SCIENTIFIC Publishing Limited
Journal of Intelligent Communication
https://ojs.ukscip.com/index.php/jic

Review

BERT and Beyond: A Comprehensive Survey of Natural Language Processing Techniques for Information Retrieval

Mokhtar H. Al-Sarori [1,*], Mubarak Mohammed Al-Ezzi Sufyan [2], Mahfoudh Al-Asaly [3] and Ghassan Abdullah Abdulwasea Al-Maamari [1]

[1] Faculty of Information Technology & Computer Science, University of Saba Region, Marib, Yemen
[2] Department of Computer Information Systems, Al-Jawf Faculty, University of Saba Region, Marib, Yemen
[3] Department of Information Technology, College of Computer, Qassim University, Buraydah 51174, Saudi Arabia
[*] Correspondence: msorori201201@gmail.com

Received: 10 July 2025; **Revised:** 18 August 2025; **Accepted:** 21 August 2025; **Published:** 6 September 2025

Abstract: Information Retrieval (IR) has undergone a profound transformation in the field of Natural Language Processing (NLP), shifting from traditional keyword-based approaches to neural architectures and, more recently, to advanced generative Large Language Models (LLMs). Transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) have substantially improved semantic understanding and retrieval accuracy by enabling contextualized embeddings, deeper interaction modeling, and more effective ranking mechanisms. The rise of Retrieval-Augmented Generation (RAG) represents a significant development by integrating retrieval with generation to produce factually-grounded, context-aware, and explainable outputs, while reducing the likelihood of hallucinations commonly associated with LLMs. This survey provides a comprehensive review of modern IR techniques, focusing on BERT-based retrieval models, emerging generative retrieval frameworks, evaluation methodologies, and key application domains. We provide a structured taxonomy of IR methods and conduct a comparative analysis of state-of-the-art research to highlight performance trends and methodological distinctions. Ongoing challenges, including scalability, computational efficiency, interpretability, and limitations of current evaluation benchmarks, are critically discussed. Additionally, the survey explores emerging directions such as multimodal and cross-modal retrieval, hybrid dense-sparse architectures, knowledge-graph-enhanced retrieval, and the integration of LLMs as unified retriever-generator systems. These advancements illustrate the rapid evolution of IR. They also underscore the need for adaptive, reliable, and transparent retrieval solutions in increasingly complex information environments. This work aims to provide researchers and practitioners with a clear and organized overview of the evolution, current landscape, and future research opportunities in IR within the era of LLM-driven NLP.

Keywords: Information Retrieval; Generative Models; Retrieval-Augmented Generation; Natural Language Processing; Transformer Models; Semantic Search; Deep Retrieval Architectures

1. Introduction

Information Retrieval (IR) is a fundamental component of Natural Language Processing (NLP), enabling diverse applications, such as search engines, question answering, recommendation systems, and domain-specific knowledge discovery. Traditional IR approaches, including Boolean retrieval, vector space models (VSM), and BM25, primarily relied on keyword matching and statistical co-occurrence. While effective for basic search, these methods are limited in capturing semantic meaning, contextual nuances, and user intent [1,2]. The advent of neural IR models and transformer-based pre-trained language models (PLMs), particularly Bidirectional Encoder Representations from Transformers (BERT), revolutionized the field by introducing deep contextual embeddings. These models enabled more accurate passage ranking, document retrieval, and semantic similarity assessment, especially when fine-tuned on domain-specific datasets, further improving performance [1–4]. More recently, Large Language Models (LLMs) and generative retrieval techniques have extended IR beyond simple ranking to the generation of knowledge-intensive, contextually coherent outputs. In particular, Retrieval-Augmented Generation (RAG) combines dense retrieval with generative modeling, supporting interactive querying, structured data retrieval, and multimodal information access. Furthermore, IR techniques are now applied across healthcare, legal, scientific, and conversational domains [5–15]. Despite these advances, most surveys focus on either BERT-based IR or generative retrieval without providing a unified perspective that covers the evolution from traditional methods to generative LLMs. **Figure 1** illustrates this progression.

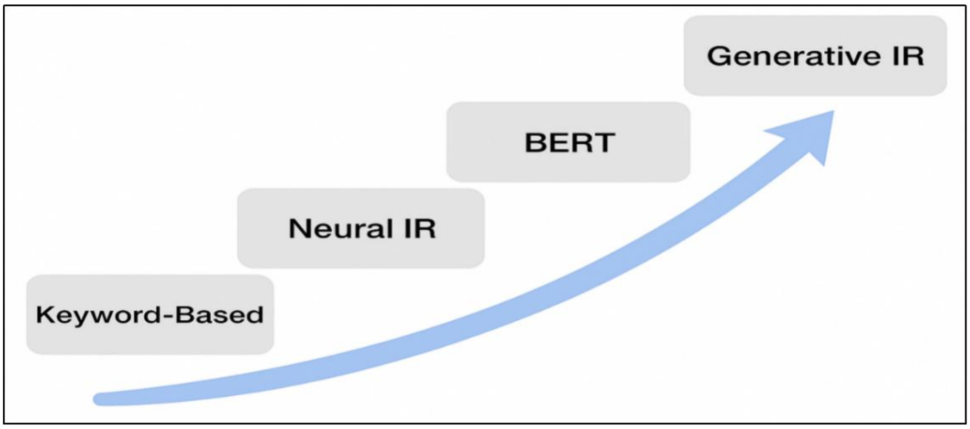

Figure 1. Evolution of TR in NLP from traditional methods to BERT and generative LLMs.

Despite these advances, significant challenges remain. High computational cost and scalability issues limit the practical deployment of large models in real-world applications. In addition, generative IR approaches frequently suffer from factual inconsistencies or hallucinations, especially in knowledge-intensive domains. Furthermore, existing evaluation metrics, such as Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE), are often insufficient for measuring semantic understanding and retrieval quality, making it difficult to benchmark and compare models effectively. While traditional IR techniques relying on keyword matching and shallow representations struggle to capture semantic meaning and context, modern PLMs and generative models offer powerful tools to understand, retrieve, and generate knowledge-rich content. However, integrating these models introduces new challenges related to scalability, factual consistency, and evaluation. This gap in the literature has motivated our comprehensive and structured survey to help researchers and practitioners understand the evolution, strengths, limitations, and future directions of IR in the era of large language models. This work provides an in-depth review of IR techniques, highlighting their architectures, evaluation frameworks, applications, and open challenges. The main contributions of this survey are summarized as follows:

- To review the evolution of Information Retrieval in NLP, from traditional methods to BERT and generative large language models.
- To analyze modern models by examining transformer-based PLMs and generative LLMs, highlighting architectures, embedding strategies, and applications in IR.
- Comparative Insights: Summarize performance, strengths, and limitations of traditional, contextual, and generative IR approaches.
- Challenges and Future Directions: identify open issues such as efficiency, factual consistency, and evaluation, and outline promising research avenues.

For reference, **Table 1** lists the abbreviations used in this work.

Table 1. List of abbreviations.

Acronym	Description
AI (XAI)	Artificial Intelligence (Explainable AI)
ALBERT	A Lite BERT
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
BM25	Okapi BM25
CNNs	Convolutional Neural Networks
DPR	Dense Passage Retrieval
GPUs	Graphics Processing Units
GRAG	Graph-enhanced RAG
IR	Information Retrieval
KGs	Knowledge Graphs
LLaMA	Large Language Model Meta AI
LLMs	Large Language Models
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
MAP	Mean Average Precision
MLM	Masked Language Modeling
NLP	Natural Language Processing
NDCG	Normalized Discounted Cumulative Gain
NSP	Next Sentence Prediction
NLP IR	Natural Language Processing Information Retrieval
PaLM	Pathways Language Model
PLMs	Pre-trained Language Models
QA	Question Answering
RAG	Retrieval-Augmented Generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RNNs	Recurrent Neural Networks
RoBERTa	Robustly Optimized BERT
SVD	Singular Value Decomposition
TF-IDF	Term Frequency-Inverse Document Frequency
TPU	Tensor Processing Unit
VSM	Vector Space Models

The paper is structured to deliver a comprehensive overview of the evolution and recent advancements in Information Retrieval (IR). Section 2 introduces the background and foundational concepts, followed by the research methodology in Section 3. Sections 4 through 7 present the key thematic areas: applications of IR and generative models across domains; BERT-based and Pre-trained Language Model (PLM) approaches; generative IR techniques; and the transition toward generative Large Language Models (LLMs). Section 8 further elaborates on the role of generative LLMs in IR. Section 9 outlines evaluation frameworks for IR and Retrieval-Augmented Generation (RAG), while Section 10 provides a comparative analysis. Section 11 offers a detailed discussion, and Section 12 addresses challenges, open issues, and future research directions. The paper concludes in Section 13.

2. Background and Foundations

The foundations of Information Retrieval are rooted in traditional keyword-based approaches. These gradually evolved into contextual and transformer-based models, laying the groundwork for modern generative methods illustrated in **Figure 2**, while the following subsections provide detailed insights into this evolution.

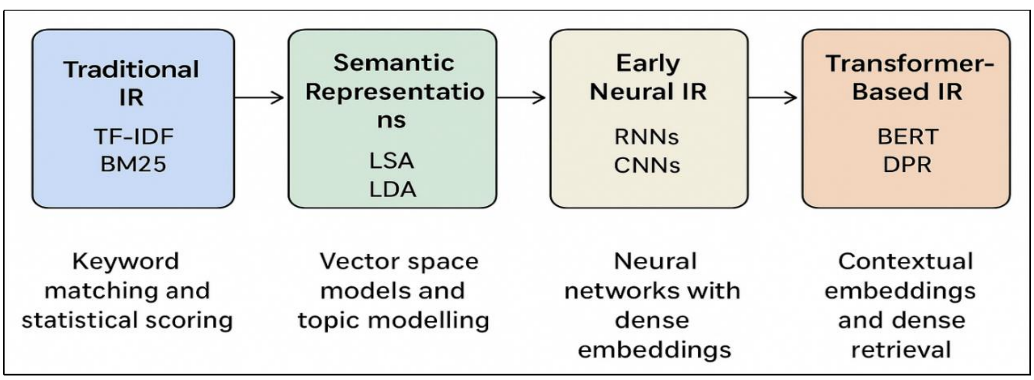

Figure 2. The foundations of information retrieval.

2.1. Traditional Information Retrieval Approaches

Early IR systems were predominantly based on keyword matching and statistical scoring techniques to assess the relevance of documents to user queries. Among these, Term Frequency-Inverse Document Frequency (TF-IDF) and BM25 (Okapi BM25) remain the most widely recognized methods [1, 2]. TF-IDF quantifies the importance of a term within a document relative to its occurrence across a corpus. Terms that appear frequently in a document receive higher scores, but this is counterbalanced by their prevalence across the entire collection, thereby reducing the impact of commonly used words. Despite its effectiveness for basic retrieval tasks, TF-IDF is inherently limited in understanding semantic relationships, struggling to retrieve relevant documents when synonyms or paraphrased expressions are used [1,2]. BM25, an evolution of TF-IDF, enhances document ranking by incorporating term frequency saturation and document length normalization. This prevents longer documents from being unduly favored and moderates the influence of very high term frequencies within a document. Moreover, BM25 operates on the bag-of-words assumption, lacking deep semantic comprehension and failing to capture nuanced contextual dependencies between words [1, 2]. Other traditional IR models include Boolean retrieval, which retrieves documents based on logical keyword operations (AND, OR, NOT), and Vector Space Models (VSM), which represent queries and documents as vectors and compute similarity via metrics such as cosine similarity. While computationally efficient, these approaches remain limited in handling polysemy, synonymy, and context-sensitive meaning, highlighting the need for semantic-aware retrieval methods [1,2].

2.2. Theoretical Foundations of Semantic Representations

The shortcomings of classical IR techniques motivated the development of semantic representations, aimed at encoding words, phrases, or documents into continuous vector spaces where semantic similarity is reflected by geometric distance. Early approaches included Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) [1,16]. LSA applies singular value decomposition (SVD) to term-document matrices, uncovering latent semantic structures. It partially addresses synonymy and polysemy but is limited by linearity constraints and high computational costs on large corpora [1, 16]. On the other hand, LDA is a generative probabilistic model representing documents as mixtures of topics and topics as distributions over words. Although effective in discovering abstract topics, LDA produces static representations that cannot dynamically capture word meaning in varying contexts [1,16].

2.3. NLP's Transition to Deep Learning

The emergence of deep learning marked a paradigm shift in NLP, moving beyond statistical and rule-based methods and toward models capable of learning complex hierarchical patterns from large datasets. Early neural architectures, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), demonstrated promise in tasks like sentiment analysis, machine translation, and named entity recognition [1,10,17]. RNNs, particularly Long Short-Term Memory (LSTM) networks, excel at modeling sequential dependencies in text, while CNNs effectively capture local features such as n-grams. However, both architectures struggle with modeling long-range dependencies and processing lengthy sequences efficiently [1,10,18]. The advent of the Transformer architecture revolutionized NLP by employing self-attention mechanisms, enabling simultaneous consideration of all tokens in a sequence. This innovation facilitated the creation of PLMs, which form the backbone of modern IR systems and enable sophisticated semantic search and question-answering capabilities [1,3,4,10].

In this regard, **Figure 3** shows the Taxonomy of Information Retrieval Evolution of this work, while **Table 2** summarizes some representative works in Information Retrieval.

Table 2. Summary of the background and foundations of information retrieval.

Category	Method/Model	Key Features	Strengths	Limitations	Refs.
Traditional IR	TF-IDF	Term weighting based on frequency in document vs. corpus	Simple, interpretable	Ignores semantics, synonyms, polysemy	Wang et al. and Gardazi et al. [1,2]
	BM25	Probabilistic ranking with term frequency & document length normalization	Improved ranking over TF-IDF	Lexical-level only, no semantic understanding	Wang et al. and Gardazi et al. [1,2]
	Boolean Retrieval	Logical keyword matching (AND, OR, NOT)	Simple, computationally cheap	Cannot rank by relevance, ignores semantics	Wang et al. [1]
	VSM	Documents and queries as vectors, similarity via cosine similarity	Provides ranking, captures basic similarity	Limited semantic understanding, relies on exact term match	Wang et al. [1]

Table 2. *Cont.*

Category	Method/Model	Key Features	Strengths	Limitations	Refs.
Semantic Representations	LSA	SVD-based dimensionality reduction for latent semantics	Captures hidden semantic relations	Linear, computationally heavy for large datasets	Garcia-Carmona. [6]
	LDA	Probabilistic topic modeling	Captures abstract topics	Static embeddings, context-insensitive	Garcia-Carmona [6]
Early Neural IR	Feedforward/CNN	Dense embeddings of queries/documents, and local feature extraction	Better semantic matching	Requires large labeled datasets	Wang et al. and Wu et al. [1,10]
	RNN/LSTM	Sequential modeling, captures long-range dependencies	Handles longer queries, contextual info	Inefficient for very long sequences	Wang et al. and Wu et al. [1,10]
Transformer-Based IR	Transformer	Self-attention for global context, parallelizable	Captures long-range dependencies, supports PLMs	Requires high computational resources	Zhang et al. [12]
	BERT	Bidirectional contextual embeddings, MLM & NSP pre-training	Captures deep semantics, supports fine-tuning	Large model, computationally intensive, low interpretability	Rojas-Carabali et al. [14]
	RoBERTa	Optimized BERT: no NSP, larger data & batch, better hyperparameters	Improved performance	Large model, resource-intensive	Wang et al. and Gardazi et al. [1,2]
	ALBERT	Parameter sharing & factorized embeddings	Efficient, smaller memory footprint	May require careful hyperparameter tuning	Wang et al. and Gardazi et al. [1,2]
Dense Retrieval	Dense Passage Retrieval (DPR)	Dual-encoder dense vector representations for queries & passages	Efficient semantic retrieval, scalable	Requires embedding index maintenance	Premasiri et al. and Yoran et al. [9,19]
Evaluation	BEIR Benchmark	Multi-domain IR evaluation dataset	Standardized evaluation, supports zero-shot testing	Limited to included datasets, benchmark may evolve	Qin et al. [13]

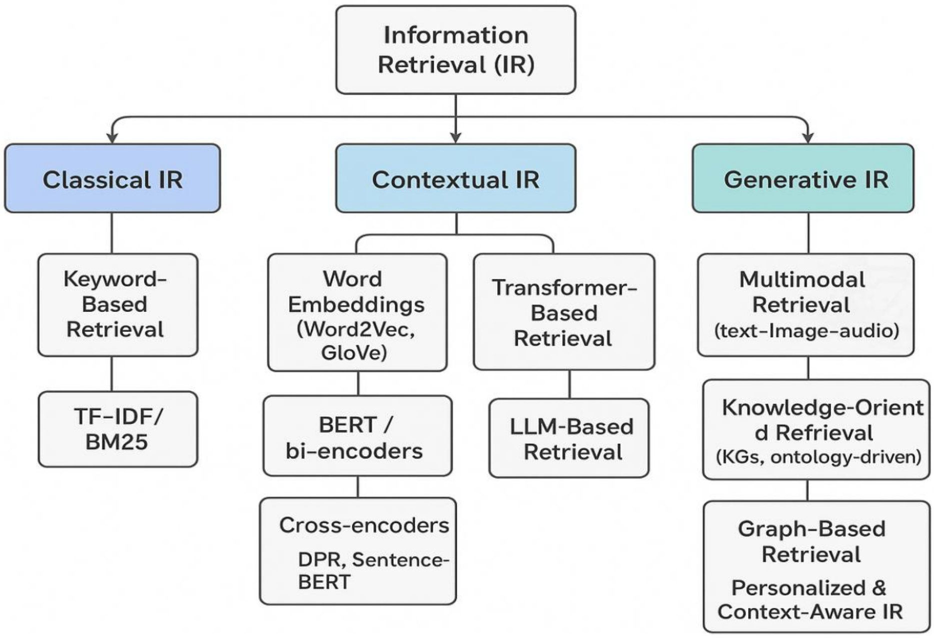

Figure 3. Taxonomy of information retrieval evolution.

3. Methodology

This survey follows a structured methodology, inspired by the PRISMA guidelines, to ensure a comprehensive, reproducible, and unbiased review of Natural Language Processing (NLP) techniques for Information Retrieval (IR), with emphasis on BERT, Transformer-based models, retrieval architectures, and recent advancements in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). The methodology is built around four phases: Identification, Screening, Eligibility, and Inclusion.

3.1. Search Strategy

A systematic search was conducted across major scientific databases: IEEE Xplore, ACM Digital Library, SpringerLink, Elsevier ScienceDirect, and arXiv (for high-impact preprints relevant to emerging LLM trends). The searches covered the period 2018–2025, corresponding to the era of Transformer dominance. The following query patterns were used: "BERT for Information Retrieval", "Transformer models NLP survey", "Large Language Models for IR", "Retrieval-Augmented Generation", "Generative Information Retrieval", and "Long-document retrieval LLMs". This process initially yielded 527 records.

3.2. Inclusion and Exclusion Criteria

Inclusion Criteria: Studies were included if they met the following conditions:

1. Address NLP techniques applied to IR (e.g., BERT, Transformers, PLMs, LLMs), e.g., Wang et al. [1], Gardazi et al. [2], Zhu et al. [4] and Xu et al. [18].
2. Survey papers focusing on retrieval models, RAG, generative IR, e.g., Wu et al. [10], Kuo et al. [16], Fan et al. [5].
3. Application-oriented IR/NLP research providing domain insights (e.g., in medicine [6,14], education [20], materials [21], and big data IR [22,23]).
4. Studies introducing benchmarks or retrieval datasets (e.g., Cocktail benchmark [24]).
5. Works examining model architectures, embedding, knowledge integration (e.g., Yang et al. [8], Y. Huang & J. X. Huang [7], and Genesis & Keane [25]).

Exclusion Criteria: Records were excluded if they: 1) Did not involve NLP or IR directly (e.g., pure linguistic studies); 2) Focused solely on non-Transformer classical IR methods; 3) Were duplicate versions of arXiv/preprint entries; or 4) Lacked technical depth or empirical relevance. And then after applying the criteria, 246 papers remained.

3.3. Screening Process

Titles and abstracts were screened to exclude papers on unrelated topics, such as speech-only systems, non-text modalities (with the exception for multimodal RAG [25]), and studies addressing general AI trends without relevance to information retrieval. This screening removed 163 records. Subsequently, a full-text evaluation of the remaining 83 papers yielded 40 studies that were ultimately included in the review, corresponding precisely to the reference list.

3.4. Quality Assessment

Each selected study was evaluated using a structured set of criteria, including technical depth (coverage of architectures, models, and algorithms), relevance to Information Retrieval (retrieval performance, LLM integration, and RAG workflows), and novelty (introduction of new frameworks, benchmarks, or taxonomies). Additional factors included the strength of empirical evidence (datasets used, evaluation metrics reported) and the clarity and reproducibility of methodological descriptions. Studies such as Wang et al., Gardazi et al., Zhu et al., Wu et al., Fan et al., Gan et al., Qin et al., Kuo et al., Xu et al., Lenadora, Genesis, Keane, and Li et al. [1,2,4,10,13,16,18,20,25–27], scored highly across these criteria due to their methodological rigor and comprehensive analyses.

3.5. Data Extraction and Synthesis

From each included study, the following key attributes were extracted:

- Model type: e.g., BERT, PLM, LLM, RAG, graph-based, multimodal, or hybrid retrieval [1,4,25–27]), Transformer reviews [3], graph-RAG [28,29], multimodal RAG [30])
- IR task focus: document ranking, QA, semantic retrieval, long-document retrieval (e.g., long-document retrieval [31])
- Application domain: medicine, education, materials science, social networks, etc.
- Evaluation metrics: nDCG, MRR, Recall@k, BLEU, ROUGE
- Challenges identified: model complexity, hallucination, incomplete grounding, scalability, etc.
- Proposed future directions: efficient LLMs, robust RAG, knowledge integration, multimodal fusion.

Information was synthesized into thematic sections: BERT and PLM evolution, Transformer-based architectures, LLMs in IR, Retrieval-Augmented Generation, and Generative IR. Additionally, we synthesized information on domain applications, Comparative frameworks, and Challenges & opportunities.

3.6. PRISMA-Inspired Flow Diagram

In this review, a total of 527 records were initially identified, after which 74 duplicates were removed, the remaining 453 records underwent title and abstract screening, leading to the exclusion of 170 records. Subsequently, 83 full-text articles were assessed for eligibility, of which 43 were excluded, resulting in 40 studies being included in the final synthesis.

4. Applications of IR and Generative Models across Domains

Information Retrieval (IR) and generative models have seen widespread adoption across diverse domains, enabling advanced search, question answering, and knowledge synthesis. The integration of pre-trained language models (PLMs) and retrieval-augmented generation (RAG) frameworks has extended these applications to knowledge-intensive and contextually complex tasks [1,5,10,13,22].

4.1. Healthcare and Biomedical IR

IR systems are widely used in healthcare to retrieve medical literature, patient records, and clinical trial information. BERT-based models and domain-specific variants such as BioBERT enhance semantic search in biomedical corpora, improving the retrieval of contextually relevant information [13, 32]. Generative models support automated summarization, clinical note generation, and decision support systems, providing evidence-grounded outputs for practitioners and patients [5, 10, 13]. RAG frameworks further ensure that generated responses are grounded in reliable sources, improving factual accuracy [6,14].

4.2. Legal and Regulatory Domains

In the legal field, IR systems facilitate case law retrieval, contract analysis, and regulatory compliance [6,8,13]. Generative IR frameworks help summarize statutes, generate reasoning chains, and produce concise legal summaries, enhancing efficiency in legal workflows [5,9,22,33]. Knowledge graphs are increasingly integrated to support fact verification and reduce hallucinations in generated content.

4.3. Scientific Literature and Academic Research

IR systems assist researchers in navigating scientific literature, extracting key insights, and performing literature reviews. BERT embeddings and PLMs enable precise semantic search and passage retrieval across large academic databases [1, 12]. Generative models can synthesize findings from multiple sources, summarize research papers, and generate knowledge-grounded literature reviews [10,16,27]. RAG approaches allow the integration of knowledge from diverse publications, supporting hypothesis generation and multi-source synthesis [6,13].

4.4. Conversational AI and Question Answering

Generative IR models, particularly RAG-based frameworks, are deployed in chatbots, virtual assistants, and customer support systems to handle open-domain queries and multi-turn dialogue [5, 9, 15, 29, 34]. By combining retrieval with generation, these systems produce contextually accurate and knowledge-grounded responses. Multimodal IR further integrates text, audio, and visual data to enhance user interactions and support richer conversational experiences [6,27,29].

4.5. Cross-Domain and Multimodal Applications

Modern IR systems increasingly combine retrieval across multiple domains and modalities. Hybrid retrieval models integrate sparse keyword-based methods with dense semantic embeddings, while multimodal frameworks allow simultaneous retrieval of text, images, audio, and video content [29, 34]. Applications include multimedia search engines, cross-modal question answering, and domain-specific knowledge integration. Multimodal RAG frameworks enable tasks such as generating text explanations from images or retrieving visual content from text queries [9,22,29].

5. BERT-Based and Pre-Trained Language Model (PLM) Approaches

The introduction of BERT and other transformer-based pre-trained language models (PLMs) represents a major milestone in Information Retrieval (IR) for NLP. By leveraging contextual embeddings, these models capture semantic nuances beyond traditional keyword matching, enabling systems to interpret user queries and understand user intent more effectively [1,2,4]. These model authors [1,2,4] provided a comprehensive survey of BERT-based

IR models, highlighting their applications in passage ranking, document retrieval, and question answering (QA). The Key applications of these models include ranking candidate passages by semantic similarity, retrieving relevant documents from large corpora, and extracting precise answers from unstructured text [2]. Fine-tuning BERT on domain-specific datasets further enhances retrieval performance by aligning the model with specialized vocabulary and domain knowledge [1,3,4]. **Figure 4** illustrates the integration of BERT and other PLMs in IR systems.

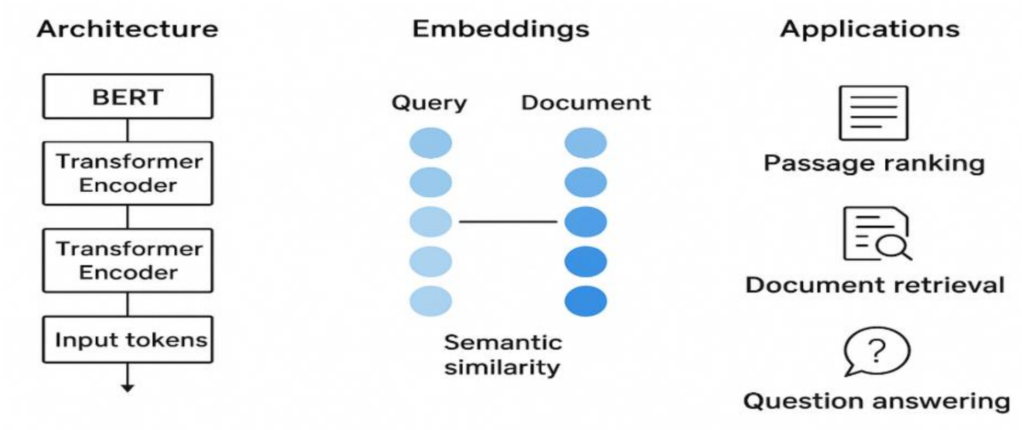

Figure 4. Integration between BERT and other PLMs in IR systems.

5.1. Architectures of BERT, RoBERTa, and ALBERT

BERT (Bidirectional Encoder Representations from Transformers) fundamentally advanced contextual understanding in NLP by processing text bidirectionally, considering both preceding and succeeding words in a sequence [1,4]. Its architecture is based on a Transformer encoder stack and two pre-training objectives:

- Masked Language Modeling (MLM): predicts masked tokens using contextual information.
- Next Sentence Prediction (NSP): models relationships between sentence pairs, supporting QA and natural language inference tasks [1,4].

Subsequent variants have been developed to address efficiency, scalability, and robustness:

- RoBERTa (Robustly Optimized BERT): removes NSP, optimizes hyperparameters, and leverages larger datasets for improved benchmark performance [1,2,4,18].
- ALBERT (A Lite BERT): reduces memory consumption via parameter sharing and factorized embeddings while maintaining performance comparable to BERT [1,2,4].

BERT and its derivatives enhance semantic search by generating dense vector embeddings for queries and documents, enabling retrieval based on semantic similarity rather than literal keyword overlap [1,2,4]. In extractive QA, these models identify the start and end positions of answers within documents, achieving state-of-the-art results on datasets such as SQuAD. A key strength of these models lies in capturing rich contextual semantics, disambiguating word meanings, and encoding dense embeddings that improve ranking and relevance in IR tasks.

Despite their strengths, BERT and related models present significant challenges:

- High computational and memory requirements, which hinder scalability for large-scale retrieval systems [1,2,4].
- Black-box nature limits interpretability, especially in high-stakes domains such as legal or medical IR [1,2,4].

5.2. Neural IR and Transformer-Based Retrieval

The evolution of dense retrieval models, such as Dense Passage Retrieval (DPR), has further advanced IR by encoding queries and documents into a shared vector space using dual encoders, enabling efficient semantic matching at scale [9,26]. Standardized evaluation benchmarks like BEIR support zero-shot testing and cross-domain generalization, facilitating more rigorous assessment of retrieval models [2,31].

Overall, the transition from traditional keyword-based approaches to deep contextual embeddings and transformer architectures forms the foundation for modern IR systems and Retrieval-Augmented Generation (RAG) frameworks, bridging BERT-based retrieval with generative LLMs [1, 3, 4, 10, 13, 22]. **Figure 5** illustrates the integration of BERT and PLMs in IR, and **Table 3** summarizes key aspects of BERT-based IR approaches.

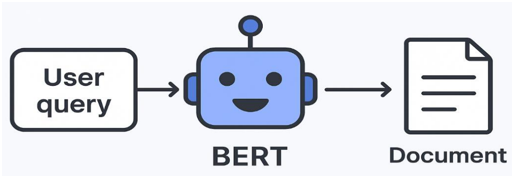

Figure 5. Integration between BERT and PLMs in IR.

Table 3. Summary of information retrieval with BERT and Pre-Trained Language Models.

Category	Details	Refs.
Models/Architectures	BERT: Bidirectional Transformer encoder with MLM and NSP for contextual semantics. RoBERTa: optimized BERT variant, removes NSP, larger datasets, fine-tuned hyperparameters. ALBERT: parameter-sharing and factorized embeddings to reduce memory usage while maintaining performance.	Wang et al., Gardazi et al., Salem et al. and Zhu et al. [1–4]
Applications in IR	Semantic Search: dense embeddings enable semantic similarity retrieval. QA: extractive QA identifies precise answer spans in documents (e.g., SQuAD). Document & Passage Retrieval: ranking passages and retrieving relevant documents. Domain-Specific Fine-Tuning: enhances performance for specialized vocabularies and domain knowledge.	Wang et al., Gardazi et al. and Zhu et al. [1,2,4]
Strengths	- Rich contextual understanding - Disambiguation of word meanings- Dense embeddings improve ranking and relevance-Effective transfer learning via pre-training on large corpora	Wang et al., Gardazi et al. and Zhu et al. [1,2,4]
Limitations	- Computationally intensive (requires GPUs/TPUs)- High memory consumption (especially BERT)- Black-box nature limits interpretability in critical domains (e.g., legal, medical IR)	Wang et al., Gardazi et al. and Zhu et al. [1,2,4]
Neural IR/Transformer-Based Retrieval	DPR: dual-encoder models encode queries and documents into shared vector space for semantic matching. Benchmarks like BEIR support zero-shot and cross-domain evaluation. Forms foundation for RAG frameworks integrating BERT-based retrieval with generative LLMs.	Wang et al., Fan et al., Garcia-Carmona et al., Wu et al., Alvarado-Maldonado and Wang. [1,5,6,10,11,21]

6. Generative Information Retrieval Approaches

Generative IR represents the next phase in the evolution of retrieval systems, combining traditional retrieval mechanisms with Large Language Models (LLMs) to generate contextually coherent and knowledge-grounded outputs. This integration allows IR systems not only to retrieve relevant information but also to produce synthesized responses that are fluent, accurate, and tailored to specific user queries [10,15,16,26]. The shift from Dense Retrieval to Generative IR, began with models like Dense Passage Retrieval (DPR) encoded both queries and passages into dense vector representations, enabling efficient semantic search over large-scale corpora [16,25]. Building upon DPR, Retrieval-Augmented Generation (RAG) frameworks integrated retrieved passages with generative models, producing responses that are informative and contextually grounded [5, 10]. This retrieval-augmented approach mitigates hallucinations commonly observed in standalone generative models and enhances factual correctness, particularly in open-domain question answering scenarios [28]. Several recent surveys have systematically reviewed generative IR and its applications. For example, Kuo et al. [16] provided an extensive overview of model architectures, applications, and emerging challenges. Wu et al. [10] focused on RAG in NLP, analyzing knowledge integration, evaluation metrics, and efficiency considerations. Meanwhile, Fan et al. [5] explored hybrid RAG and LLM architectures, emphasizing knowledge-grounded retrieval for large-scale applications. Collectively, these surveys illustrate a clear shift from purely BERT-based retrieval systems toward hybrid architectures that combine dense retrieval with generative capabilities. Recent research has extended generative IR through specialized variants and innovative integrations:

- Knowledge-Oriented RAG: integrates domain-specific knowledge into RAG models, enhancing reasoning and factual accuracy [28].

- Multimodal RAG: extends retrieval-augmented generation to multiple modalities, including text, images, and audio [29].
- Graph-Based RAG (GRAG): incorporates structured knowledge via graphs for improved handling of complex queries [7,12,31].
- Long-Document Retrieval: addresses retrieval and processing of long documents, preserving context and efficiency in LLM and PLM frameworks [35].
- LLM-Based Embeddings for Specialized Retrieval: leverages embeddings from LLMs for tasks such as legal case retrieval, ensuring grounded and precise outputs [9].

These innovations demonstrate the growing sophistication of generative IR systems, highlighting their capacity to integrate structured knowledge, handle diverse data modalities, and leverage LLMs for complex information retrieval tasks. RAG enhances factual accuracy by grounding LLM outputs in retrieved documents. In this process, the retriever identifies relevant passages, and the LLM generator synthesizes them into coherent, knowledge-grounded responses [10,25,29]. Hybrid Retrieval Models (Sparse + Dense) combine dense semantic retrieval with sparse keyword-based retrieval (e.g., TF-IDF, BM25), often employing ranking fusion techniques to improve recall and precision [25]. Multimodal IR (Text + Audio + Image + Video) enables cross-modal queries and retrieval, with joint embeddings allowing semantic understanding across multiple data types [28,33]. Integration with Knowledge Graphs (KGs) further improves retrieval by supporting query expansion, re-ranking, and fact verification, mitigating hallucinations in LLM-based systems [28,33]. **Table 4** summarized generative information retrieval.

Table 4. Summary of generative IR variants and approaches.

Generative IR Variant	Key Features	Applications/Use Cases	References
Dense Passage Retrieval (DPR)	Dense embeddings of queries and passages, efficient semantic search	Large-scale semantic search, QA, passage retrieval	Kuo et al., Genesis and Keane [16,25]
Retrieval-Augmented Generation (RAG)	Combines retrieval with generative LLMs, contextually grounded responses	Open-domain QA, knowledge synthesis, conversational AI	Fan et al., Wu et al. and Cheng et al. [5,10,28]
Knowledge-Oriented RAG	Integrates domain-specific knowledge	Domain-specific QA, reasoning, factual accuracy	Cheng et al. [28]
Multimodal RAG	Handles multiple modalities: text, images, audio	Cross-modal retrieval, multimodal QA, multimedia synthesis	Wu et al., Genesis et al. and Mei et al. [10,25,29]
Graph-Based RAG (GRAG)	Incorporates structured knowledge graphs	Complex query handling, structured knowledge reasoning	Huang et al., Zhang et al. and Zhu et al. [7,12,31]
Long-Document Retrieval	Maintains context over long passages/documents	Legal, scientific, technical document retrieval	Li et al. [35]
LLM-Based Embeddings for Specialized Retrieval	Embeddings from LLMs for precise, grounded retrieval	Legal case retrieval, domain-specific IR	Premasiri et al. [9]
Hybrid Retrieval (Sparse + Dense)	Combines dense semantic retrieval with sparse keyword-based retrieval	Improved recall and precision across domains	Wu et al., Genesis et al. and Mei et al. [10,25,29]
Multimodal IR (Text + Audio + Image + Video)	Joint embeddings for cross-modal semantic understanding	Multimedia QA, video retrieval, audio-visual search	Wu et al., Genesis et al. and Mei et al. [10,25,29]
Knowledge Graph Integration	Query expansion, re-ranking, fact verification	Factually consistent outputs, mitigates LLM hallucination	Cheng et al. and Wu et al. [28,33]

7. Transition to Generative Large Language Models (LLMs)

The field of NLP and IR has experienced a major transformation with the advent of Generative LLMs. Models such as GPT, Large Language Model Meta AI (LLaMA), and Pathways Language Model (PaLM) have expanded the capabilities of language technologies beyond representation learning, enabling complex reasoning, knowledge synthesis, and fluent, context-aware text generation. This evolution has shifted IR from merely retrieving relevant documents to generating coherent responses that leverage retrieved knowledge effectively [1,5,10,13,15,16].

7.1. Emergence of GPT, LLaMA, PaLM, and Other LLMs

The foundation of LLMs lies in the Transformer architecture, which underpins models like OpenAI's GPT series. Beginning with GPT-1 and progressing to GPT-3 and GPT-4, these decoder-only Transformers demonstrated remarkable capabilities in generating human-like text and performing diverse language tasks with minimal supervision [1,4]. Trained on massive text corpora, they predict the next token in a sequence, enabling fluent and coherent outputs. Following GPT's success, other LLMs emerged from leading research organizations [13,20]. Google's PaLM and Meta AI's LLaMA series illustrate alternative approaches, often differing in scale, architecture, or training methodology. These models typically consist of hundreds of billions or even trillions of parameters and are capable of zero-shot or few-shot learning, allowing generalization to unseen tasks with minimal task-specific data [7,8].

7.2. The Shift in Retrieval Concept: From Matching to Generation

LLMs have redefined the concept of information retrieval. Classical IR focused on query-to-document matching, while BERT-based models introduced semantic matching capabilities [1,2]. In contrast, LLMs enable RAG, synthesizing information from multiple sources to produce direct, contextually coherent answers. This paradigm is particularly valuable for complex or open-ended queries that require the aggregation and summarization of information across diverse sources. Instead of returning ranked document lists, RAG models provide concise, generated answers, often citing sources. Such capabilities enhance the user experience by enabling intelligent and informative IR systems [16,28,29].

7.3. Generative Models in Conversational Search

Conversational search benefits significantly from generative LLMs. Traditional search engines are optimized for single-turn queries, whereas human information-seeking is often multi-turn and context-dependent. Generative models maintain context across turns, comprehend follow-up questions, and produce coherent conversational responses [5,25].

In this context, LLMs serve as intelligent agents, retrieving relevant information and generating responses aligned with user intent. This facilitates interactive search experiences, allowing users to refine queries, ask clarifying questions, and explore topics through natural dialogue. For example, an LLM can provide summaries of historical events and answer subsequent questions about dates, figures, or implications within the same conversational thread [6].

7.4. Comparison with BERT

When comparing LLMs to BERT-based models, a trade-off between performance and computational requirements becomes evident. LLMs excel at tasks demanding deep semantic understanding, reasoning, and generative output, producing human-like, contextually informed responses [1,5,10,28]. Conversely, BERT models are primarily designed for representation learning and classification tasks, performing less effectively in generative applications [1,2]. However, the enhanced performance of LLMs comes at a substantial computational cost, requiring high-end GPUs, large memory capacities, and potentially higher inference latency. BERT-based models, while less capable generatively, are computationally efficient and more suitable for traditional IR tasks, offering a practical balance between resources and performance [1,4].

Figure 6 and **Table 5** summarize the transition from traditional and BERT-based IR to generative LLMs.

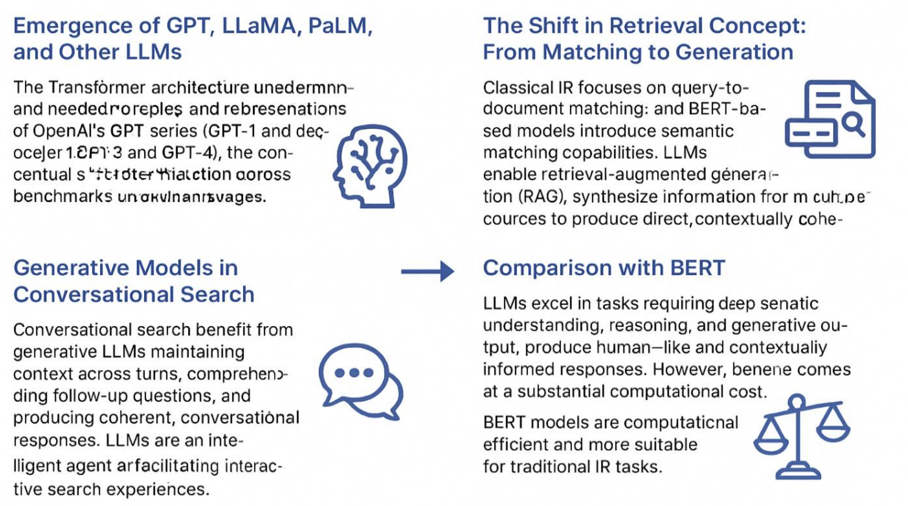

Figure 6. Transition from traditional and BERT-based IR to generative LLMs.

Table 5. Summary of generative LLM-based IR.

Category	Key Features	Applications/Use Cases	References
GPT Series	Decoder-only Transformers, next-token prediction, fluent text generation	Open-domain QA, content generation, reasoning	Wang et al., Zhu et al., Kuo et al. and Subi et al. [1,4,16,30]
PaLM	Transformer-based LLM, large-scale pre-training, few-shot learning	Cross-domain reasoning, multilingual tasks	Wang et al., Zhu et al., Kuo et al. and Subi et al. [1,4,16,30]
LLaMA	Transformer LLM, efficient scaling, optimized architecture	Research applications, specialized IR tasks	Wang et al., Zhu et al., Kuo et al. and Subi et al. [1,4,16,30]
RAG (Generative IR)	Retrieval-augmented generation, synthesizes info from multiple sources	Knowledge-grounded QA, complex queries	Fan et al. and Wu et al. [5,10]
Conversational LLMs	Maintains multi-turn context, generates coherent responses	Conversational search, virtual assistants, interactive QA	Fan et al., Garcia-Carmona et al. and Genesis et al. [5,6,25]
Comparison with BERT	LLMs: generative, reasoning, high semantic understanding. BERT: representation learning, classification	Trade-off between performance and computational cost, task-specific deployment	Wang et al., Gardazi et al., Fan et al., Wu et al. and Cheng et al. [1,2,5,10,28]

8. Generative Large Language Models in Information Retrieval

The rise of generative large language models (LLMs) has significantly transformed the landscape of Information Retrieval (IR). Models such as GPT, LLaMA, and PaLM enable complex reasoning, knowledge synthesis, and fluent text generation, shifting IR from simple document retrieval to retrieval-augmented generation of coherent responses [4,7,8,36].

8.1. Emergence of GPT, LLaMA, PaLM, and Other LLMs

LLMs are built upon the Transformer architecture. OpenAI's GPT series, beginning with GPT-1 and progressing to GPT-3 and GPT-4, demonstrated remarkable generative capabilities and few-shot learning, allowing the models to generalize across diverse language tasks [4,7,8]. Similarly, Google's PaLM and Meta AI's LLaMA introduced architectural and training variations, supporting hundreds of billions of parameters and enabling effective cross-domain generalization [4,7,8].

8.2. Shift in Retrieval: From Matching to Generation

Traditional IR and BERT-based models primarily focus on semantic matching between queries and documents [1,2]. In contrast, LLMs enable a paradigm shift where retrieval is augmented by generative synthesis. Instead of returning ranked lists of documents, retrieval-augmented generation (RAG) models integrate retrieved passages with generative capabilities to produce direct, coherent, and source-grounded answers. This approach is particularly effective for complex or open-ended queries, enhancing both accuracy and user experience [5,10,36].

8.3. Generative Models in Conversational Search

Conversational search benefits substantially from LLMs, which can maintain multi-turn context, interpret follow-up queries, and generate coherent natural language responses [5,16]. These models serve as intelligent agents, allowing users to refine queries, ask clarifying questions, and explore topics through interactive dialogue. Generative IR thus provides more dynamic, context-aware, and human-like search interactions.

8.4. Comparison with BERT

LLMs outperform BERT-based models in tasks that require reasoning, information synthesis, and fluent text generation. However, this advantage comes with higher computational costs, including increased memory requirements and longer inference times. BERT-based models remain efficient and practical for traditional retrieval tasks, whereas LLMs are preferred when advanced generative capabilities and conversational intelligence are needed [4,8]. **Table 6** shows the comparison among the BERT and Generative in the term of IR.

Figure 7 illustrates the advantages and challenges of BERT-based models versus generative LLMs in modern IR systems.

This comparison highlights the trade-offs between computational efficiency and generative power, guiding the choice of models based on application requirements.

Table 6. Comparison of BERT-Based Models vs. Generative LLMs in IR.

Feature/Aspect	BERT-Based Models	Generative LLMs (GPT, LLaMA, PaLM)
Architecture	Transformer encoder	Transformer decoder/encoder-decoder
Semantic Understanding	Strong contextual embeddings, effective for ranking QA	Advanced reasoning, synthesis, and generative capabilities
Output Type	Ranked documents, extracted answers	Synthesized, coherent, source-grounded responses
Computational Cost	Moderate (GPU/TPU required)	High (large memory, multiple GPUs/TPUs)
Inference Latency	Low to moderate	Higher, may affect real-time applications
Interpretability	Limited, but better than LLMs	Black-box, less interpretable
Ideal Use Case	Traditional IR, passage/document retrieval, extractive QA	Conversational search, open-domain QA, multi-turn dialogue, knowledge synthesis

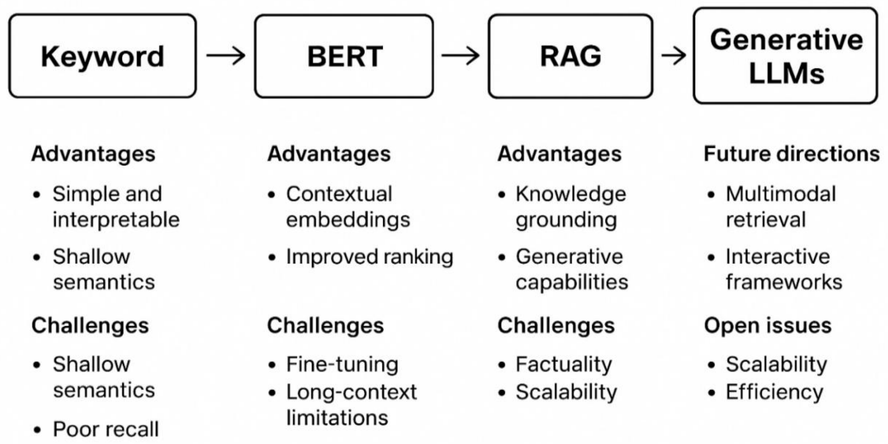

Figure 7. Advantages and challenges of BERT-based models versus generative LLMs in modern IR systems.

9. Evaluation Frameworks for IR and RAG

Evaluation plays a pivotal role in Information Retrieval (IR), as it determines the effectiveness of retrieval methods and guides the selection of appropriate models for downstream applications. With the shift from traditional keyword-based retrieval to neural architectures, and more recently to generative approaches, evaluation frameworks have evolved to meet new requirements and challenges.

9.1. Evaluation Frameworks

The evaluation of IR frameworks has progressed as follows:

- **Traditional IR Metrics:** Conventional evaluation methods in IR focused on retrieval accuracy and ranking performance. Widely adopted metrics include:
  - **Mean Average Precision (MAP):** Summarizes ranking effectiveness across queries.
  - **Normalized Discounted Cumulative Gain (NDCG):** Rewards highly relevant results at top ranks.
  - **Precision and Recall:** Measure retrieval accuracy and completeness.

These metrics are effective for classical and BERT-based retrieval systems but are less suited to generative IR outputs, where the response may be synthesized rather than a direct document match [1,4].

- **Generative Evaluation Issues:** Generative IR systems, including RAG and LLM-driven retrieval, present distinct evaluation challenges:
  - **Factual correctness:** Outputs may be fluent yet inaccurate.
  - **Diversity and completeness:** Responses vary in style and coverage.
  - **Relevance–fluency trade-off:** Coherent answers may not fully align with the retrieved knowledge.

Automated metrics like BLEU, ROUGE, and METEOR capture surface-level similarity but often fail to assess factuality or reasoning quality. Hence, human evaluation remains indispensable for complex or domain-specific scenarios [4,26].

- **Hybrid Evaluation Frameworks:** A systematic review by Gan et al. [26], emphasized hybrid frameworks combining:

  1. Traditional IR metrics (MAP, NDCG),
  2. Generative output metrics (BLEU, ROUGE), and
  3. Human-centered evaluation (factual accuracy, interpretability, usability).

  Such hybrid frameworks enable holistic assessment of retrieval effectiveness and generative quality.

- **Benchmarks for IR Evaluation:** Benchmark datasets play a central role in standardizing evaluation. BEIR is widely adopted for testing retrieval methods across 18 heterogeneous datasets from domains including science, news, biomedical, and social media [27]. It supports zero-shot evaluation, enabling rigorous comparison of classical IR, BERT-based retrieval, and LLM embeddings [24,28].
- **Emerging Hybrid Approaches: Recent** advancements combine dense retrieval metrics (MAP, NDCG, Precision, Recall) with generative metrics (BLEU, ROUGE, METEOR, BERTScore), supplemented by **factuality and grounding checks** via knowledge bases or human judgment. These approaches are particularly critical in knowledge-intensive domains such as healthcare, law, and scientific IR [10,14,26].

  Figure 8 illustrates the evolution of evaluation frameworks for IR and RAG.

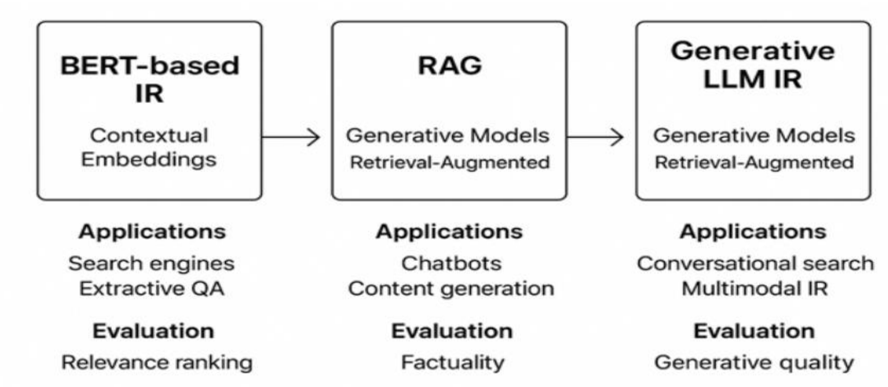

Figure 8. Evolution of evaluation frameworks for IR and RAG.

Table 7 provides comparative overview of evaluation frameworks in traditional, BERT-based, and generative IR systems.

Table 7. Comparative of evaluation frameworks in traditional, BERT-based, and generative IR systems.

Evaluation Category	Metric/Approach	Purpose	Advantages	Limitations/Challenges	Applicability
Traditional IR Metrics	MAP	Measure ranking accuracy across queries	Well-established, interpretable, captures overall ranking	Does not handle generative outputs	Classical IR, BERT-based retrieval
	NDCG	Evaluate ranking quality emphasizing top results	Rewards highly relevant top results	Limited for outputs beyond exact document matches	Classical IR, BERT-based retrieval
	Precision & Recall	Measure fraction of relevant docs retrieved & completeness	Simple, widely understood	Insufficient for generative responses	Classical IR, BERT-based retrieval
Generative Output Metrics	BLEU, ROUGE, METEOR	Evaluate n-gram overlap with reference outputs	Standardized, automatic, fast	Cannot fully assess factuality, reasoning, or relevance-fluency trade-off	LLM-based IR, RAG systems
	BERTScore	Semantic similarity between generated and reference text	Captures semantic alignment	Requires reference texts, may miss factual errors	LLM-based IR, RAG systems
Factuality & Grounding	Knowledge-base or human evaluation	Assess correctness and factual grounding	Accurate for domain-specific and complex queries	Time-consuming, expensive, less scalable	RAG, domain-specific LLM retrieval
Hybrid Frameworks	Combined metrics (MAP/NDCG + BLEU/ROUGE + human evaluation)	Holistic assessment of retrieval + generative quality	Balances retrieval accuracy, output quality, and factuality	Complexity of integration, evaluation cost	BERT + RAG + LLM systems
Benchmark Datasets	BEIR	Standardized evaluation across multiple domains	Enables zero-shot and cross-domain comparison	Limited generative evaluation metrics	Classical IR, BERT-based IR, LLM embeddings

9.2. Evaluation Based on Applications of IR with LLMs

The integration of LLMs into IR has enabled significant advancements across multiple domains, enhancing both precision and contextual relevance [21,23,37,38]. The Key application areas include:

- **Conversational IR with ChatGPT:** Dialogue-based LLMs reformulate queries, retrieve contextually relevant passages, and generate concise summaries, representing a shift toward interactive, user-centric retrieval [7,39,40].
- **Structured Information Retrieval:** LLMs map unstructured text into structured formats (tables, schemas), improving accuracy over classical symbolic methods, especially in biomedical literature and EHRs [6,14].
- **Knowledge Base Integration:** LLMs combined with knowledge graphs reduce hallucinations, enhance factual correctness, and support domain-specific reasoning [8,19].
- **Biomedical and Scientific Retrieval:** LLMs extract and link information across scientific publications to support hypothesis generation, literature review, and clinical QA [6].
- **Legal and Case Law Retrieval:** LLM embedders efficiently retrieve prior cases and legal references, with RAG architectures grounding outputs in verified legal documents [9].

  Tables 8 and **9** illustrate the evaluation framework Based on Applications of IR with LLMs, and Combined Evaluation Frameworks and Applications of IR + LLMs, respectively.

Table 8. Evaluation framework based on applications of IR with LLMs.

Application Area	Description/Function	Refs.
Conversational IR	Dialogue-driven query reformulation and answer generation	Y. Huang and J. X. Huang [7]
Structured Information Extraction	Map unstructured text to structured formats (tables, schemas)	Garcia-Carmona et al. [6]
Knowledge Graph Integration	Ground generative outputs in knowledge graphs for factual accuracy	Yang et al. [8]
Biomedical/Scientific Retrieval	Extract and link information for literature review, hypothesis generation, clinical QA	Garcia-Carmona et al. [6]
Legal/Case Law Retrieval	Retrieve and summarize precedent cases, support legal analysis	Premasiri et al. [9]

Table 9. Combined evaluation frameworks and applications of IR + LLMs.

Dimension	Category	Key Methods/Metrics	Applications & Use Cases	Refs.
Evaluation Frameworks	Traditional IR Metrics	MAP, NDCG, Precision, Recall	Ranking quality, retrieval accuracy	Wang et al. and Zhu et al [1,4]
	Generative Output Metrics	BLEU, ROUGE, METEOR, BERTScore	Assess fluency, semantic similarity	Wu et al. and Gan et al. [10,26]
	Factuality & Grounding	Human evaluation, knowledge-based checks	Domain-specific correctness (healthcare, law, science)	Wu et al., Alvarado-Maldonado et al. and Gan et al. [10,11,26]
	Benchmarks	BEIR (18 heterogeneous datasets, zero-shot IR)	Cross-domain evaluation	Cheng et al. [28]
	Hybrid Frameworks	Combination of retrieval + generative metrics + human evaluation	Holistic system-level assessment	Wu et al. and Gan et al. [10,26]
Applications of IR + LLMs	Conversational IR	Dialogue-driven retrieval & answer generation	Chat-based assistants, query reformulation	Huang et al. [7]
	Structured Information Retrieval	Embedding-based retrieval, hybrid indexing	Biomedical literature, EHRs, finance	Garcia-Carmona et al. [6]
	Knowledge Graph Integration	Grounding with symbolic knowledge	Reasoning, domain-specific QA	Yang et al. [8]
	Biomedical & Scientific IR	Domain-specific embeddings + RAG	Hypothesis generation, literature review, clinical QA	Garcia-Carmona et al. [6]
	Legal/Case Law Retrieval	LLM embedders + RAG grounding	Precedent retrieval, legal summarization	Premasiriet al. [9]

This structured overview emphasizes how evaluation frameworks have evolved from traditional IR metrics to hybrid strategies, addressing the needs of generative LLM-enhanced IR across multiple domains.

10. Comparative Analysis

This section offers a structured perspective on the evolution of Information Retrieval (IR), by comparing traditional, BERT-based, and generative IR approaches across multiple dimensions, including semantic capability, scalability, computational requirements, and suitability for various applications. The comparison emphasizes the trade-offs between accuracy, efficiency, and versatility in different retrieval scenarios. The progression of IR in NLP reflects a clear shift from contextually focused models (such as BERT and its variants, to generative Large Language Models (LLMs) that integrate retrieval mechanisms.) Although both paradigms have significantly advanced IR, they differ in design principles, operational mechanisms, and performance trade-offs. This section presents a systematic comparison of BERT-based and LLM-based IR approaches, considering criteria such as accuracy, efficiency, scalability, interpretability, and practical applicability [1,2,4–10,12,16,19,22,25,28,29,33,35].

10.1. Traditional IR Models

Traditional IR models such as TF-IDF, BM25, Boolean retrieval, and VSM are computationally efficient and interpretable. They perform well for keyword-based search tasks but are limited in capturing semantics, context, and user intent [1,2]. These methods are best suited for applications where exact matching is sufficient, or where computational resources are constrained.

10.2. Comparing BERT-Based IR and LLM-Based IR

- **Accuracy and Effectiveness**
  - *BERT-based models* transform IR through bidirectional contextual embeddings, capturing semantic relationships between queries and documents. They excel in semantic search and extractive question answering, achieving high precision [1,2].
  - *LLM-based IR*, particularly via Retrieval-Augmented Generation (RAG), goes beyond retrieval by generating coherent, knowledge-grounded responses. It synthesizes information from multiple sources, providing comprehensive answers to complex queries [5,10,28,29]. While hallucinations remain a challenge, RAG mitigates this by grounding outputs in retrieved evidence.
- **Efficiency and Computational Cost**
  - BERT models are relatively efficient with smaller parameter sizes, enabling faster inference and lower memory usage, suitable for real-time deployment [1,2].
  - LLMs often encompass billions of parameters, requiring high-end GPUs or distributed systems. They offer advanced capabilities at a higher computational cost, with latency considerations for real-time applications [5,10,16].
- **Scalability**
  - BERT-based systems scale effectively with optimizations such as approximate nearest neighbor search for dense embeddings [2,22].
  - LLM-based systems face additional challenges in maintaining large knowledge bases and generating responses at scale. Research in distributed computing, model quantization, and efficient indexing is ongoing [9,28,29,33].
- **Interpretability**
  - BERT-based IR has limited interpretability, yet ranking decisions are somewhat explainable through similarity-based matching [1,2].
  - LLMs pose higher interpretability challenges due to potential hallucinations. Explainable AI (XAI) techniques are critical, including attention visualization, influential feature identification, and human-readable justifications for generation outcomes [5,10].
- **Practical Applications**
  - BERT-based IR: Search engines, recommendation systems, extractive QA [1,2].
  - LLM-based IR: Interactive chatbots, content generation, conversational search, multimodal information systems [6,7,10,28,29].
- **Hallucination in Generative Models**
  LLMs may generate plausible but factually incorrect outputs. RAG mitigates this by grounding responses in retrieved documents, but hallucinations persist if retrieval is insufficient or contradictory. Research on self-correction, fact-checking, and adversarial training aims to improve factuality [28,33].
- **Bias and Ethical Concerns**
  Both BERT and LLMs inherit societal biases from training data, potentially affecting fairness, privacy, and security. LLMs amplify ethical concerns due to their ability to generate convincing, large-scale outputs. Potential solutions include bias detection, ethical guidelines, transparency, and interdisciplinary collaboration.
  In summary, the BERT-based models are suitable for domain-specific and context-aware retrieval tasks, although they require substantial computational resources and careful tuning.

10.3. Need for More Efficient and Context-Aware Models

Challenges in scalability, interpretability, and hallucination highlight the need for efficient, context-aware models. Future IR systems should:

- Dynamically focus on salient information in long or multi-document contexts.
- Leverage user intent, conversational history, and domain-specific knowledge.
- Combine symbolic reasoning with neural approaches for more robust understanding.

Table 10 provides a comparative analysis of BERT-Based and Generative LLM-Based IR Approaches

Table 10. Comparative analysis of BERT-based and generative LLM-based IR approaches.

Feature	BERT-Based IR	LLM-Based IR (RAG)
Primary Function	Semantic Matching, Ranking, Extraction	Generative Synthesis, Conversational Answers
Core Mechanism	Contextual Embeddings, Bidirectional Encoding	Generative Models, Retrieval-Augmented Generation
Accuracy	High for relevance ranking and extraction	High for synthesized answers, RAG mitigates hallucination
Efficiency	More efficient, faster inference	Less efficient, high computational cost
Scalability	Optimized for large corpora	Challenging, KB maintenance adds complexity
Interpretability	Limited but more explainable	Limited, XAI crucial to manage hallucinations
Applications	Search engines, extractive QA, recommendations	Chatbots, content generation, conversational search

Table 11 summarizes the evolution of IR in NLP: From BERT-Based Models to RAG and Generative LLMs.

Table 11. Evolution of IR in NLP: From BERT-based Models to RAG and generative LLMs.

Refs.	Problem	Techniques Used	Contributions	Measurements	Conclusion
Wang et al. [1]	BERT-based IR	Transformer embeddings, fine-tuning	Survey of BERT IR applications	MAP, NDCG	High-quality semantic retrieval, domain adaptation necessary
Gardazi et al. [2]	BERT in NLP	Pretrained BERT models	Overview of BERT IR use-cases	Literature synthesis	BERT foundational for neural IR
Kuo et al. [16]	Generative IR	LLMs, RAG	Survey of generative IR trends	BLEU, ROUGE	Integrates retrieval and LLM generation effectively
Wu et al. [10]	RAG in NLP	DPR + LLM	Comprehensive survey on RAG	MAP, BLEU, ROUGE	Improves factuality and QA accuracy
Cheng et al. [28]	Knowledge-oriented RAG	Knowledge integration, RAG	Domain-specific generative IR	BLEU, human evaluation	Knowledge grounding enhances accuracy
Mei et al. and Nawaz. [29,34]	Multimodal RAG	Text + Image embeddings	Survey of multimodal RAG	Human eval	Enables multimodal retrieval and generation
Wu et al. [33]	Graph-based retrieval	Graph embeddings + RAG	Structured reasoning in IR	MAP, human eval	Supports complex queries and reasoning
Yang et al. [9]	Legal retrieval	LLM embeddings, RAG	Efficient retrieval of prior cases	MAP, NDCG	LLM embeddings improve retrieval accuracy
Fan et al. and Hu et al. [5,36]	RAG + LLM	DPR + generator	Survey of retrieval-augmented LLMs	MAP, BLEU	Combines retrieval and generation effectively
Zhang et al. [12]	Evaluation frameworks	Retrieval + generative metrics	Survey on hybrid evaluation	MAP, NDCG, BLEU, human eval	Hybrid metrics assess both retrieval and generative quality

This comparative analysis highlights the trade-offs between performance, interpretability, and computational demands across the evolution of IR systems. Traditional IR remains effective in resource-constrained settings, whereas BERT-based and dense retrieval methods enhance semantic understanding. Generative LLMs, in turn, facilitate advanced reasoning, conversational interaction, and multimodal retrieval. This section systematically contrasts BERT-based IR with LLM-based IR, emphasizing their respective strengths, limitations, and emerging applications in generative retrieval.

11. Discussion

The progression of Information Retrieval (IR) in NLP has followed a clear trajectory: from traditional keyword-based approaches, to neural IR using contextual embeddings, and more recently to generative retrieval with LLMs. Transformer-based models, particularly BERT, have been instrumental in this transition by encoding rich contextual information and improving ranking performance across various tasks [1,2]. However, BERT-based approaches have inherent limitations: a heavy reliance on task-specific fine-tuning, difficulty in maintaining long-range contextual dependencies, and challenges in domain adaptation and large-scale deployment [2,22]. The advent of generative IR, especially through RAG (Retrieval-Augmented Generation), has addressed some of these challenges by integrating dense retrieval mechanisms with generative capabilities. This allows models to produce fluent, knowledge-grounded responses while retaining retrieval relevance [5,10]. Extensions of RAG frameworks have further incorporated knowledge-oriented, multimodal, and graph-based architectures, enabling reasoning over both structured and unstructured information sources [13,28,29]. Despite these advancements, several open issues persist:

- Scaling LLMs efficiently to handle extensive corpora.
- Ensuring factual consistency in generated outputs.
- Supporting multimodal or long-document retrieval effectively [9,12].

Hybrid evaluation strategies have emerged as practical approaches to assess both retrieval accuracy and generative quality. These combine traditional IR metrics (e.g., MAP, NDCG) with generative evaluation metrics (e.g., BLEU, ROUGE) and human-centered assessments [12]. Key Insights from Comparative Analysis.

1. **Evolution of Techniques:**
    - NLP IR has progressed from keyword-based retrieval → BERT contextual embeddings → LLM-based generative IR.
    - Each step addresses prior limitations:
        - BERT mitigates shallow semantic representations.
        - RAG introduces knowledge-grounded reasoning**.**
        - LLM-based IR supports interactive, multimodal, and structured retrieval tasks [1,5,10,28,36].
2. **Evaluation Challenges:**
    - Conventional IR metrics are sufficient for evaluating ranking.
    - Generative IR requires hybrid evaluation frameworks that capture semantic correctness, factual grounding, and end-user satisfaction [12].
3. **Domain Adaptation and Applications:**
    - Legal, biomedical, and scientific applications necessitate domain-specific embeddings, retrieval augmentation, and fine-tuning to maintain relevance [8,9].
4. **Scalability and Efficiency:**
    - LLM-augmented pipelines improve retrieval performance but incur high computational costs.
    - Techniques such as model distillation, approximate nearest neighbor search, and lightweight retrieval modules are crucial for deployment [4,5,36].
5. **Future Directions:**
    - The integration of knowledge graphs, interactive conversational frameworks, and multimodal retrieval is expected.
    - Future IR systems should balance accuracy, interpretability and user-centric design, while also supporting complex reasoning **over** diverse data types [7,8,13,32].

Table 12 consolidates the main discussion points, challenges, and research directions, providing a clear roadmap for understanding current and future IR developments.

Table 12. Summary of discussion: Comparative insights and future directions.

Dimension	Key Insights	Challenges	Future Directions	References
Evolution of Techniques	Keyword-based → BERT → LLM + RAG	Shallow semantic understanding, limited reasoning	Knowledge-grounded, multimodal, interactive IR	Wang et al., Wu et al., Fan et al., Cheng et al. and Hu and Lu. [1,5,10,28,36]
Evaluation	Hybrid metrics combining MAP/NDCG + BLEU/ROUGE + human evaluation	Assessing factuality, semantic correctness	Domain-adapted, automated + human evaluation	Zhang et al. [12]

Table 12. *Cont.*

Dimension	Key Insights	Challenges	Future Directions	References
Domain Adaptation	Domain-specific embeddings improve performance	Specialized knowledge, fine-tuning requirements	Domain-adaptive LLMs, task-specific RAG	Yang et al. and Premasiri et al. [8,9]
Scalability & Efficiency	Efficient retrieval via ANN search, lightweight modules	High computational cost for LLMs	Model distillation, approximate search, distributed deployment	Zhu et al., Fan et al. and Hu et al. [4,5,36]
Generative Quality	LLMs generate fluent, context-aware responses	Hallucinations, factual inconsistencies	RAG grounding, knowledge integration, fact-checking	Fan et al., Wu et al. and heng et al. [5,10,28]
Future Directions	Integrate symbolic reasoning, conversational frameworks	Multimodal/long-document retrieval	Knowledge-graph + generative pipelines, context-aware IR	Yang et al., Premasiri et al., Qin et al. and Le [7,8,13,32]

12. Challenges, Open Issues, and Future Directions

Despite substantial progress in BERT-based and generative IR systems, several challenges remain, affecting their scalability, effectiveness, and adoption in real-world applications. At the same time, emerging solutions and research directions offer pathways to address these issues and advance the capabilities of IR systems. **Table 13** summarizes the challenges, Open Issues, Recommendations, and Future Directions of this work.

Table 13. Challenges, open issues, recommendations, and future directions.

Challenge/Issue	Description	Recommendation/Future Direction
Scalability & Efficiency	High GPU/TPU cost, latency in real-time retrieval	Lightweight models, pruning, quantization, efficient dense retrieval, model compression and distillation, sparse attention, approximate nearest neighbor search.
Domain Adaptation	Difficulty adapting models to specialized fields	Few-shot learning, domain-adaptive pretraining, knowledge injection, integration with domain-specific knowledge graphs.
Factuality & Hallucination	Fluent but incorrect outputs	Retrieval grounding, factual verification, trust calibration, robust evaluation frameworks including human assessment and grounding checks.
Long-Document Retrieval	Maintaining context across paragraphs	Sliding window, hierarchical embeddings, segment-level retrieval, context-aware processing and hierarchical reasoning
Multimodal Retrieval	Aligning text, images, structured data	Multimodal encoders, graph-based integration, cross-attention fusion, multimodal IR combining text, image, audio, and video.
Evaluation Frameworks	Metrics fail to capture semantic alignment	Hybrid evaluation (retrieval + generation), domain-specific benchmarks, multi-dimensional metrics (semantic fidelity, factual accuracy, user satisfaction).
Bias & Fairness	Risk of unethical or discriminatory outcomes	Bias detection frameworks, diverse training datasets, fairness-aware evaluation.
Interpretability	Limited transparency in decision-making	Explainable IR methods, interpretability toolkits, XAI techniques for visualization, feature importance, and human-readable justifications
Interactive and Conversational IR	Supporting multi-turn, context-dependent queries	Dialogue-driven retrieval using ChatGPT-like models, reinforcement learning for adaptive retrieval, conversational search with LLMs.
Real-World Applications	Domain-specific requirements and reliability	Domain-specific IR in healthcare, law, enterprise search, integration with knowledge graphs, structured reasoning, trust, security, and interpretability-focused designs.

In Summary, this combined perspective emphasizes that addressing technical challenges such as computational efficiency, domain adaptation, factuality, multimodal alignment, and interpretability must be paired with strategic research directions, including:

1. Multimodal IR with LLMs: leveraging embeddings across text, image, audio, and video.
2. Interactive Conversational IR: enabling adaptive, dialogue-driven information retrieval.
3. Robust and Hybrid Evaluation Frameworks: integrating traditional IR metrics, generative output metrics, and human-centered assessments.
4. Knowledge-Enhanced Retrieval: grounding outputs in knowledge graphs or structured databases.
5. Efficient Retrieval Architectures: model compression, distillation, sparse attention, and hierarchical embeddings.
6. Domain-Specific Applications: ensuring performance, trustworthiness, and usability in critical areas such as healthcare, law, and enterprise search.

By combining these challenges with the recommended directions, IR systems can evolve to be scalable, context-aware, knowledge-grounded, and user-centric, paving the way for next-generation generative and multimodal retrieval solutions.

13. Conclusions

This survey has provided a comprehensive overview of the evolution of Information Retrieval (IR) in Natural Language Processing (NLP). We have traced its progression from early keyword-based methods to neural approaches, followed by the introduction of BERT and transformer-based models, and finally the emergence of generative IR with Large Language Models (LLMs). The development of Retrieval-Augmented Generation (RAG) marks a significant paradigm shift, enabling systems that not only retrieve relevant information but also generate contextually grounded and semantically coherent responses. While BERT and other Pre-trained Language Models (PLMs) improved semantic retrieval and passage ranking through contextual embeddings and fine-tuning, RAG architectures integrate retrieval with generation. This integration supports knowledge-grounded outputs for question answering, long-document retrieval, and domain-specific applications. Recent research has extended generative IR to multimodal, graph-based, and knowledge-oriented models, highlighting its versatility. Despite these advances, challenges remain in scalability, interpretability, hallucination mitigation in generative models, evaluation, and ethical considerations such as bias and fairness. By synthesizing insights from numerous studies, this survey has provided a taxonomy of IR models, comparative analyses, a discussion of evaluation frameworks, and a roadmap for future directions. These directions include explainable IR, domain-specific LLMs, efficient few-shot and zero-shot learning, multimodal integration, and deployment on edge devices. In summary, the convergence of retrieval and generation promises to advance the development of IR systems that are accurate, interpretable but also human-centered. This progression will ultimately provide users with more intelligent and impactful ways for users to access and interact with information.

Author Contributions

M.H.A.-S.: Conceptualization, methodology, literature review, supervision, and writing original draft and final revision. M.M.A.-E.S.: Data curation, visualization, comparative analysis, writing, editing, and finalizing the technical review. M.A.-A. and G.A.A.A.-M.: Validation, technical review, and proofreading.

Funding

This work received no financial support.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Wang, J.; Gardazi, N.M.; Daud, A.; et al. Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *ACM Comput. Surv.* **2024**, *56*, 1–33.
2. Gardazi, N.M.; Daud, A.; Malik, M.K.; et al. BERT Applications in Natural Language Processing: A Review. *Artif. Intell. Rev.* **2025**, *58*, 1–49.
3. Salem, M.; Mohamed, A.; Shaalan, K. Transformer Models in Natural Language Processing: A Comprehensive Review and Prospects for Future Development. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics (AISI 2025). AISI 2025. Lecture Notes on Data Engineering and Communi-*

cations Technologies; Hassanien, A.E., Rizk, R.Y., Darwish, A., Eds.; Lecture Notes on Data Engineering and Communications Technologies; Springer: Cham, Switzerland, 2025; Vol. 238, pp. 463–472. [CrossRef]

4. Zhu, Y.; Liu, X.; Zhang, H.; et al. Large Language Models for Information Retrieval: A Survey. *ACM Trans. Inf. Syst.* **2023**, *44*, 1–54. [CrossRef]
5. Fan, W.; Chen, Z.; Yang, P.; et al. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; pp. 6491–6501.
6. Garcia-Carmona, A.M.; Prieto, M.-L.; Puertas, E.; et al. Leveraging Large Language Models for Accurate Retrieval of Patient Information From Medical Reports: Systematic Evaluation Study. *JMIR AI* **2025**, *4*. [CrossRef]
7. Huang, Y.; Huang, J.X. Exploring ChatGPT for Next-Generation Information Retrieval: Opportunities and Challenges. *Web Intelligence* **2024**, *22*, 31–44. [CrossRef]
8. Yang, W.; Some, L.; Bain, M.; et al. A Comprehensive Survey on Integrating Large Language Models with Knowledge-Based Methods. *Knowl.-Based Syst.* **2025**, *318*. [CrossRef]
9. Premasiri, D.; Ranasinghe, T.; Mitkov, R. LLM-Based Embedders for Prior Case Retrieval. In Proceedings of Recent Advances in Natural Language Processing, Varna, Bulgaria, 8–10 September 2025; pp. 980–988.
10. Wu, S.; Zhang, Y.; Chen, L.; et al. Retrieval-Augmented Generation for Natural Language Processing: A Survey. *arXiv preprint* **2025**, *arXiv.2407.13193*. [CrossRef]
11. Alvarado-Maldonado, D.; Johnston, B.; Brown, C.J. Natural Language Processing Tools for Pharmaceutical Manufacturing Information Extraction from Patents. *arXiv preprint* **2025**, *arXiv.2504.20598*. [CrossRef]
12. Zhang, Q.; Liu, M.; Zhou, Y.; et al. A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models. *arXiv preprint* **2025**, *arXiv.2501.13958*. [CrossRef]
13. Qin, L.; Zhao, X.; Wang, H.; et al. Large Language Models Meet NLP: A Survey. *FCS* **2024**, *arXiv.2405.12819*. [CrossRef]
14. Rojas-Carabali, W.; Liu, S.; Lee, J.; et al. Natural Language Processing in Medicine and Ophthalmology: A Review for the 21st-Century Clinician. *Asia-Pac. J. Ophthalmol.* **2024**, *13*. [CrossRef]
15. Zhao, S.; Yang, Y.; Wang, Z.; et al. Retrieval-Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make Your LLMs Use External Data More Wisely. *arXiv preprint* **2024**, *arXiv.2409.14924*. [CrossRef]
16. Kuo, T.-L.; Chiu, T.-W.; Lin, T.-S.; et al. A Survey of Generative Information Retrieval. *arXiv preprint* **2024**, *arXiv.2406.01197*. [CrossRef]
17. Demotte, P.; Wijegunarathna, K.; Meedeniya, D.; et al. Enhanced Sentiment Extraction Architecture for Social Media Content Analysis Using Capsule Networks. *Multimed. Tools Appl.* **2023**, *82*, 8665–8690.
18. Xu, Z.; Li, Q.; Zhang, Y.; et al. A Survey of Model Architectures in Information Retrieval. *arXiv preprint* **2025**, *arXiv.2502.14822*. [CrossRef]
19. Yoran, O.; Wolfson, T.; Ram, O.; et al. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. *arXiv preprint* **2023**, *arXiv.2310.01558*. [CrossRef]
20. Wu, H.; Li, S.; Gao, Y.; et al. Natural Language Processing in Educational Research: The Evolution of Research Topics. *Educ. Inf. Technol.* **2024**, *29*, 23271–23297.
21. Wang, X. The Application of NLP in Information Retrieval. *Appl. Comput. Eng.* **2024**, *42*, 290–297.
22. Mansour, E.; Alsaud, A.B.F.B.M. Advanced Information Retrieval Techniques in the Big Data Era: Trends, Challenges, and Applications. *Metall. Mater. Eng.* **2025**, *31*, 466–483.
23. Yan, H.; Xiao, J.; Zhang, B.; et al. The Application of Natural Language Processing Technology in the Era of Big Data. *J. Ind. Eng. Appl. Sci.* **2024**, *2*, 20–27.
24. Dai, S.; Yang, X.; Liu, H.; et al. Cocktail: A Comprehensive Information Retrieval Benchmark with LLM-Generated Documents Integration. In *Findings of the Association for Computational Linguistics: ACL 2024*, Ku, L.-W., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 7052–7074.
25. Genesis, J.; Keane, F. Integrating Knowledge Retrieval with Generation: A Comprehensive Survey of RAG Models in NLP. *arXiv preprint* **2025**, *202504.0351*. Available online: https://www.preprints.org/manuscript/202504.0351
26. Gan, A.; Li, B.; Zhou, Q.; et al. Retrieval-Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey. *arXiv preprint* **2025**, *arXiv.2504.14891*. [CrossRef]
27. Lenadora, D.; Gamage, G.; Haputhanthri, D.; et al. Exploratory Analysis of a Social Media Network in Sri Lanka During the COVID-19 Virus Outbreak. *arXiv preprint* **2020**, *arXiv.2006.07855*. [CrossRef]

28. Cheng, M.; Wang, J.; Li, S.; et al. A Survey on Knowledge-Oriented Retrieval-Augmented Generation. *arXiv preprint* **2025**, *arXiv.2503.10677*. [CrossRef]
29. Mei, L.; Mo, S.; Yang, Z.; et al. A Survey of Multimodal Retrieval-Augmented Generation. *arXiv preprint* **2025**, *arXiv.2504.08748*. [CrossRef]
30. Subi, S.; Shanthini, B.; SilpaRaj, M.; et al. Natural Language Processing Techniques for Information Retrieval Enhancing Search Engines with Semantic Understanding. *ITM Web Con.* **2025**, *76*. [CrossRef]
31. Zhu, Z.; Huang, T.; Wang, K.; et al. Graph-Based Approaches and Functionalities in Retrieval-Augmented Generation: A Comprehensive Survey. *arXiv preprint* **2025**, *arXiv.2504.10499*. [CrossRef]
32. Le, D.-V.-T.; Bigo, L.; Herremans, D.; et al. Natural Language Processing Methods for Symbolic Music Generation and Information Retrieval: A Survey. *ACM Comput. Surv.* **2025**, *57*, 1–40.
33. Wu, Y.; Liao, L.; Fang, Y. Retrieval-Augmented Generation for Dynamic Graph Modeling. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, Padua, Italy, 13–18 July 2025; pp. 1434–1443.
34. Nawaz, A. Advancements in Natural Language Processing for Multilingual Information Retrieval Systems. *Multidiscip. Res. Comput. Inf. Syst.* **2025**, *5*, 197–216.
35. Li, M.; Zhang, Y.; Fang, X.; et al. A Survey of Long-Document Retrieval in the PLM and LLM Era. *arXiv preprint* **2025**, *arXiv.2509.07759*. [CrossRef]
36. Hu, Y.; Lu, Y. RAG and RAU: A Survey on Retrieval-Augmented Language Model in Natural Language Processing. *arXiv preprint* **2024**, *arXiv.2404.19543*. [CrossRef]
37. Jiang, X.; Wang, W.; Tian, S.; et al. Applications of Natural Language Processing and Large Language Models in Materials Discovery. *npj Comput. Mater.* **2025**, *11*. [CrossRef]
38. Le, K.D.R.; Tay, S.B.P.; Choy, K.T.; et al. Applications of Natural Language Processing Tools in the Surgical Journey. *Front. Surg.* **2024**, *11*. [CrossRef]
39. Alomari, E. Unlocking the Potential: A Comprehensive Systematic Review of ChatGPT in Natural Language Processing Tasks. *Comput. Model. Eng. Sci.* **2024**, *141*, 43–85. [CrossRef]
40. Kim, H.-W.; Shin, D.-H.; Kim, J.; et al. Assessing the Performance of ChatGPT's Responses to Questions Related to Epilepsy: A Cross-Sectional Study on Natural Language Processing and Medical Information Retrieval. *Seizure* **2024**, *114*, 1–8.