SURVEY PAPER

PREPARED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE BIG DATA ANALYTICS COURSE

Semantic Search Engines over Big Data Using LLMs

Searching Large Text Corpora with Contextual Understanding

SUBMITTED BY

Khouloud BEN YOUNES

Montaha GHABRI

EVALUATED BY

Pr. Manel ABDELKADER

ACADEMIC YEAR

2025-2026

Abstract

This survey explores the domain of semantic search over big data using large language models (LLMs), focusing on dense retrieval techniques and retrieval-augmented generation (RAG) systems. Semantic search has evolved from traditional keyword-based methods to advanced embedding-based approaches that capture meaning and context, enabling more accurate information retrieval from massive corpora. The scope includes foundational concepts in dense retrieval, scalability in vector search, hybrid methods, and RAG architectures, drawing from key papers published between 2018 and 2025. Main findings reveal that dense retrievers like DPR and ColBERT improve semantic understanding but face challenges in scalability and vulnerability to attacks, while RAG systems like SELF-RAG enhance LLM reliability by integrating external knowledge. Key trends include unsupervised adaptation of LLMs for retrieval (e.g., Llama2Vec), hardware acceleration for RAG (e.g., Chameleon), and query enhancement techniques (e.g., HyDE, Query2doc). However, gaps persist in robustness against poisoning attacks, efficient compression for edge deployment, and ethical concerns like bias in retrieved content. This survey identifies opportunities for multimodal RAG and agentic systems, providing insights for future advancements in handling big data with LLMs.

Contents

List of Tables

INTRODUCTION

1.1 Background and Context

The exponential growth of digital data presents unprecedented challenges for information retrieval. Modern systems must process billions of documents, from Wikipedia's millions of articles to enterprise knowledge bases, while understanding semantic relationships beyond exact keyword matches. Traditional term-based methods like BM25, while efficient, suffer from vocabulary mismatch: relevant documents may use different terminology than queries, leading to missed information.

The emergence of Large Language Models (LLMs) and transformer architectures has revolutionized this landscape. Pre-trained models like BERT enable contextual understanding, capturing synonyms, paraphrases, and semantic relationships that elude classical methods. Dense retrieval systems can now map queries and documents to continuous vector spaces where semantic similarity corresponds to geometric proximity, fundamentally changing how we approach large-scale search.

1.2 Importance and Relevance

Neural semantic search addresses three critical limitations of classical retrieval: (1) **Vocabulary Mismatch**, understanding "physician" and "doctor" as equivalent; (2) **Semantic Understanding**, recognizing that "bad guy" matches "villain" in context; (3) **Compositionality**, interpreting multi-word queries as coherent semantic units rather than isolated terms.

However, transitioning to neural retrieval introduces new challenges. Dense representations require sophisticated indexing (approximate nearest neighbor search), raising questions about scalability to billion-document corpora. Storage requirements explode from compressed inverted indices (gigabytes) to dense vector databases (terabytes). Security concerns emerge as adversarial attacks can manipulate retrieval outputs. These trade-offs determine whether neural systems can safely replace classical methods in production environments.

1.3 Scope of the Survey

This survey focuses on semantic search over textual big data using LLMs, covering work from 2020-2025. We analyze 16 papers across five categories:

Included: Dense retrieval architectures (dual-encoder, late interaction), training methodologies (negative sampling, distillation), scalability systems (incremental indexing, vector quantization), RAG integration, sparse neural methods, hybrid approaches, and security analysis.

Excluded: Pure re-ranking (not first-stage retrieval), cross-encoders (unsuitable for large-scale deployment), multimodal retrieval (images, video), conversational search, and non-neural classical methods except as baselines.

1.4 Research Questions

1. **RQ1:** What are the fundamental neural approaches to semantic retrieval, and how do sparse, dense, and hybrid paradigms compare in effectiveness, efficiency, and scalability?
2. **RQ2:** What training methodologies (negative sampling, distillation, pre-training) most impact retrieval quality, and how do they enable sample-efficient learning?
3. **RQ3:** How do neural retrievers handle billion-scale deployment challenges including indexing latency, storage costs, incremental updates, and concurrent read-write workloads?
4. **RQ4:** What security vulnerabilities exist in dense retrieval systems, and what defenses protect against adversarial corpus poisoning?
5. **RQ5:** How do retrieval systems integrate with LLM generation in RAG architectures, and what hardware accelerations enable practical deployment?

1.5 Structure of the Paper

Chapter 2 describes our literature selection methodology, including databases searched, keywords, and selection criteria. Chapter 3 establishes background concepts including problem formalization, neural architectures, indexing strategies, and evaluation metrics. Chapter 4 reviews the 16 selected papers organized by contribution type. Chapter 5 provides critical analysis comparing paradigms and identifying design principles. Chapter 6 discusses emerging trends including RAG integration, multimodal retrieval, and efficiency advances. Chapter 7 concludes with key findings and recommendations.

LITERATURE SELECTION

To ensure a comprehensive and high-quality review of semantic search over big data with large language models (LLMs), we followed a systematic literature selection process. The goal was to identify influential, relevant, and recent works that directly contribute to the core topics: dense retrieval foundations, vector search scalability, RAG architectures, and hybrid/advanced methods.

2.1 Databases Searched

We conducted searches across the following major academic databases and repositories:

- IEEE Xplore
- ACM Digital Library
- SpringerLink
- Elsevier (ScienceDirect)
- arXiv (preprints, focusing on CS.CL, CS.IR, and cs.LG categories)

These databases cover the primary venues for information retrieval, natural language processing, machine learning, and systems research.

2.2 Keywords Used

The search strings combined the following keywords and synonyms (using Boolean operators):

- Core terms: “dense retrieval”, “semantic search”, “dense passage retrieval”, “vector search”, “embedding retrieval”, “retrieval-augmented generation”, “RAG”, “late interaction retrieval”
- Additional filters: “contrastive learning”, “hard negatives”, “vector quantization”, “ANN index”, “billion-scale”, “incremental update”, “query expansion”, “LLM retrieval”, “self-reflection RAG”, “unsupervised retrieval”

We also used venue-specific filters (e.g., SIGIR, ACL, EMNLP, NeurIPS, ICML) and citation thresholds in Google Scholar for cross-validation.

2.3 Time Window

Publications were restricted to the period 2018–2025. This range captures the rise of transformer-based dense retrieval (post-BERT 2018) through the rapid growth of LLM-integrated RAG systems (2023–2025).

2.4 Inclusion and Exclusion Criteria

Inclusion criteria:

- Peer-reviewed conference or journal paper (or highly cited arXiv preprint with strong venue follow-up)

- Directly addresses dense retrieval, vector database scalability, RAG architectures, or hybrid retrieval methods
- Contains empirical evaluation (experiments on standard datasets like MS MARCO, NQ, BEIR, etc.)
- Citation count:
  - 100+ citations for 2020–2022 papers
  - 20+ citations for 2023 papers
  - 10+ citations for 2024–2025 papers
- Published in top-tier venues (SIGIR, ACL, EMNLP, NeurIPS, ICML, VLDB, WWW) or highly influential arXiv papers widely cited in follow-up work

Exclusion criteria:

- Workshop, demo, poster, or short papers
- Purely theoretical work without experiments
- Non-English publications
- Redundant with earlier seminal works
- Focused only on non-text modalities or unrelated tasks

2.5 Selection Process and Results

Initial keyword searches returned over 800 results. After title/abstract screening, 180 papers were shortlisted. Full-text review and citation verification reduced this to 35 highly relevant papers. From these, we selected 15 core papers that best represent each category, meet citation thresholds, appear in top venues or are highly influential, and directly advance the survey topics.

The final set includes seminal works such as DPR [1] and ColBERTv2 [2], and recent breakthroughs including SIMLM [3], Llama2Vec [4], SELF-RAG [5], Chameleon [6], and HAKES [7]. It also incorporates influential surveys by Guo et al. [8] and Zhao et al. [9], as well as critical vulnerability studies regarding poisoning attacks [10]. These papers collectively cover the evolution from 2020 foundations to 2025 scalability and RAG optimizations.

BACKGROUND

This chapter provides the conceptual foundation for semantic search over big data with large language models.

3.1 Definitions and Terminologies

3.1.1 Core Retrieval Paradigms

- **Embeddings:** Dense, fixed-length vector representations of text produced by neural encoders (e.g., BERT). Typically ranging from 768 to 1536 dimensions, they capture semantic relationships through geometric proximity, where similarity is measured by dot product or cosine similarity.
- **Dense Retrieval:** A retrieval method utilizing learned embeddings. It typically employs dual-encoder architectures to map queries and passages to vectors independently, facilitating similarity scoring via approximate nearest neighbor search. This stands in contrast to sparse retrieval (e.g., BM25), which relies on lexical term overlap.
- **Retrieval-Augmented Generation (RAG):** A framework that integrates retrieval with generative LLMs. By fetching relevant documents from external sources to serve as context for the LLM, RAG grounds generation in factual evidence and reduces model hallucinations.

3.1.2 Optimization and Search Mechanics

- **Contrastive Learning:** A training objective that pulls positive query-passage pairs closer in vector space while pushing negative pairs apart.
- **InfoNCE Loss:** The standard contrastive loss function used to maximize the probability of retrieving correct positives.
- **Hard Negatives:** Passages that are semantically similar to the query but factually incorrect; these are used during training to improve the model's discriminative precision.
- **Late Interaction:** A mechanism that matches individual query tokens to passage tokens (e.g., ColBERT), allowing for fine-grained alignment.
- **Approximate Nearest Neighbor (ANN) Search:** Efficient algorithms like FAISS and HNSW designed for high-speed, high-dimensional vector search at scale.
- **In-Batch Negatives:** A memory-efficient training strategy where positive passages for other queries within the same computational batch serve as negatives for the current query.
- **Vector Quantization (VQ):** Compression methods that replace full-precision vectors with discrete codes to reduce memory overhead.

3.1.3 Evaluation and Infrastructure

- **FAISS (Facebook AI Similarity Search):** An open-source library optimized for efficient similarity search and clustering of dense vectors. It is the foundational infrastructure for billion-scale vector search, supporting both exact and approximate search (ANN).

- **MS MARCO (Microsoft Machine Reading Comprehension):** A large-scale dataset based on real Bing search queries. It serves as the primary benchmark for training and evaluating modern passage ranking and retrieval models.
- **BEIR Benchmark:** A diverse evaluation suite used to measure the **zero-shot** generalization of retrieval models across multiple domains (e.g., medical, legal, financial) without task-specific fine-tuning.
- **Bi-Encoder vs. Cross-Encoder:** A fundamental architectural distinction in IR. *Bi-Encoders* map queries and documents to a shared vector space for fast retrieval, while *Cross-Encoders* process query-document pairs simultaneously for higher accuracy at the cost of significantly higher computational latency.

3.1.4 Evaluation Metrics

- **MRR@k (Mean Reciprocal Rank):** A measure of search quality that calculates the reciprocal of the rank of the first relevant document.
- **nDCG (Normalized Discounted Cumulative Gain):** A metric that accounts for the graded relevance of documents and their positions in the result list, rewarding systems that place highly relevant documents at the top.

3.2 Theoretical Foundations

3.2.1 Contrastive Representation Learning

Contrastive learning creates embedding spaces where semantic similarity reflects geometric proximity. The InfoNCE loss is standard:

$$\mathcal{L}(q) = -\log \frac{\exp(\text{sim}(q, p^+)/\tau)}{\exp(\text{sim}(q, p^+)/\tau) + \sum_{i=1}^{N} \exp(\text{sim}(q, p_i^-)/\tau)}$$

where $\text{sim}(\cdot, \cdot)$ is typically dot product, τ is temperature, p^+ is the positive passage, and p_i^- are negatives. This encourages higher similarity to positives than negatives. Hard negative mining (retrieving top-k wrong passages using the current model) improves discrimination.

3.2.2 Billion-Scale Vector Search

Efficiently searching across massive datasets requires **ANN algorithms** to bypass the computational cost of exhaustive search.

- **IVF-PQ (Inverted File with Product Quantization):** Partitions the vector space into clusters and compresses vectors into compact codes for accelerated distance computation.
- **HNSW (Hierarchical Navigable Small World):** Utilizes graph-based indexing to provide high-recall, low-latency search capabilities.

REVIEW OF EXISTING WORK

4.1 Review of Existing Work

4.1.1 Organization

This chapter reviews selected papers in four categories: Dense Retrieval Foundations (core embedding methods), Scalability and Vector Search (billion-scale deployment), RAG Architectures and Optimization (retrieval-generation integration), and Hybrid and Advanced Methods (combining techniques). Each paper is described with its method, datasets, and results, followed by category-level comparison.

4.2 Dense Retrieval Foundations

4.2.1 Dense Passage Retrieval

Karpukhin et al. [1] introduced DPR using dual BERT encoders to map queries and passages into a shared 768-dimensional vector space with similarity computed via inner product. Training uses contrastive learning with one positive and multiple negatives per query (random passages, BM25 non-answers, in-batch negatives).

On Natural Questions with 21 million Wikipedia passages, DPR achieves 78.4% top-20 accuracy (BM25: 59.1%) and 41.5% exact match for end-to-end QA (ORQA: 33.3%). Training with just 1,000 pairs beats BM25. FAISS-accelerated inference processes 995 questions/second, though indexing requires 17 hours on 8 GPUs.

4.2.2 Representation Bottleneck Pre-training

Wang et al. [3] proposed SIMLM, compressing passage hidden states to 128 dimensions during pre-training. The model must predict original words from this bottleneck via masked language modeling, forcing document-level semantic capture. Pre-training is self-supervised, followed by supervised fine-tuning.

SIMLM achieves 41% MRR@10 on MS MARCO and shows strong zero-shot transfer on BEIR, consistently beating standard BERT initialization.

4.2.3 Comprehensive Survey of Semantic Models

Guo et al. [8] surveyed three decades of semantic retrieval research, organizing methods into lexical (BM25, sparse vectors with inverted indices), dense (neural approaches like DPR with continuous vectors and ANN search), and hybrid (combining both). Key findings: dense methods excel at semantics but need more computation; hybrid approaches often balance effectiveness and efficiency best. Six future directions identified including retrieval-specific pre-training and learned indexing.

4.2.4 Adversarial Vulnerabilities in Dense Retrieval

Zhong et al. [10] exposed security vulnerabilities through corpus poisoning. Attacks use gradient-based optimization to iteratively replace tokens, maximizing similarity between adversarial embeddings and target queries. K-means clustering generates diverse adversarial passages.

Results show severe vulnerabilities: unsupervised retrievers (Contriever) achieve 84.2% attack success with one passage, reaching 99.4% with 50. Supervised retrievers (DPR) show 60% success with 500 passages (0.02% of corpus). Attacks transfer across domains (FiQA: 94.1%, Quora: 97.2%). Multi-vector retrievers (ColBERT) show better resistance (11.5%). Defenses include GPT-2 likelihood filtering and embedding norm clipping (reduces attack from 99.4% to 0.6% with 6% quality loss).

4.2.5 Comparative Analysis

Table 4.1 compares foundational papers. Training method matters more than architecture complexity—DPR's simple dual-encoder beats complex designs through better negative sampling. Generalization receives increasing emphasis (SIMLM, survey). Security analysis shows adversarial robustness is critical. Progression moves from purely supervised toward combining self-supervised pre-training with supervised fine-tuning.

Table 4.1. Comparison of Dense Retrieval Foundation Papers

Paper	Primary Contribution	Datasets	Key Result
[1]	Dual-encoder with in-batch negatives	NQ, TriviaQA, WebQuestions	41% EM on NQ, 19+ point gain over BM25
[3]	Bottleneck pre-training	MS MARCO, BEIR	41% MRR@10, strong zero-shot transfer
[8]	Comprehensive survey	MS MARCO, TREC, BEIR	Taxonomy of 100+ papers, 6 future directions
[10]	Corpus poisoning attacks	NQ, MS MARCO, BEIR	99% attack success (unsupervised), 60% (supervised)

4.3 Scalability and Vector Search

4.3.1 Incremental Updates for Billion-Scale Search

Xu et al. [11] introduced SPFresh with Lightweight Incremental Rebalancing (LIRE) for efficient updates. When partitions overload, LIRE splits overflowing partitions locally and reassigns only boundary vectors whose nearest centroid changes, avoiding global recomputation.

On Yandex Deep1B (one billion 96-dimensional vectors), SPFresh achieves 2-5x faster updates than FreshDiskANN while maintaining 95% recall@10, using 1/100th resources of full rebuilds.

4.3.2 Learning Retrieval-Oriented Vector Quantization

Xiao et al. [12] proposed Distill-VQ, where quantization codes are trained to preserve a frozen teacher retriever's ranking using ListNet or LambdaRank losses, optimizing retrieval quality rather than reconstruction

error.

On MS MARCO and Natural Questions, Distill-VQ shows 2-5% MRR and recall improvements versus standard product quantization at identical compression ratios, particularly for highly compressed representations.

4.3.3 Compressed Concatenation of Embedding Models

Ayoub et al. [13] explored combining multiple small models (e.g., E5, GTE with 33M parameters) through compressed concatenation. Lightweight MLP decoders compress concatenated vectors using Matryoshka Representation Learning loss, followed by per-dimension percentile-based quantization.

On MTEB tasks (NFCorpus, SciFact), concatenating four models and compressing to 1/48th original size recovers 89% performance, outperforming single models with less storage than uncompressed concatenation.

4.3.4 Disaggregated Architecture for Vector Databases

Hu et al. [7] introduced HAKES with two-stage search: fast filtering using compressed representations (learned dimensionality reduction plus 4-bit quantization), and precise refinement using full-precision vectors. Compression parameters optimize similarity distribution preservation via KL-divergence loss.

The architecture separates IndexWorkers (filtering) from RefineWorkers (refinement) for independent scaling. On DPR-768 and OpenAI-1536 embeddings under concurrent workloads, HAKES achieves 16x throughput versus Weaviate and Milvus.

4.3.5 Comparative Analysis

Table 4.2 compares scalability solutions. Learning-based optimization increasingly replaces hand-engineering (Distill-VQ learns quantization, HAKES learns compression, ColBERTv2 learns compression-robust embeddings). Systems decouple components by requirements (HAKES: write path vs read path). Production requires co-design across algorithms, data structures, and architecture—HAKES shows multiplicative gains from combining learned compression with disaggregation.

Table 4.2. Comparison of Scalability and Vector Search Papers

Paper	Primary Contribution	Datasets	Key Result
[11]	Incremental rebalancing	Yandex Deep1B, custom	2-5x faster updates, 95% recall maintained
[12]	Ranking-aware quantization	MS MARCO, NQ	2-5% MRR/recall improvement
[13]	Compressed concatenation	MTEB subsets	89% performance at 48x compression
[7]	Disaggregated database	DPR-768, OpenAI-1536	16x throughput improvement

4.4 RAG Architectures and Optimization

4.4.1 Survey of RAG for AI-Generated Content

Zhao et al. [9] surveyed RAG across modalities and applications, organizing into foundations (retrieval mechanisms, integration approaches), enhancements (efficiency, robustness, hallucination reduction), and applications (text, image, code). Key trends: adaptive retrieval, multi-hop reasoning, multimodal expansion. Identifies hallucination reduction as central RAG motivation.

4.4.2 Heterogeneous Hardware Acceleration

Jiang et al. [6] paired FPGAs for vector search (IVF-PQ with near-memory processing) with GPUs for transformer inference, coordinated by CPUs for independent scaling. On databases from 1GB to 92TB, Chameleon achieves 2.16x latency reduction and 3.18x throughput versus CPU-GPU baselines.

4.4.3 Self-Reflective Generation

Asai et al. [5] proposed SELF-RAG, training models to generate reflection tokens: [Retrieve] triggers retrieval, [Relevant]/[Irrelevant] assess documents, [Supported]/[Not Supported] check grounding, [Utility] evaluates quality. A critic model annotates training data automatically.

On PubMedQA, PopQA, TriviaQA, SELF-RAG improves factuality and citation accuracy while reducing unnecessary retrievals by 50%, substantially reducing hallucinations.

4.4.4 Zero-Shot Query Expansion

Wang et al. [14] proposed Query2doc: LLMs generate hypothetical relevant documents, concatenated with queries before retrieval. Zero-shot (no training), works with sparse (BM25) and dense retrievers.

On MS MARCO and TREC DL, Query2doc provides up to 15% improvement on BM25, particularly for short, ambiguous queries.

4.4.5 Unsupervised LLM Adaptation

Liu et al. [4] adapted Llama-2 for retrieval without labeled data. Generates synthetic query-passage pairs (LLM creates questions for passages), trains with contrastive loss using adapter layers.

On BEIR, Llama2Vec matches or exceeds supervised retrievers on out-of-domain tasks despite entirely synthetic training data.

4.4.6 Comparative Analysis

Table 4.3 compares RAG papers. Adaptive retrieval replaces static approaches (SELF-RAG learns when to retrieve, Query2doc expands intelligently). LLMs improve retrieval through expansion and synthetic data.

Practical deployment requires system design (Chameleon's heterogeneous architecture). Pre-trained LLM knowledge with retrieval appears synergistic.

Table 4.3. Comparison of RAG Architecture and Optimization Papers

Paper	Primary Contribution	Datasets	Key Result
[9]	RAG survey and taxonomy	Various (surveyed)	Organizes RAG landscape, identifies hallucination mitigation as key
[6]	FPGA-GPU heterogeneous system	Custom RALM configs	2.16x latency reduction, 3.18x throughput
[5]	Self-reflective generation	PubMedQA, PopQA, TriviaQA	Improved factuality, 50% fewer retrievals
[14]	LLM-based query expansion	MS MARCO, TREC DL	Up to 15% improvement on BM25
[4]	Unsupervised LLM adaptation	BEIR	Matches supervised retrievers zero-shot

4.5 Hybrid and Advanced Methods

4.5.1 Enhanced Sparse Neural Retrieval

Formal et al. [15] demonstrated sparse retrieval can match dense methods with modern training. SPLADE++ represents documents as sparse vocabulary vectors with learned weights, applying distillation from cross-encoders (MarginMSE loss), ensemble hard negative mining, and CoCondenser pre-training. FLOPS regularization controls sparsity.

Achieves 38.0% MRR@10 on MS MARCO (competitive with dense) and 50.5-50.7 nDCG@10 on BEIR (exceeds ColBERTv2's 49.7), suggesting better out-of-domain generalization while maintaining inverted index compatibility.

4.5.2 Hypothetical Document Embeddings

Gao et al. [16] introduced HyDE: LLMs generate hypothetical answer documents, encode these (not queries), search for similar real documents. Zero-shot, requires no training.

On BEIR, HyDE provides 5-15% improvement over direct query encoding, larger gains on knowledge-intensive tasks. With unsupervised encoders, sometimes outperforms supervised retrievers.

4.5.3 Late Interaction with Residual Compression

Santhanam et al. [2] improved ColBERT through residual compression: cluster corpus token embeddings, store centroid index plus quantized residual (difference from centroid). MaxSim operates on reconstructed approximations. Training uses straight-through estimators for compression-robust embeddings.

Reduces index size 6-10x versus uncompressed ColBERT while maintaining or improving accuracy on MS MARCO and BEIR, demonstrating multi-vector retrieval practicality at scale.

4.5.4 Comparative Analysis

Table 4.4 compares hybrid methods. Sparse-dense distinction blurs (SPLADE++ achieves dense-competitive performance with sparse structure). Zero-shot emphasis increases across papers. Compression proves critical (ColBERTv2's residual compression enables multi-vector at scale). Creative combinations yield benefits (HyDE's generation-retrieval composition, SELF-RAG's reflection).

Table 4.4. Comparison of Hybrid and Advanced Methods Papers

Pape	Primary Contribution	Datasets	Key Result
[15]	Enhanced sparse retrieval	MS MARCO, BEIR	38.0% MRR@10, 50.7 nDCG@10 (BEIR)
[16]	Hypothetical documents	BEIR	5-15% zero-shot improvement
[2]	Late interaction with compression	MS MARCO, BEIR	6-10x storage reduction

4.6 Synthesis Across Categories

Several themes emerge across categories. Zero-shot and out-of-domain performance become standard criteria. Algorithmic innovations require system-level optimizations for web-scale deployment. Retrieval and generation increasingly intertwine. Combining approaches often outperforms pure methods.

Learning-based techniques optimize traditionally hand-engineered components (Distill-VQ: quantization, HAKES: compression, SPLADE++: term weights, ColBERTv2: compression-robust embeddings). Training methodology grows sophisticated (DPR's in-batch negatives $\rightarrow$ dynamic hard negatives $\rightarrow$ ensemble negatives). Simple architectures with sophisticated training consistently beat complex architectures with basic training.

Security and robustness emerge as first-order concerns (corpus poisoning reveals severe vulnerabilities). RAG emphasis on factuality reflects awareness that deployed systems must be trustworthy beyond benchmarks.

LLM integration expands: query expanders (Query2doc), synthetic data generators (Llama2Vec), hypothetical document generators (HyDE), self-reflective generators (SELF-RAG). The retrieval-generation boundary dissolves, suggesting future systems will integrate both seamlessly rather than treating as separate stages.

CRITICAL ANALYSIS AND DISCUSSION

This chapter provides a critical evaluation of the surveyed approaches, highlighting their strengths and limitations, identifying major research gaps, and discussing key challenges in deployment, robustness, and ethics.

5.1 Strengths of Existing Approaches

Dense retrieval methods have established strong semantic matching capabilities. DPR demonstrated the effectiveness of supervised contrastive training with dual encoders, consistently outperforming sparse baselines on knowledge-intensive tasks. SIMLM further improved generalization through bottleneck pre-training, enabling better zero-shot performance across diverse domains.

Scalability solutions effectively support real-world deployment. SPFresh offers practical incremental updates with low overhead and stable recall on billion-scale corpora. HAKES achieves high throughput under concurrent read-write loads via disaggregated architecture and learned compression. Compression techniques such as Distill-VQ and Compressed Concatenation enable significant memory reduction (up to 48×) while preserving retrieval quality, facilitating deployment in resource-constrained environments.

RAG architectures markedly enhance generative reliability. SELF-RAG introduces self-reflection tokens that allow dynamic retrieval decisions, relevance assessment, and output critique, leading to improved factuality. Query2doc and HyDE provide effective zero-shot query enhancement using LLMs, frequently surpassing supervised retrievers on out-of-domain benchmarks. Llama2Vec shows unsupervised adaptation of open LLMs into high-quality dense retrievers, reducing dependence on labeled data.

Hybrid methods balance precision, efficiency, and robustness. ColBERTv2 preserves token-level detail through late interaction and lightweight compression. SPLADE++ successfully applies dense training strategies to sparse models, achieving competitive in-domain and zero-shot results. Comprehensive RAG surveys offer valuable synthesis and point to emerging directions such as multimodal and agentic extensions.

Overall, the literature reflects a clear evolution from embedding accuracy to scalable infrastructure, adaptive generation, and production-oriented systems.

5.2 Weaknesses and Limitations

Dense retrieval approaches remain computationally intensive for training and indexing, particularly when incorporating hard negative mining or large-scale contrastive objectives. Zero-shot performance, while advanced, shows inconsistency on highly specialized or long-tail domains.

Scalability techniques often require specialized hardware or complex pipelines, limiting accessibility. Incremental update methods can introduce temporary accuracy fluctuations during rebalancing. Compression strategies occasionally degrade performance on challenging edge cases.

RAG systems are vulnerable to propagation of retrieval noise into generation. Reflection-based methods increase inference latency. Query enhancement techniques can inherit LLM biases or inconsistencies. Multimodal extensions remain limited, with most work focused on text-only settings.

Security vulnerabilities are a major concern. The poisoning attack study shows that a very small fraction of adversarial passages (0.02% of corpus) can mislead dense retrievers on in-domain and out-of-domain queries, with unsupervised models exhibiting extreme susceptibility. This poses significant risks for open-access platforms.

Ethical issues, including bias amplification from retrieved content and lack of transparency in retrieval-generation pipelines, are insufficiently addressed.

5.3 Research Gaps

Several critical gaps remain:

- Robust defenses against corpus poisoning and embedding attacks (adversarial training, detection, certification). - Multimodal and cross-modal retrieval capabilities for richer applications. - Systematic analysis of bias, fairness, and ethical implications in embeddings and RAG outputs. - Sustainable trillion-scale systems with real-time updates and low energy consumption. - Efficient on-device semantic search and RAG for edge environments. - Support for long-context documents and multi-hop reasoning over large corpora.

5.4 Interpretation of Trends and Insights

The field shows a clear shift toward unsupervised and zero-shot paradigms, production-oriented engineering (hardware acceleration, disaggregation, incremental updates), adaptive self-improving retrieval, and hybridization for balanced performance.

The progression from accuracy-focused research to deployable, reliable systems is evident. Future advancements will likely combine agentic reasoning, multimodal understanding, and strong robustness measures. Addressing poisoning vulnerabilities, ethical concerns, and sustainability will be essential for responsible adoption in large-scale real-world environments.

TRENDS AND FUTURE DIRECTION

6.1 Emerging Technologies and Methods

Recent developments demonstrate technological advances addressing practical deployment challenges across hardware acceleration, unsupervised learning, compression, and retrieval-generation integration.

Hardware acceleration through heterogeneous architectures enables production deployment. Chameleon combines FPGAs (vector search with IVF-PQ) and GPUs (transformer inference), achieving 2.16x latency reduction and 3.18x throughput on databases up to 92TB. HAKES separates filtering (IndexWorkers with compressed representations) from refinement (RefineWorkers with full precision), achieving 16x throughput under concurrent workloads through learned compression preserving similarity distributions.

Unsupervised and zero-shot techniques reduce labeled data dependence. Llama2Vec generates synthetic training data via LLM, achieving competitive zero-shot BEIR performance matching supervised baselines. Query2doc and HyDE leverage LLM generation for query expansion and hypothetical document creation, improving zero-shot retrieval by 5-15% by bridging vocabulary gaps.

Compression and efficiency optimizations enable resource-constrained deployment. ColBERTv2 reduces multi-vector storage 6-10x through residual compression (centroid plus quantized residual). SPLADE++ achieves 38.0% MRR@10 on MS MARCO and superior BEIR transfer through distillation, hard negatives, and proper pre-training. Compressed concatenation recovers 89% performance at 48x compression using Matryoshka Representation Learning.

Retrieval-generation integration advances through self-reflection. SELF-RAG generates reflection tokens controlling retrieval, assessing relevance, verifying support, and evaluating quality, reducing unnecessary retrievals by 50% while improving factuality. Surveys identify trends including multimodal RAG, agentic multi-step reasoning, and hierarchical retrieval structures.

Security considerations emerge following vulnerability revelations. Corpus poisoning demonstrates 50 adversarial passages mislead 94% of queries even out-of-domain, motivating defenses including embedding norm clipping, likelihood filtering, and hybrid architectures.

6.2 Open Problems for Future Research

Several fundamental challenges present opportunities for investigation:

Scalability beyond billion-scale requires new indexing and maintenance approaches. Trillion-scale deployment faces computational cost and energy challenges. Research directions include efficient encoding models reducing dimensionality, learned index structures jointly optimizing encoding and indexing, and distributed architectures maintaining search quality.

Robustness against adversarial manipulation is critical for user-contributed platforms. Dense retriever vulnerability to corpus poisoning, particularly unsupervised models and out-of-domain transfer, necessitates defensive techniques including adversarial training, certified defenses, anomaly detection, and robust architectural choices like multi-vector or hybrid systems.

Bias and fairness require systematic investigation. Retrieved content can reflect and amplify training data biases, affecting RAG outputs and lending credibility to biased information through citations. Research directions include bias measurement methodologies, debiasing techniques for embeddings, and fairness-aware ranking with demographic parity constraints.

Multimodal retrieval remains underdeveloped versus text-only systems. Extending techniques like SIMLM's bottleneck or ColBERT's late interaction to multimodal settings requires addressing representation and indexing of mixed-modality documents, cross-modal queries, and multimodal evaluation metrics.

Zero-shot generalization shows promise but inconsistent effectiveness across datasets and query types. Long-tail queries involving rare entities exhibit poor performance. Research directions include understanding transfer success factors, domain adaptation with minimal labeled data, and architectures optimized for transfer.

Sustainable deployment becomes critical as systems scale. Encoding billions of documents and serving millions of queries requires substantial energy. Questions include environmental impact versus classical methods and optimizing effectiveness-efficiency-sustainability trade-offs.

Evaluation beyond benchmarks is necessary for real-world assessment. MS MARCO and BEIR don't capture incremental updates, distribution shift, adversarial robustness, and cost-effectiveness. Comprehensive evaluation frameworks would better guide research toward practical impact.

6.3 Opportunities and Applications

Advances enable applications requiring reliable information access and reasoning:

Agentic retrieval systems perform multi-step information gathering, including medical diagnosis support (clinical guidelines, literature, patient records), legal research (case law synthesis across jurisdictions), and scientific review (cross-disciplinary connections).

Healthcare applications ground responses in medical literature and databases, enabling clinical decision support, drug interaction checking, and patient education with verifiable citations.

Financial applications leverage real-time retrieval for market intelligence, regulatory compliance, and fraud detection through historical pattern analysis.

Educational technology provides personalized learning through intelligent tutoring systems, research assistants, and curriculum development tools with adaptive content retrieval.

Edge deployment through compressed models enables mobile search without cloud connectivity, privacy-preserving local processing, and embedded systems for vehicles and IoT devices.

Enterprise knowledge management provides unified semantic search across documents, emails, databases, and code repositories for technical support, project management, and compliance tracking.

Scientific research acceleration assists systematic reviews, research trend analysis, and interdisciplinary connection discovery across exponentially growing literature.

Integration of hybrid sparse-dense retrieval, self-reflective generation, and disaggregated architecture could yield production systems balancing effectiveness, efficiency, and reliability. Unsupervised methods reduce deployment barriers while hardware acceleration enables real-time processing at scale.

CONCLUSION

7.1 Summary of Contributions

This survey has provided a comprehensive examination of semantic search over big data using large language models, synthesizing findings across thirteen papers spanning dense retrieval foundations, scalability optimizations, retrieval-augmented generation, and hybrid methods. The primary contributions of this work include establishing a unified framework for understanding and comparing diverse retrieval approaches, providing detailed analysis of training methodologies and their impact on effectiveness, synthesizing findings on the efficiency-effectiveness trade-offs inherent in different paradigms, revealing security vulnerabilities in dense retrieval systems and potential defenses, and identifying emerging trends including unsupervised learning, hardware acceleration, and retrieval-generation integration.

The categorization of reviewed works into four coherent groups enables systematic comparison and understanding of the landscape. Dense retrieval foundations (DPR, SIMLM, survey papers, corpus poisoning) establish core principles and reveal fundamental characteristics of neural retrieval including training methodology dominance, sample efficiency, and adversarial vulnerability. Scalability and vector search papers (SPFresh, Distill-VQ, compressed concatenation, HAKES) address practical deployment challenges through incremental indexing, learned compression, model ensembling, and disaggregated architecture. RAG architectures and optimization works (surveys, Chameleon, SELF-RAG, Query2doc, Llama2Vec) demonstrate increasing sophistication in retrieval-generation integration through adaptive retrieval, query enhancement, and unsupervised adaptation. Hybrid and advanced methods (SPLADE++, HyDE, ColBERTv2) show that combining paradigms or applying sophisticated optimizations yields effectiveness and efficiency gains.

The analysis has revealed several key insights. Training methodology including in-batch negative training, hard negative mining, and distillation appears more impactful than architectural complexity for retrieval effectiveness. Sample efficiency is remarkable, with DPR demonstrating that 1,000 examples suffice to outperform classical methods. Generalization capability varies substantially across methods, with sparse neural approaches often exhibiting better zero-shot transfer than dense methods despite lower in-domain performance. Security vulnerabilities in dense retrieval are severe, particularly for unsupervised models, with implications for deployment on platforms accepting user-contributed content. Hybrid approaches combining sparse and dense retrieval consistently outperform pure methods and provide defense-in-depth against adversarial attacks.

7.2 Research Gaps and Future Directions

Despite substantial progress, significant gaps remain. Multimodal retrieval beyond text is underdeveloped, with most work focusing exclusively on textual content. Ethical considerations including bias measurement, fairness constraints, and transparency mechanisms receive insufficient attention given the societal impact of deployed retrieval systems. Robustness against adversarial manipulation requires further investigation including certified defenses and adversarially robust training procedures. Evaluation methodologies inadequately capture important deployment considerations including cost-effectiveness, energy efficiency, and performance under distribution shift.

Future research directions include developing retrieval techniques for trillion-scale collections through more efficient encoding, learned index structures, and distributed architectures. Investigating robustness mechanisms providing provable guarantees against adversarial attacks would enable safer deployment. Extending successful techniques to multimodal settings covering text, images, tables, code, and structured data would broaden applicability. Improving zero-shot generalization through better understanding of domain characteristics and development of effective domain adaptation techniques requiring minimal labeled data would democratize access to neural retrieval. Addressing sustainability through energy-efficient architectures and training procedures would reduce environmental impact.

The shift toward unsupervised methods and zero-shot approaches creates opportunities for deployment in domains and languages currently lacking large labeled datasets. Hardware acceleration through heterogeneous and disaggregated architectures signals increasing maturity and production readiness. The integration of retrieval and generation in RAG systems enables applications requiring factual grounding and knowledge updates beyond model parameters. These trends suggest that semantic search over big data is transitioning from research exploration to practical deployment addressing real-world information access challenges.

7.3 Concluding Remarks

Semantic search over big data using large language models has progressed substantially from early neural ranking approaches to sophisticated systems integrating retrieval and generation. The reviewed works demonstrate that dense retrieval can outperform classical term-based methods through learned representations capturing semantic similarity, that billion-scale deployment is feasible through careful system design and compression techniques, that retrieval-augmented generation improves factuality and enables knowledge updates, and that hybrid approaches combining multiple paradigms often achieve optimal trade-offs.

However, challenges remain including computational costs, adversarial vulnerabilities, generalization inconsistencies, and ethical considerations. The field must address these limitations while building on established strengths to create systems that are not only effective but also efficient, robust, fair, and sustainable. The integration of trends identified in this survey including unsupervised learning, hardware acceleration, hybrid architectures, and self-reflective generation suggests promising directions for future development.

As these systems increasingly mediate information access for billions of users, careful attention to effectiveness, efficiency, robustness, fairness, and transparency becomes paramount. The foundations established by the reviewed works provide a solid basis for continued innovation addressing both technical challenges and societal implications of semantic search at scale.

BIBLIOGRAPHY

[1] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[2] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, "ColBERTv2: Effective and efficient retrieval via lightweight late interaction," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2022.

[3] L. Wang et al., "SIMLM: Pre-training with representation bottleneck for dense passage retrieval," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[4] Z. Liu, C. Li, S. Xiao, Y. Shao, and D. Lian, "Llama2vec: Unsupervised adaptation of large language models for dense retrieval," *arXiv preprint arXiv:2312.15503*, 2023.

[5] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "SELF-RAG: Learning to retrieve, generate, and critique through self-reflection," *arXiv preprint arXiv:2310.11511*, 2024.

[6] W. Jiang, M. Zeller, R. Waleffe, T. Hoefler, and G. Alonso, "Chameleon: A heterogeneous and disaggregated accelerator system for retrieval-augmented language models," in *Proceedings of the VLDB Endowment*, 2023.

[7] G. Hu et al., "HAKES: Scalable vector database for embedding search service," *arXiv preprint arXiv:2505.12524*, 2025.

[8] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng, "Semantic models for the first-stage retrieval: A comprehensive review," *ACM Transactions on Information Systems*, 2022.

[9] P. Zhao et al., "Retrieval-augmented generation for AI-generated content: A survey," *Data Science and Engineering*, 2026.

[10] Z. Zhong, Z. Huang, A. Wettig, and D. Chen, "Poisoning retrieval corpora by injecting adversarial passages," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[11] Y. Xu et al., "SPFresh: Incremental in-place update for billion-scale vector search," in *Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP)*, 2023.

[12] S. Xiao et al., "Distill-VQ: Learning retrieval oriented vector quantization by distilling knowledge from dense embeddings," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

[13] B. A. M. Ayoub, M. Dinzinger, K. G. Dastidar, J. Mitrović, and M. Granitzer, "Compressed concatenation of small embedding models," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2025.

[14] L. Wang, N. Yang, and F. Wei, "Query2doc: Query expansion with large language models," *arXiv preprint arXiv:2303.07678*, 2023.

[15] T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant, "From distillation to hard negative sampling: Making sparse neural IR models more effective," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

[16] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.