Pearson correlation: measures the linear relationship between the scores of both raters and is sensitive to the consistency of the direction, even if the absolute levels differ.
- Exact agreement percentage: calculates the proportion of pairs with identical scores.
- Agreement within ± 1 point: calculates the proportion of pairs with a maximum discrepancy of one level on the rubric.

5.6.2. Hypothesis Testing

Performance differences were initially evaluated using paired *t*-tests to facilitate comparison with prior LLM evaluation studies. Given the ordinal nature of the rubric and the observed ceiling effects, all analyses were additionally replicated using Wilcoxon signed-rank tests as a robustness check. The paired design is justified by the use of the same set of 100 questions for both conditions, which allows for the control of variables associated with the difficulty of each individual question.

The tests are run on the full dataset ($n = 100$) and on subsets by category ($n = 40$ for Categories 1 and 2; $n = 20$ for Category 3). In Category 3, where both conditions consistently achieve the maximum score according to the H4 prediction, a *t*-test is not calculated because the statistic is inapplicable when the variance is zero.

H1 is interpreted as absence of detectable difference rather than formal equivalence. Hypotheses H2, H3, and H4 make specific predictions regarding differences or asymmetries, which can be tested using standard significance tests with direct interpretation.

5.6.3. Qualitative Analysis

Quantitative results are complemented by qualitative analysis of the following:

- Extreme scores with full agreement.
- Cases with >2 point differences between conditions.

This qualitative analysis is particularly relevant for interpreting hypotheses H2 and H3: extreme cases allow us to identify the specific mechanisms responsible for the observed differences and to link the theoretical predictions in Section 3.5 to specific empirical patterns.

5.7. Replication Package and Artifact Availability

To facilitate independent verification and replication, all experimental artifacts underlying the study are provided in a public repository.

The replication package contains the following:

- The complete evaluation corpus used in both experimental conditions.
- The full set of 100 evaluation questions (P001–P100).
- Ground-truth annotations, including source-document references used for verification.
- The complete prompts and system instructions used in Condition A (full-context retrieval) and Condition B (agentic knowledge curation).
- The curated knowledge base generated during the ingestion phase of Condition B.
- The complete set of generated responses for both conditions.
- Individual scores assigned by each evaluator.
- The aggregated evaluation dataset used for statistical analysis.
- The scripts required to reproduce the reported descriptive statistics and hypothesis tests.

The repository is intended to support full methodological transparency. Readers can inspect every stage of the experimental workflow, from corpus construction and question design to response generation, scoring, and statistical analysis.

Because the study evaluates retrieval architectures through prompt-based interaction and human assessment, artifact availability is particularly important. Providing the complete experimental record allows independent researchers to verify reported findings,

examine alternative interpretations of individual cases, reproduce the statistical analyses, and extend the benchmark to additional models, corpora, or retrieval paradigms.

The repository also enables direct examination of the singular cases discussed in Section 6.4 (P028, P050, P052, and P055), which play a central role in the interpretation of the results.

Repository URL: “<https://zenodo.org/records/20752159> (accessed on 22 June 2026)”.

6. Results

This section presents the experimental results in the following order: characterization of inter-rater agreement, comparative performance between the two experimental conditions, both globally and by category, including statistical contrasts, detailed analysis of score distributions, qualitative examination of singular cases with unanimous extreme ratings, and formal evaluation of the four hypotheses in light of the empirical evidence.

6.1. Inter-Rater Agreement

Table 4 summarizes the four agreement indicators computed across the 200 evaluation pairs (100 questions $\times$ two conditions). Cohen’s κ is structurally appropriate for the two-reviewer design.

Table 4. Inter-rater agreement in the blind evaluation (n = 200 response-condition pairs).

Metric	Value	Interpretation
Cohen’s κ (E1 vs. E2)	0.38	Fair agreement according to ref. [39]; disagreements concentrate in three questions (P009, P013, P079).
Pearson correlation	$r = 0.52; p < 0.001$	Highly significant positive association; evaluators agree on score directionality.
Exact agreement	72.0%	144 of 200 response pairs received identical scores.
Agreement within ± 1 point	88.5%	177 of 200 response pairs differed by at most one rubric level.

The κ value of 0.38 falls within the conventional range of fair agreement. However, the practical agreement indicators provide a substantially more favorable interpretation: 72.0% exact agreement and, more importantly, 88.5% agreement within ± 1 point suggest that inter-rater variability was low in operational terms and concentrated in a small subset of questions. This apparent discrepancy between κ and practical agreement metrics warrants closer examination.

Question-level inspection reveals that nearly all substantial disagreements, defined as discrepancies of two or more rubric levels, are concentrated in only three items: P009 (question about the LiteLLM proxy version), P013 (mandatory CSV input columns), and P079 (question regarding context-window mechanisms outside the documented scope of the corpus). Examination of these cases reveals a consistent pattern: disagreements did not arise from divergent judgments regarding response quality itself, but from differences in the evaluators’ verification of ground truth within the corpus.

In P009, one evaluator located the relevant information in `docker-compose.yml`, while the other did not identify it during review. In P013, evaluators interpreted the CSV specification in `etl/README.md` differently. In P079, originally classified as a Category 2 question, one evaluator reasonably considered it a Category 3 absence-reasoning item due to the wording of the query.

This pattern also reveals a limitation of the evaluation protocol. Evaluators were required not only to assess answer quality but also to independently verify the supporting evidence within a heterogeneous technical corpus. In a small number of cases, disagreement arose from differences in source localization rather than response assessment. If the

three substantively disputed questions are excluded, κ for the remaining 97 questions rises into the moderate-to-substantial agreement range.

Two methodological implications follow from this observation. First, future replications should incorporate a prior calibration session using a pilot subset of questions, in which evaluators collaboratively verify the presence or absence of relevant information before proceeding to independent blind scoring. Second, interpretation of the main findings should weigh practical agreement indicators at least as heavily as absolute κ values, particularly for findings based on unanimous extreme scores, where the structural effect of evaluator count is inherently limited.

6.2. Comparative Performance of the Two Conditions

Table 5 presents arithmetic means, standard deviations, and paired-sample Student's *t*-tests for the three experimental categories and the global dataset. Scores correspond to the mean of both evaluators' ratings for each response.

Table 5. Comparative performance between full-context retrieval (Condition A) and agentic knowledge curation (Condition B), by category and globally. Two-tailed paired *t*-tests, $\alpha = 0.05$.

Category	n	Full-Context Retrieval (M $\pm$ SD)	Curation (M $\pm$ SD)	Δ	t; p	Result
Category 1—Direct retrieval	40	4.35 $\pm$ 0.96	4.69 $\pm$ 0.78	0.34	−2.86; 0.007	Curation > RAG ($p < 0.01$)
Category 2—Multi-document synthesis	40	4.23 $\pm$ 1.08	4.09 $\pm$ 0.91	−0.14	0.72; 0.476	No evidence of difference
Category 3—Absence reasoning	20	5.00 $\pm$ 0.00	5.00 $\pm$ 0.00	0	—	Identical perfect performance
Global	100	4.43 $\pm$ 0.96	4.51 $\pm$ 0.84	0.08	−0.87; 0.385	No evidence of global difference

The most important finding of Table 5 is therefore not the aggregate outcome, but the differentiated pattern across categories. The absence of global difference masks three distinct operational regimes: significant advantage in direct retrieval, parity in multi-document synthesis, and identical performance in absence detection.

To address the ordinal nature of the evaluation rubric and the observed ceiling effects, all paired comparisons were additionally replicated using Wilcoxon signed-rank tests. The Wilcoxon analyses yielded exactly the same pattern of conclusions as the paired *t*-tests: a significant advantage of agentic curation in Category 1, no significant differences in Category 2, and no global difference between paradigms. These results indicate that the findings are robust to the choice of parametric versus non-parametric statistical procedure (Table 6).

Table 6. Robustness analysis using Wilcoxon signed-rank tests. Category 3 was not tested because both conditions obtained identical scores for all items (all differences = 0).

Category	W	p-Value	Conclusion
Category 1—Direct retrieval	9	0.0019	Curation > RAG
Category 2—Multi-document synthesis	125	0.194	No difference
Global	407	0.572	No difference

The aggregated pattern combines three qualitatively distinct outcomes that require joint interpretation. In Category 1 (direct retrieval), agentic knowledge curation significantly outperforms the full-context retrieval condition, with a difference of 0.34 rubric points and robust statistical significance ($p = 0.007$). For descriptive purposes, the observed difference

corresponds to a paired-samples Cohen's *d* of approximately 0.39, computed using the standard deviation of the paired score differences. According to conventional benchmarks, this represents a small-to-moderate effect.

In Category 2 (multi-document synthesis), the direction of the difference reverses slightly in favor of the full-context retrieval condition (-0.14 points), but the difference does not reach statistical significance ($p = 0.476$). No evidence therefore supports a systematic advantage of either paradigm under this regime.

In Category 3 (absence reasoning), both conditions achieved identical perfect performance across all questions, producing zero variance in both paradigms.

At the global level, the 0.08-point advantage of agentic curation does not reach statistical significance ($p = 0.385$). This result should be interpreted as absence of detectable difference at the available statistical power rather than formal equivalence. This distinction is particularly relevant for H1: the study does not demonstrate equivalence in the strict TOST sense, but it does provide evidence that any global difference, if present under the evaluated conditions, lies below the detection threshold of the current design.

6.3. Distribution of Scores

The central tendency and dispersion statistics presented in Section 6.2 do not fully capture the differential behavior between paradigms. Examination of the complete score distributions reveals additional qualitative distinctions summarized in Table 7.

Table 7. Distribution of final scores (mean of two evaluators) by condition and rubric level.

Condition	Score 1	Score 2	Score 3	Score 4	Score 5	Total
Full-context retrieval (Condition A)	2	5	5	28	60	100
Agentic knowledge curation (Condition B)	0	7	4	24	65	100

Three observations merit detailed discussion. First, agentic knowledge curation received no score of 1 on any question, whereas full-context retrieval received two. These minimum scores correspond to two specific items (P028 and P050) in which both evaluators independently and unanimously assigned a score of 1 to the full-context retrieval response, while the corresponding responses in Condition B received unanimous scores of 5. These two cases are qualitatively the most informative in the entire experiment and are analyzed in detail in Section 6.4. The complete absence of score-1 responses in the curation condition, combined with the presence of unanimous catastrophic failures in full-context retrieval, constitutes strong asymmetric evidence regarding the distribution of severe retrieval failures between paradigms.

Second, agentic curation accumulated a greater number of perfect scores (65 vs. 60), consistent with the advantage observed in Category 1. However, this gain at the upper extreme of the rubric is accompanied by a slight increase in score-2 responses (7 vs. 5), concentrated in multi-document synthesis questions. This relative polarization (more perfect answers but also more identifiable deficiencies) is interpretable as the manifestation of the structural trade-off predicted by the framework in Section 3.5: the same mechanism that improves direct retrieval also introduces localized costs in textual fidelity for specific classes of questions.

Third, the score distribution of full-context retrieval is more uniform across the upper range of the rubric ($28 + 60 = 88$ responses scoring 4 or 5), whereas the distribution of agentic curation is more polarized ($24 + 65 = 89$). In terms of total high-quality responses (scores 4 and 5 combined), the paradigms are nearly indistinguishable (88 vs. 89). The

difference therefore lies not in the quantity of high-quality answers, but in how that quality is distributed across the highest rubric levels and in the residual presence of severe failures.

Taken together, these observations support a substantive interpretation: the two paradigms are not functionally equivalent despite similar aggregate means. They fail and succeed on different types of questions, and they do so through distinguishable mechanisms. Section 6.4 examines the concrete cases underlying this interpretation.

6.4. Qualitative Analysis of Singular Cases

The aggregated quantitative analysis presented in previous sections is complemented by detailed examination of questions whose scores reached rubric extremes with unanimous evaluator agreement, or whose divergence between paradigms exceeded two rubric points. These cases are qualitatively informative for two reasons: unanimous extreme ratings indicate structurally unambiguous successes or failures, and large divergences between paradigms allow for the identification of the mechanisms responsible for differential behavior. Table 8 summarizes the four most informative cases.

Table 8. Singular cases with unanimous extreme scores.

ID	Cat.	Full-Context Retrieval Condition	Curation	Analysis
P028	1	1.0 (unanimous)	5.0 (unanimous)	Question about the LiteLLM administration panel URL (http://localhost:4000/ui). Information explicitly present in manual-usuario.md, Section 4. Full-context retrieval returned [NO RESULT]; curation retrieved the URL correctly. Confirmed false negative in the full-context retrieval condition.
P050	2	1.0 (unanimous)	5.0 (unanimous)	Question regarding the L2-normalization condition underlying cosine/Euclidean equivalence, documented in manual-usuario.md §9.3. Full-context retrieval failed to retrieve it; curation stored it as an independent entry in context. Second confirmed false negative in the full-context retrieval condition.
P052	2	5.0 (unanimous)	2.5 (E1 = 1; E2 = 4)	Question regarding the exact WRRF formula implemented in search_engine.py. Full-context retrieval reproduced the formula literally: $\text{Score}(d) = 0.5/(60 + \text{rankBM25}(d)) + 0.5/(60 + \text{rankVector}(d))$ Curation paraphrased it with reduced precision, omitting the denominator. Evidence of localized fidelity loss.
P055	2	5.0 (unanimous)	2.0 (E1 = 1; E2 = 3)	Question regarding the exact seven-step initialization sequence in entrypoint.sh. Full-context retrieval reproduced the ordered sequence correctly. Curation omitted at least one step during synthesis. Second fidelity-loss case.

The four cases provide direct qualitative evidence for the two central theoretical mechanisms proposed in the conceptual framework.

6.4.1. P028 and P050: Evidence for Structural Visibility

P028 and P050 constitute the strongest qualitative evidence in the experiment due to both evaluator unanimity and mechanistic clarity. In both cases, the correct answer is unequivocally present in the corpus, yet located in structurally non-prominent regions of the largest document in the dataset (manual-usuario.md). In P028, the information appears as a table cell within a “quick access” section; in P050, it appears as a technical implementation note near the end of a subsection. Neither location possesses explicit structural salience.

The full-context retrieval condition, despite full access to the corpus context, returned [NO RESULT] in both cases. Agentic curation answered correctly because the ingestion process had redistributed the information into dedicated entries organized around retrieval-relevant concepts. Information that was structurally peripheral in the original corpus became directly navigable in the curated representation.

This pattern is consistent with the type of positional retrieval failures described in the lost-in-the-middle literature [3], although the present design does not permit causal attribution of the observed failures to document position alone: native attention over large contexts degrades systematically when relevant information occupies non-prominent positions. Agentic curation eliminates positional dependence by reorganizing the corpus into an explicit retrieval-oriented structure.

6.4.2. P052 and P055: Evidence for Deferred Local Fidelity

P052 and P055 illustrate the complementary mechanism. In both cases, correct answering depends on exact reproduction of ordered expressions: a mathematical formula in P052 and a procedural sequence in P055. The full-context retrieval condition, with direct access to source code, reproduced both responses literally and obtained unanimous perfect scores. Agentic curation, operating over ingestion-generated summaries, reproduced them with lower precision.

This pattern corresponds directly to H3. If semantic compression during ingestion is the mechanism responsible for the structural visibility observed in P028 and P050, it necessarily introduces localized costs in exact textual fidelity. The four cases jointly suggest that the two mechanisms are inseparable aspects of the same architectural trade-off.

An additional observation is noteworthy. In P052 and P055, evaluators differed not on whether a fidelity problem existed, but on its severity. Both agreed that the curated representation introduced imprecision; disagreement concerned the magnitude of the penalty. This pattern is consistent with the inter-rater agreement analysis in Section 6.1.

6.5. Formal Evaluation of Hypotheses

Table 9 synthesizes the evaluation of the four hypotheses formulated in Section 4 against the empirical evidence presented in the preceding sections.

Table 9. Formal evaluation of hypotheses against the empirical evidence.

H	Hypothesis	Empirical Evidence	Result
H1	No detectable global performance difference in context-fitting corpora	Global: $p = 0.385$ (no significant difference); Wilcoxon $p = 0.572$; Category 1: $p = 0.007$ favoring curation; Category 2: $p = 0.476$; Category 3: perfect parity	Supported as absence of detectable difference; formal equivalence not established
H2	Advantage of agentic curation in direct retrieval over long documents	Category 1: $\Delta = +0.34$; $t = -2.86$; $p = 0.007$. Cases P028 and P050 confirm mechanism qualitatively	Confirmed with statistical and qualitative support
H3	Disadvantage of agentic curation in tasks requiring exact textual fidelity	Category 2: $\Delta = -0.14$; $p = 0.476$. Cases P052 and P055 confirm mechanism qualitatively	Qualitatively confirmed; aggregate effect not significant
H4	Parity in absence detection with fewer false negatives in curation	Category 3: 5.00 vs. 5.00 parity. Confirmed false negatives: full-context retrieval = 2; curation = 0	Confirmed in both components

Taken together, the results support the four hypotheses with varying degrees of statistical and qualitative strength. The strongest support corresponds to H2 and H4. H2 combines statistical significance ($p = 0.007$), unanimous qualitative evidence (P028 and P050), and direct mechanistic correspondence with the theoretical literature on positional bias. H4 is confirmed in precisely the asymmetric form predicted: parity in genuine absence detection combined with fewer false negatives under agentic curation.

H1 receives a more cautious interpretation. While no detectable global difference emerged between the paradigms, suggesting that semantic reorganization preserves sufficient factual content, the design does not establish formal equivalence. Instead, it indicates that any global difference lies below the detection threshold of the present study. Similarly,

H3 is supported qualitatively rather than statistically. Cases P052 and P055 suggest that semantic compression introduces localized fidelity degradation precisely in tasks requiring exact symbolic or procedural reproduction, although this effect does not reach significance at the aggregate level.

More importantly, the integrated reading of these results supports the conceptual framework as a coherent whole. The three theoretical mechanisms proposed earlier find corresponding empirical manifestations in the observed data:

- Structural visibility is evidenced by the Category 1 advantage and the elimination of full-context retrieval's false negatives.
- Deferred local fidelity is visible in cases P052 and P055, and in the residual increase in score-2 responses under curation.
- Reversible compression is manifested in the global parity and the perfect reasoning regarding information absence.

The convergence between theoretical predictions and empirical observations constitutes the central finding of this study. The results suggest that retrieval architectures should not be compared solely through aggregate accuracy metrics, as the principal difference between paradigms lies in the structure of their failures, specifically, what kinds of information become invisible and under what representational conditions. From this perspective, agentic knowledge curation does not universally outperform full-context retrieval; rather, it shifts the balance between semantic accessibility and literal fidelity by relocating computational effort from query-time retrieval to ingestion-time organization.

7. Discussion

7.1. Naturalistic Observations Consistent with the Lost-in-the-Middle Phenomenon

The work [3] demonstrated that language models exhibit a characteristic U-shaped retrieval curve over long contexts: information placed at the beginning or end of the context is retrieved substantially more reliably than information located in intermediate positions. Their experiments relied necessarily on synthetic manipulation of passage position within controlled contexts, allowing positional effects to be isolated experimentally.

A central open question in the subsequent literature [24,25] has been whether the phenomenon persists under naturalistic conditions involving authentic technical documentation rather than adversarially constructed benchmarks. Real-world corpora differ fundamentally from synthetic evaluation settings: documents are heterogeneous, information distribution emerges organically rather than experimentally, and operational queries rarely resemble benchmark-style retrieval tasks.

The P028 and P050 cases observed in the present study provide observations directly relevant to this gap. Both questions have a single, verifiable answer based on the repository's documentation; both answers appear verbatim in the corpus, and both are found in unremarkable positions within the longest document in the corpus (manual-usuario.md, ~10,000 tokens). In P028, the information refers to a cell within a quick-reference table embedded in an intermediate section of the manual. In P050, it is a technical note located in a minor subsection. Neither location was chosen deliberately; both are the result of the natural process of drafting the manual.

Despite full corpus availability within context, the full-context retrieval condition declared absence of information in both cases and received unanimous minimum scores from both evaluators. Agentic curation succeeded because ingestion had transformed the same information into explicitly retrievable entries. The qualitative structure of these failures is compatible with explanations based on positional retrieval bias, although other retrieval-related factors may also have contributed. The finding suggests that retrieval failures qualitatively similar to those reported in lost-in-the-middle studies can emerge

in realistic technical documentation environments. However, the present study does not isolate positional effects experimentally and therefore cannot establish that document position was the causal factor underlying the observed failures.

This contribution should nevertheless be interpreted with appropriate caution. The present experiment does not estimate a population-level false-negative rate for long-context retrieval systems. The sample size is limited, and the question set was not explicitly stratified by positional prominence. The significance of P028 and P050 lies elsewhere: they show that catastrophic false negatives may occur under realistic conditions in a manner consistent with positional retrieval failures.

From a systems perspective, this observation has practical importance. It suggests that direct access to long context does not eliminate retrieval invisibility in technical corpora with heterogeneous information distribution. Instead, positional bias appears to remain a structural property of attention-based retrieval even when all relevant information is technically available to the model.

7.2. Structural Visibility as a Retrieval Mechanism

The significant advantage of agentic curation in Category 1 (+0.34 points; $p = 0.007$, see Table 5) is best interpreted through the mechanism introduced in Section 3.5 as structural visibility.

During ingestion, the agent does not merely summarize the corpus. It actively redistributes information according to semantic relevance within the domain, transforming peripheral details embedded inside larger documents into dedicated retrieval units with explicit organizational identity. This transformation changes the topology of retrieval itself. In the original corpus, information competes for attention within a continuous sequence of approximately 19,300 tokens. In the curated representation, information becomes accessible through deterministic navigation over semantically labeled entries. Retrieval ceases to depend primarily on attentional salience and becomes instead a problem of structural localization.

The importance of this shift is illustrated by the two severe failures observed in the full-context retrieval condition. A URL for accessing the admin panel, which in `user-manual.md` was a table cell within a quick-start section, is transformed, in agentic curation, into a standalone entry under `reference/` with its own header, explicit description and citation of the source file. A technical note on L2 normalization, which in the manual occupied a paragraph at the end of a chapter on embeddings, becomes a dedicated entry in `context/` with the explicit status of a documented design decision. The redistribution has a bidirectional effect: it increases the structural prominence of specific facts and, simultaneously, transforms the retrieval task into a deterministic navigation operation involving paths and names.

This mechanism, which we refer to as structural visibility, provides a plausible interpretative explanation for the patterns observed in Category 1 and the two specific cases discussed in Section 6.4.1. The advantage does not stem from any superior capacity of the agentic curation paradigm regarding the intrinsic content of the corpus, but rather from a prior transformation of the content's organization that eliminates the variable "position" from the retrieval problem. The agent in Condition B, when working on the wiki, does not face the problem of uniformly attending to 19,300 tokens; it faces the problem of locating the correct entry by index and reading its content in full.

This interpretation aligns conceptually with Anthropic's proposal of contextual retrieval [9], where contextual descriptions are added to fragments before embedding generation to preserve semantic locality. Agentic curation can be interpreted as a more radical

extension of the same principle: instead of enriching fragments while preserving vector retrieval, it reorganizes the corpus entirely around explicit semantic accessibility.

Importantly, the present study does not establish this mechanism causally in the strict experimental sense. The evidence is mechanistically consistent with the structural visibility hypothesis, but alternative explanations cannot be completely excluded. A factorial design isolating positional prominence from deterministic navigation, for example, a reranked long-context condition forcing relevant information toward context edges, would be required for strict causal decomposition. Nevertheless, the convergence between quantitative results, qualitative cases, and prior literature strongly suggests that structural visibility may be an important mechanism underlying the observed reduction in false negatives.

7.3. Semantic Navigability Versus Literal Fidelity

The same ingestion process responsible for structural visibility also explains the localized fidelity degradation observed in P052 and P055. The relationship between both effects is not incidental but structural. Agentic curation operates through semantic compression. During ingestion, the agent transforms the corpus into conceptually organized representations optimized for navigability and retrieval. This process preserves meaning effectively at the conceptual level but may degrade representations whose informational value depends on exact syntax, ordering, or symbolic precision.

The distinction is particularly visible in formulas and procedural sequences. A summarized description of Weighted Reciprocal Rank Fusion preserves the conceptual role of the algorithm but not necessarily its exact mathematical structure. Likewise, a narrative description of an initialization script preserves operational logic while potentially omitting ordering dependencies or low-level implementation details.

This trade-off resembles broader distinctions studied in both cognitive science and information theory. In cognitive terms, the paradigm behaves more like semantic memory than episodic memory [40]: it preserves conceptual organization better than verbatim reproduction. In informational terms, the process resembles lossy rather than lossless compression. The agentic curation paradigm operates as a semantic transformation of the corpus: it organizes meaning but does not fully preserve form. The full-context retrieval, which has access to the entire corpus, preserves form at the expense of structural prominence.

The experiment suggests that no single retrieval paradigm optimizes simultaneously for semantic accessibility and exact textual fidelity across all query regimes. Systems emphasizing navigability improve robustness for distributed factual retrieval but incur localized fidelity costs. Systems preserving raw corpus structure retain verbatim precision but remain vulnerable to positional invisibility. Consequently, the optimal architecture depends on the expected query distribution over the corpus. Corpora dominated by dispersed factual queries benefit from semantic restructuring, whereas corpora dominated by formulaic, procedural, or code-level queries benefit from preserving direct access to source material.

Most realistic technical environments likely contain both query types simultaneously. This observation suggests that hybrid architectures may outperform either pure paradigm individually: curated semantic layers could support navigability, while direct corpus access remains available for high-fidelity reconstruction tasks. The present study does not evaluate such hybrid systems, but the hypothesis follows naturally from the mechanisms identified experimentally.

7.4. Agentic Memory, Claude Code, and the Reorganization of Knowledge

The findings also connect naturally with the emerging literature on agentic memory systems. Frameworks such as MemGPT [5], A-MEM [6], and Zep [7] share a common principle: agents should actively construct and manage knowledge structures rather than passively consume retrieved fragments. The core convergence across these approaches is the idea that raw context availability is structurally inferior to organized memory. Whether implemented as memory graphs, evolving contextual nodes, temporal knowledge stores, or structured text files, the underlying thesis remains the same: retrieval quality depends not only on what information exists, but on how that information is reorganized before future access. Similar intuitions appear in recent multi-agent approaches such as GraphCogent [31], where explicit external structures are used to compensate for limitations of internal working memory during complex reasoning tasks. Although GraphCogent focuses on graph understanding rather than corpus curation, it reflects the same broader trend toward external organization of knowledge as a mechanism for improving retrieval and reasoning performance.

The paradigm studied here differs from most agentic-memory systems in three important respects. First, it relies on structured plain text rather than specialized graph or vector infrastructures. This distinction is operationally significant because it preserves direct human readability, auditability, and version control compatibility. Second, it operates primarily over authoritative external corpora rather than experiential conversational memory. The agent reorganizes existing knowledge instead of accumulating new episodic traces during operation. Third, and perhaps most importantly, the paradigm already possesses an industrial-scale precedent through Claude Code and its CLAUDE.md shared-memory mechanism. Developers using Claude Code in their projects collaboratively edit a CLAUDE.md file that sets out design decisions, project conventions and operational rules; the agent reads this file at the start of each session and uses it as a shared memory. The file is human-readable, can be versioned with Git, can be edited manually and can be audited just like any other code file. Developers can ask the agent to extract relevant information from a session and incorporate it into the file, or they can edit it directly to correct the conventions.

This precedent matters because it demonstrates that the paradigm is not merely theoretically plausible but operationally viable at scale. The convergence between the empirical findings of the present study and the real-world usage patterns of Claude Code strengthens the plausibility that the observed mechanisms are not corpus-specific artifacts but manifestations of a broader organizational principle for LLM-assisted knowledge systems.

7.5. Design Implications and Threats to Validity

The findings carry several implications for the design of applied knowledge-management systems supported by Large Language Models (LLMs) in applied environments. First, retrieval architectures should be selected according to corpus structure and query distribution rather than aggregate benchmark accuracy alone. The experiment suggests that paradigms with equivalent mean performance may exhibit radically different operational failure profiles.

Second, false negatives deserve greater attention as a retrieval failure mode. The two documented false negatives observed in the full-context retrieval condition (P028 and P050) were qualitatively more dangerous than the fidelity degradations observed under agentic curation because they rendered existing information invisible to the user. In practical environments, invisible failures are often operationally costlier than visibly imperfect answers, because a non-expert user has no way of knowing that the information existed.

Third, the operational asymmetry between paradigms extends beyond answer quality. Agentic curation:

- (a) may reduce vector infrastructure requirements (embedding model, managed vector store, reindexing pipeline) with the direct and indirect cost implications discussed in Section 2.1.
- (b) is architecturally compatible with localized maintenance through localized editing. In conventional vector-based retrieval systems, any modification to the corpus requires re-indexing at least the affected part, with a cost proportional to the size of the corpus or the modified fragment; in agentic curation, the modification is carried out by locally editing the affected entry, with a cost proportional to the size of the change.
- (c) provides human-readable knowledge representations that may facilitate auditing and traceability. Vector-based retrieval systems represent information through embeddings that cannot be read by humans without specialized tools; the agent-based curation paradigm stores knowledge in text files that are immediately readable, editable and traceable back to their original source.

These considerations derive from the architectural properties of the paradigm rather than from direct measurements performed in the present study. No benchmarks of latency, token consumption, ingestion cost, maintenance effort, or operational reliability were conducted.

These properties are particularly relevant in regulated or organizational contexts requiring transparency and traceability, such as those set out in Regulation (EU) 2024/1689 [41] on artificial intelligence for systems classified as high-risk.

The study nevertheless presents important limitations that constrain generalization:

- The corpus belongs to a single technical domain and remains sufficiently small to fit entirely within the model context window. Replication across larger corpora and additional domains is necessary before broader claims can be established.
- The experiment used a single foundational model. Although the mechanisms identified are architecture-independent, replication across model families remains necessary.
- The study does not evaluate the stability of the ingestion process itself. During corpus construction, different ingestion procedures produced knowledge bases with different coverage levels, suggesting sensitivity to document presentation and prompting conditions. Since agentic knowledge curation shifts substantial responsibility from retrieval to ingestion, the robustness and reproducibility of ingestion outcomes constitute an important open question that remains outside the scope of the present experiment.
- The study was conducted using two external and independent assessors. The characterization of agreement using Cohen's κ is structurally appropriate for this design, but the robustness of the estimate would benefit from triangulation with a third rater, which would allow for the use of multi-rater statistics (Fleiss's κ or Krippendorff's alpha) and the resolution of the three questions with substantive disagreement (P009, P013, P079).
- The operationalization of the full-context retrieval condition deserves careful interpretation. The experiment isolates retrieval organization effects by providing the full corpus directly within context rather than implementing a production retriever pipeline. Consequently, the findings should be interpreted primarily as evidence about long-context retrieval behavior rather than as direct benchmarking of optimized industrial RAG systems.
- The quality of the ingestion depends on the specific capabilities of the agent used. The extent to which the results generalize to models of comparable capability remains unknown and requires empirical verification.

A related limitation concerns the maturity of the conceptual framework itself. The constructs introduced in Section 3 (retrieval failure topology, structural visibility, reversible compression, and deferred local fidelity) are currently operationalized qualitatively rather than through dedicated quantitative metrics. The present study suggests their usefulness as explanatory dimensions for interpreting observed retrieval behavior, but does not provide formal measures linking corpus structure, ingestion transformations, and retrieval outcomes. Consequently, the framework should be understood as a first analytical step toward a more formal framework for understanding knowledge organization in long-context systems rather than as a completed theoretical model.

More broadly, the experimental scope is deliberately narrow: a single corpus (~19,300 tokens), a single application domain, a single foundation model, and a single curated knowledge base. Furthermore, the corpus fits entirely within the model's context window. Consequently, the findings should be interpreted as evidence about the comparison between agentic knowledge curation and a full-context retrieval implementation under a specific but ecologically realistic set of conditions, rather than as a basis for general claims regarding the replacement of RAG systems, the superiority of any architectural paradigm, or their performance across domains, corpus scales, and operational environments.

The study does not directly measure positional prominence or experimentally manipulate document position. Consequently, references to lost-in-the-middle should be interpreted as consistency with the phenomenon rather than direct causal evidence of it.

The ingestion process was evaluated through manual coverage verification, but no formal source-to-entry traceability matrix was constructed. Consequently, the study cannot quantify which portions of the corpus were copied verbatim, summarized, transformed, or omitted during curation. Future work should incorporate explicit traceability analyses linking source fragments to curated entries in order to characterize information redistribution during ingestion more rigorously.

Inter-rater reliability was moderate ($\kappa = 0.38$). Although practical agreement indicators were substantially higher (72.0% exact agreement; 88.5% within one rubric level), future studies would benefit from evaluator calibration procedures and additional raters.

Taken together, these limitations suggest that the study should be interpreted not as definitive evidence of paradigm superiority, but as a well-delimited empirical contribution supporting a broader research program on knowledge organization for language-model systems.

8. Conclusions and Future Research

8.1. Main Contributions

The contributions of the study operate at four distinct levels: conceptual, empirical, theoretical, and applied.

At the conceptual level, the article conceptualizes agentic knowledge curation as a distinct architectural paradigm within the broader landscape of LLM-assisted knowledge systems. Unlike classical retrieval-augmented generation, which relies on retrieval-time selection of corpus fragments, and unlike full-context retrieval, which relies on direct attention over the original corpus, agentic knowledge curation performs an explicit reorganization of knowledge during ingestion and exposes it through a structured, navigable representation. The study therefore contributes to expanding the conceptual design space of retrieval architectures beyond the commonly discussed dichotomy between RAG and long-context approaches. To support this distinction, the paper proposes a comparative framework spanning five paradigms and five analytical dimensions, from which three operational mechanisms are derived: reversible compression, structural visibility, and deferred local fidelity.

At the empirical level, the study provides externally evaluated quantitative evidence broadly consistent with the four hypotheses derived from the conceptual framework. The results suggest no detectable global performance differences between full-context retrieval and agentic knowledge curation, while revealing a statistically significant advantage of agentic curation in direct retrieval tasks and parity in absence detection. The same pattern of conclusions was obtained using both paired *t*-tests and Wilcoxon signed-rank tests, indicating robustness to the choice of statistical procedure. Qualitative analysis further identified localized fidelity losses in specific synthesis tasks requiring exact textual reconstruction. Importantly, the strongest findings rest not only on aggregate statistics but also on qualitative evidence with unanimous agreement among reviewers in the singular cases P028, P050, P052, and P055. The study additionally highlights ingestion robustness as a potentially critical dependency of curation-based systems, highlighting an area that remains largely unexplored in the current literature.

At the theoretical level, the study provides naturalistic observations consistent with the existence of positional retrieval failures in long-context systems. The false negatives observed in P028 and P050 emerged in authentic technical documentation without adversarial manipulation of passage placement, extending prior evidence on the lost-in-the-middle phenomenon beyond synthetic benchmark settings. More broadly, the results suggest that retrieval architectures should be analyzed not only through aggregate accuracy metrics but also through the topology of their failures, namely which kinds of information become inaccessible under different representational regimes.

At the methodological level, the study contributes a reproducible evaluation protocol combining verified ground truth, blind human assessment, and a naturalistic technical corpus. Such evaluation designs remain relatively uncommon in retrieval-architecture research, which is dominated by automated metrics and benchmark datasets. The protocol provides a complementary methodology for studying retrieval behavior under realistic documentation conditions.

At the applied level, the work derives several implications for the engineering of retrieval systems. The findings indicate that retrieval architectures should be selected according to corpus structure and dominant query regimes rather than aggregate benchmark accuracy alone, that false negatives constitute a particularly dangerous class of invisible retrieval failure, that operational considerations such as auditability, maintenance complexity, and infrastructure cost must be incorporated into architectural evaluation, and that ingestion quality becomes a critical dependency in curation-based systems. More broadly, the study suggests that human-readable structured memory may offer advantages in regulated or auditable environments where transparency and traceability are operational requirements.

8.2. Interpretative Synthesis

The central finding of the study is that the observed results are broadly consistent with a unified account of retrieval failure topology. The four hypotheses were designed as observable consequences of this broader prediction rather than as independent theoretical claims. Full-context retrieval and agentic knowledge curation are not functionally equivalent even when their aggregate means appear statistically indistinguishable. They succeed and fail on different classes of questions, and they do so through structurally distinct mechanisms. Agentic curation appears to improve robustness where semantic organization and structural accessibility dominate retrieval effectiveness; full-context retrieval preserves superior fidelity where exact symbolic or procedural reconstruction is required. The findings establish the existence of this trade-off under one realistic configuration; they do not establish its prevalence across the broader space of retrieval systems.

This distinction has implications that extend beyond the binary comparison evaluated experimentally. The design space of LLM-assisted knowledge architectures is not organized linearly around a single aggregate quality metric. Instead, it is structured around competing architectural trade-offs whose dimensions must be made explicit: navigability versus literal fidelity, structural visibility versus verbatim preservation, ingestion-time organization versus inference-time retrieval. From this perspective, the contribution of the present study is not to declare a universally superior paradigm, but to empirically map one of the central trade-offs governing the emerging architecture space of long-context knowledge systems. The experiment suggests that retrieval performance depends not only on model capability or corpus availability, but also on how knowledge is reorganized before future access. Because the comparison isolates full-context access rather than a production retrieval pipeline, the findings should be interpreted as evidence about alternative forms of knowledge organization and access, not as a direct benchmark against optimized industrial RAG systems. Operational implications discussed in Section 7.5 should be interpreted as design-level consequences inferred from the architecture rather than as experimentally validated performance outcomes.

The broader implication is methodological as much as technical. Aggregate accuracy metrics alone are insufficient to characterize knowledge-access architectures because they obscure differences in failure topology. Two systems with comparable average performance may produce radically different operational behaviors depending on what kinds of information become invisible, under what structural conditions, and with what consequences for users.

8.3. Future Research

The research agenda emerging from the present work can be organized into four principal directions:

- **Multi-model replication:** the experiment was conducted using a single foundational model. Replication across models with different architectures, sizes, retrieval mechanisms and context-management strategies is necessary to determine whether the observed patterns reflect paradigm-level properties or partially depend on characteristics specific to Claude 3.5 Sonnet. This constitutes the most immediate and informative extension for assessing generalizability.
- **Multi-evaluator validation:** future replications should incorporate a third external evaluator, enabling the use of multi-rater agreement statistics such as Fleiss' κ or Krippendorff's α and reducing uncertainty in the small subset of disputed cases identified in Section 6.1. Although the substantive conclusions of the study rely primarily on unanimously evaluated extreme cases, additional evaluator triangulation would strengthen the robustness of aggregate metrics.
- **Hybrid retrieval architectures:** the trade-off identified between semantic navigability and literal fidelity suggests that hybrid architectures may outperform either pure paradigm in heterogeneous technical environments. In such systems, curated semantic memory could support conceptual retrieval while direct corpus access remains available for questions requiring verbatim precision. The conceptual framework proposed in this article naturally suggests the plausibility of such architectures, although their empirical evaluation remains beyond the scope of the present study.
- **Comparison with production RAG pipelines:** the present study isolates the contrast between full-context retrieval and agentic knowledge curation. Future work should evaluate the same hypotheses against fully implemented retrieval pipelines incorporating chunking, embeddings, hybrid retrieval, reranking, and contextual retrieval

techniques. Such comparisons would clarify which observed effects are specific to long-context access and which generalize to contemporary RAG architectures.

- Systematic characterization of positional bias in naturalistic settings: the two documented false negatives provide strong qualitative observations consistent with lost-in-the-middle effects that occur in realistic technical documentation, but they do not permit estimation of the population-level frequency or severity of the phenomenon. A dedicated large-scale study stratified by positional prominence within long technical documents would allow for quantitative characterization of positional retrieval bias under naturalistic conditions and could contribute both to theoretical understanding of long-context attention and to the design of mitigation strategies for applied systems.
- Formal operationalization of knowledge-organization constructs: future work should develop quantitative measures for the analytical constructs introduced in this article. Examples include positional-prominence metrics for information within long documents, measures of information redistribution during ingestion, retrieval-path complexity metrics for structured knowledge bases, fidelity-preservation indices comparing source content and curated representations, and corpus-level characterizations of retrieval failure distributions. Such measures would enable formal testing of the mechanisms proposed here and support the development of a more rigorous theory of retrieval behavior across knowledge-organization paradigms.
Operational benchmarking: future work should compare paradigms using direct measurements of ingestion cost, token consumption, update effort, latency, infrastructure requirements, and maintenance complexity in order to quantify the practical implications discussed qualitatively in the present study.
- Ingestion robustness and reproducibility: future work should systematically evaluate how knowledge-base quality varies as a function of prompt design, document format, corpus size, model family, and ingestion strategy. Since curation-based systems relocate much of the retrieval burden to ingestion, understanding ingestion stability is essential for assessing their practical viability.
- Traceability and ingestion analysis: future work should develop source-to-entry traceability frameworks capable of quantifying how information is redistributed during curation. Such analyses could distinguish between verbatim preservation, semantic compression, transformation, and omission, enabling more precise characterization of the relationship between ingestion processes and downstream retrieval performance.

8.4. Final Remarks

Over the last four years, the consolidation of RAG as the dominant retrieval architecture has directed much of the research literature toward incremental refinement of vector retrieval pipelines and chunking strategies. The emergence of long-context models and increasingly capable agents able to construct and maintain explicit knowledge structures expands the accessible architectural space and reopens a more fundamental question: under what conditions vector infrastructure remains beneficial relative to alternative forms of knowledge organization?

The present study contributes a bounded but empirically grounded answer to that question. Under the evaluated conditions, neither full-context retrieval nor agentic knowledge curation emerges as universally superior. However, under the conditions evaluated here, it appears to reduce a specific class of positional retrieval failures while maintaining globally comparable performance. The paradigm therefore appears not as a replacement for retrieval augmentation, but as an alternative allocation of where computational and organizational effort occurs within the retrieval process itself.

For system designers operating in 2026, this distinction is increasingly consequential. As context windows continue to expand and agentic workflows become operationally viable, architectural decisions should not be guided solely by assumptions inherited from early-generation RAG systems. The contribution of this work is to replace part of that intuition with preliminary empirically grounded evidence regarding the structural trade-offs that govern retrieval behavior in long-context LLM systems.

The present work should therefore be interpreted as an exploratory empirical study and a conceptual investigation of an emerging knowledge-organization paradigm rather than as a definitive evaluation of retrieval architectures. Broader conclusions will require replication across models, corpora, retrieval settings, and operational environments.

Author Contributions: Conceptualization, C.A.M. and R.M.A.; methodology, C.A.M., R.M.A. and S.A.; software, C.A.M., M.A.B. and J.M.T.; validation, C.A.M., M.A.B. and J.M.T.; writing—original draft preparation, S.A. and R.M.A.; supervision, C.A.M.; funding acquisition, S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by CajaCanarias Foundation and La Caixa Foundation, grant number 2023DIG11.

Institutional Review Board Statement: The study involved two adult expert evaluators who anonymously assessed AI-generated responses using a predefined evaluation rubric. No personal, sensitive, or health-related data were collected, and participation involved minimal risk. According to Spanish legislation governing research involving human subjects, including Spanish Law 14/2007 on Biomedical Research: BOE-A-2007-12945 (ref. [42]), this type of anonymous expert evaluation study does not require formal approval by a Research Ethics Committee.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Participants were informed about the purpose of the research, the voluntary nature of their participation, the anonymous treatment of the collected data, and the intended use of the evaluations for scientific publication.

Data Availability Statement: All data and materials necessary to reproduce the findings of this study are publicly available in the accompanying replication repository. The repository is provided to enable independent verification, reanalysis, and extension of the experimental protocol. Repository URL: “<https://zenodo.org/records/20752159> (accessed on 22 June 2026)”.

Acknowledgments: During the preparation of the manuscript, the authors used Claude Sonnet 5 (Anthropic) as an assistant for linguistic editing and argument structuring. The authors reviewed and edited the content in its entirety and assume final responsibility for the published material. The experimental execution, methodological design, interpretation of results, and drafting of conclusions are the sole responsibility of the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NIPS’20: Proceedings of the 34th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2020; pp. 9459–9474.
2. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv* **2024**, arXiv:2312.10997.
3. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [CrossRef]
4. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing reasoning and acting in language models. *arXiv* **2022**, arXiv:2210.03629.
5. Packer, C.; Wooders, S.; Lin, K.; Fang, V.; Patil, S.G.; Stoica, I.; Gonzalez, J.E. MemGPT: Towards LLMs as operating systems. *arXiv* **2023**, arXiv:2310.08560.
6. Xu, W.; Liang, Z.; Mei, K.; Gao, H.; Tan, J.; Zhang, Y. A-MEM: Agentic memory for LLM agents. *arXiv* **2025**, arXiv:2502.12110.

7. Rasmussen, P.; Paliychuk, P.; Beauvais, T.; Ryan, J.; Chalef, D. Zep: A temporal knowledge graph architecture for agent memory. *arXiv* **2025**, arXiv:2501.13956.
8. Singh, A.; Ehtesham, A.; Kumar, S.; Khoei, T.T. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv* **2025**, arXiv:2501.09136.
9. Anthropic. Introducing Contextual Retrieval. Anthropic Research. 2024. Available online: <https://www.anthropic.com/news/contextual-retrieval> (accessed on 22 June 2026).
10. Sarthi, P.; Abdullah, S.; Tuli, A.; Khanna, S.; Goldie, A.; Manning, C.D. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In Proceedings of the International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.
11. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Larson, J. From local to global: A graph RAG approach to query-focused summarization. *arXiv* **2024**, arXiv:2404.16130.
12. Karpathy, A. LLM Wiki: A Pattern for Building Personal Knowledge Bases Using LLMs. GitHub Gist. 2026. Available online: <https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f> (accessed on 22 June 2026).
13. Li, X.; Cao, Y.; Ma, Y.; Sun, A. Long context vs. RAG for LLMs: An evaluation and revisits. *arXiv* **2024**, arXiv:2501.01880.
14. Li, Z.; Chen, X.; Yu, H.; Lin, H.; Lu, Y.; Tang, Q.; Huang, F.; Han, X.; Sun, L.; Li, Y. LaRA: Benchmarking retrieval-augmented generation and long-context LLMs—No silver bullet for production. *arXiv* **2025**, arXiv:2502.09977.
15. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), Online, 16–20 November 2020; pp. 6769–6781.
16. Izacard, G.; Grave, E. 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 2021), Online, 19–23 April 2021; pp. 874–880.
17. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), Hong Kong, China, 3–7 November 2019; pp. 3982–3992.
18. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In Proceedings of the International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.
19. Yan, S.-Q.; Gu, J.-C.; Zhu, Y.; Ling, Z.-H. Corrective retrieval augmented generation. *arXiv* **2024**, arXiv:2401.15884.
20. Trivedi, H.; Balasubramanian, N.; Khot, T.; Sabharwal, A. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Toronto, ON, Canada, 9–14 July 2023; pp. 10014–10037.
21. Barnett, S.; Kurniawan, S.; Thudumu, S.; Brannelly, Z.; Abdelrazek, M. Seven failure points when engineering a retrieval augmented generation system. *arXiv* **2024**, arXiv:2401.05856.
22. Lyu, Y.; Li, Z.; Niu, S.; Xiong, F.; Tang, B.; Wang, W.; Wu, H.; Liu, H.; Xu, T.; Chen, E. CRUD-RAG: A comprehensive Chinese benchmark for retrieval-augmented generation of large language models. *arXiv* **2024**, arXiv:2401.17043.
23. Kuratov, Y.; Bulatov, A.; Anokhin, P.; Rodkin, I.; Sorokin, D.; Sorokin, A.; Burtsev, M. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*; Neural Information Processing Systems Foundation, Inc.: Red Hook, NY, USA, 2024.
24. An, C.; Chen, S.; Xiao, Z.; Zhang, S.; Liu, Z.; Wu, Y.; Wei, F.; Wei, Z. Why does the effective context length of LLMs fall short? *arXiv* **2024**, arXiv:2410.18745.
25. Leng, Q.; Portes, J.; Havens, S.; Zaharia, M.; Carbin, M. Long context RAG performance of LLMs. *arXiv* **2024**, arXiv:2411.03538.
26. Robertson, S.; Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389.
27. Cormack, G.V.; Clarke, C.L.A.; Büttcher, S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009), Boston, MA, USA, 19–23 July 2009; pp. 758–759.
28. Jiang, Z.; Xu, F.F.; Gao, L.; Sun, Z.; Liu, Q.; Dwivedi-Yu, J.; Yang, Y.; Callan, J.; Neubig, G. Active retrieval augmented generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), Singapore, 6–10 December 2023; pp. 7969–7992.
29. Li, X.; Dong, G.; Jin, J.; Zhang, Y.; Zhou, Y.; Zhu, Y.; Zhang, P.; Dou, Z. Search-o1: Agentic search-enhanced large reasoning models. *arXiv* **2025**, arXiv:2501.05366.
30. Chhikara, P.; Khant, D.; Aryan, S.; Singh, T.; Yadav, D. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv* **2024**, arXiv:2504.19413.
31. Wang, R.; Liang, S.; Chen, Q.; Huang, Y.; Li, M.; Ma, Y.; Zhang, D.; Qin, K.; Leung, M.-F. GraphCogent: Mitigating LLMs' Working Memory Constraints via Multi-Agent Collaboration in Complex Graph Understanding. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*; Association for Computing Machinery: New York, NY, USA, 2026; pp. 3811–3822. [CrossRef]

32. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. RAGAS: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024), System Demonstrations, St. Julians, Malta, 17–22 March 2024; pp. 150–158.
33. Saad-Falcon, J.; Khattab, O.; Potts, C.; Zaharia, M. ARES: An automated evaluation framework for retrieval-augmented generation systems. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024), Mexico City, Mexico, 16–21 June 2024; pp. 338–354.
34. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2023.
35. Clark, E.; August, T.; Serrano, S.; Haduong, N.; Gururangan, S.; Smith, N.A. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021), Bangkok, Thailand, 1–6 August 2021; pp. 7282–7296.
36. Chiang, C.-H.; Lee, H.-Y. Can large language models be an alternative to human evaluations? In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Toronto, ON, Canada, 9–14 July 2023; pp. 15607–15631.
37. Mombaerts, L.; Ding, T.; Banerjee, A.; Felice, F.; Taws, J.; Borogovac, T. Meta knowledge for retrieval augmented large language models. *arXiv* **2024**, arXiv:2408.09017.
38. Anthropic. Claude Code by Anthropic. 2024. Available online: <https://www.anthropic.com/product/claude-code> (accessed on 22 June 2026).
39. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [[CrossRef](#)] [[PubMed](#)]
40. Tulving, E. Episodic and semantic memory. In *Organization of Memory*; Tulving, E., Donaldson, W., Eds.; Academic Press: Cambridge, MA, USA, 1972; pp. 381–403.
41. The European Parliament; The Council of The European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, establishing harmonized rules on artificial intelligence. In *Official Journal of the European Union*; L Series; Publications Office of the European Union: Luxembourg, 2024.
42. Spanish Law 14/2007 on Biomedical Research: BOE-A-2007-12945. Available online: <https://www.boe.es/eli/es/l/2007/07/03/14> (accessed on 22 June 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.