A Survey of Model Architectures in Information Retrieval

Anonymous ACL submission

Abstract

The period from 2022 to the present has represented one of the biggest paradigm shifts in information retrieval (IR) and natural language processing (NLP). This work surveys the evolution of model architectures in IR, focusing on two key aspects: backbone models for feature extraction and end-to-end system architectures for relevance estimation. The review intentionally separates architectural considerations from training methodologies to provide a focused analysis of structural innovations in IR systems. We trace the development from traditional term-based methods to modern neural approaches, particularly highlighting the impact of transformer-based models and subsequent large language models (LLMs). We conclude with a forward-looking discussion of emerging challenges and future directions, including architectural optimizations for performance and scalability, handling of multimodal, multilingual data, and adaptation to novel application domains such as autonomous search agents that is beyond traditional search paradigms.

1 Introduction

Information Retrieval (IR) aims to retrieve relevant information sources to satisfy users' information needs. In the past decades, IR has become indispensable for efficiently and effectively accessing vast amounts of information across various applications. Beyond its traditional role, IR now also plays a critical role in assisting large language models (LLMs) to generate grounded and factual responses. Research in IR primarily centers on two key aspects: (1) extracting better query and document feature representations, and (2) developing more accurate relevance estimators. The approaches for extracting query and document features have evolved from traditional term-based methods, such as boolean logic and vector space models, to modern solutions such as dense retrieval based on pre-trained language models (Lin et al., 2022).

Relevance estimators have evolved alongside advances in feature representations. Early approaches, including probabilistic and statistical language models, computed relevance with simple similarity functions based on term-based features. Learning-to-rank (LTR) techniques later emerged, incorporating machine learning models and multi-layer neural networks for relevance estimation (Li, 2011). The success of LTR methods can be largely attributed to their extensive use of manually engineered features, derived from both statistical properties of text terms and user behavior data collected from web browsing traffic (Qin and Liu, 2013). In 2010s, a vast literature explored neural rerankers in different architectures to capture the semantic similarity between queries and documents. Then pre-trained transformers, represented by BERT (Devlin et al., 2019), quickly revolutionized the model design, leading to an era where retrieval and ranking models adopt simpler architectures for relevance estimation, such as dot product operations and MLP layer prediction heads, which operate on learned neural representations (Karpukhin et al., 2020; Nogueira et al., 2020; Lin et al., 2022).

Recent advancements of LLMs have revolutionized applied machine learning (ML) communities, including IR. One intriguing property of LLMs is that they can be used for feature extraction and relevance estimation, achieving strong performance without extensive training (Ni et al., 2022a; Neelakantan et al., 2022; BehnamGhader et al., 2024; Sun et al., 2023; Qin et al., 2024a, *inter alia*). The rise of LLMs in IR builds upon a rich foundation of transformer-based pre-trained language models that have evolved from earlier neural architectures. These include Transformers (Vaswani et al., 2017), Recurrent Neural Networks (RNN, Elman, 1990), Attention (Bahdanau, 2014) and pre-trained static neural representations such as Word2Vec (Mikolov, 2013) and GloVe (Pennington et al., 2014).

This work reviews the evolution of model ar-

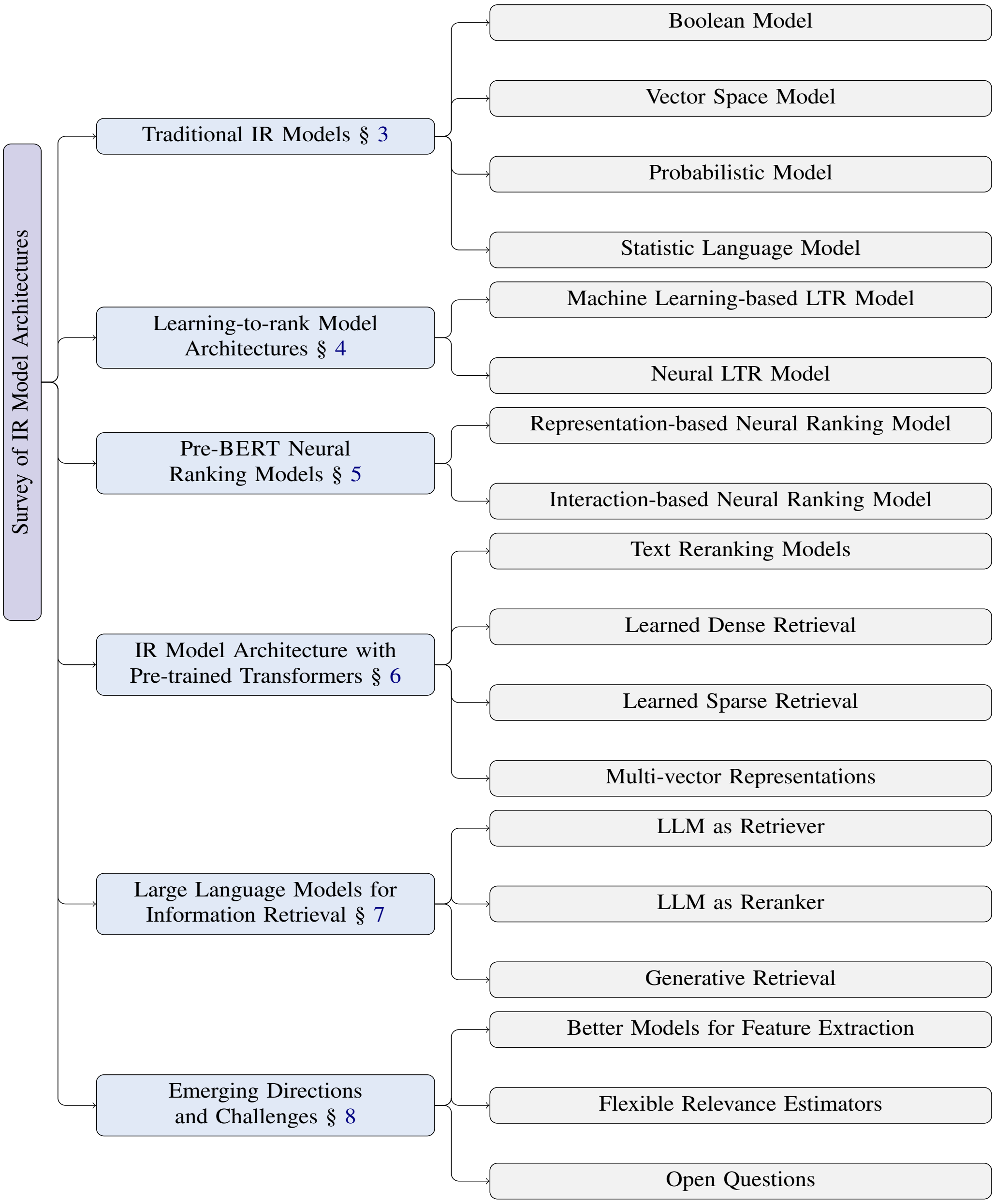


Figure 1: An overview of this survey. We focus on representative lines of works and defer details to the Appendix.

chitectures in IR (with an overview in Fig. 1). Here, the meaning of model architecture is twofold: it describes (1) backbone models for extracting query and document feature representations, and (2) end-to-end system architectures that process raw inputs, perform feature extraction, and estimate relevance. Different from prior works and surveys (Lin et al., 2022; Zhu et al., 2023), we intentionally separate our discussion of model architectures from training methodologies and deployment best practices to provide a focused architectural analysis. The shift towards neural architectures, particularly Transformer-based models, has fundamentally transformed IR by enabling rich, contextualized representations and improved handling of complex queries. While this evolution has enhanced retrieval precision, it also presents new challenges, especially with the emergence of LLMs. These challenges include the need for architectural innovations to optimize performance and scalability, handling multimodal and multilingual data, incorporating domain-specific knowledge and understanding complex instructions. Moreover, as IR systems are increasingly integrated into diverse applications — from robotics (Xie et al., 2024), protein structure discovery (Jumper et al., 2021) to autonomous agents (Wu et al., 2023; Chen et al., 2025) that are capable of reasoning and search —

the field must evolve beyond traditional search paradigms. We conclude this survey by examining these challenges and discuss their implications for the future of IR model architectures research.

2 Background and Terminology

We focus on the classical *ad hoc* retrieval task.

Task Definition and Evaluation Given a query $\mathcal{Q}$, the task is to find a ranked list of k documents, denoted as $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k\}$, that exhibit the highest relevance to $\mathcal{Q}$. This is achieved either by *retrieving* top-k documents from a large collection $\mathcal{C}$ ($|\mathcal{C}| \gg |k|$), which typically comprises millions or billions of documents, or by *reranking* the top-k candidates returned by a retriever. System performance is measured using standard IR metrics such as Mean Reciprocal Rank, Recall, and normalized Discounted Cumulative Gain (nDCG).

Query and Document A *query* expresses an information need and serves as input to the *ad hoc* retrieval system. We denote *document* as the atomic unit for retrieval and ranking. Our discussions are primarily based on text-based documents, but it can also refer to a webpage or an email, depending on the actual IR application of interest.

Disentangling Model Architecture with Training Strategies Similar to other applied ML domains, the design of IR model architectures is paired with training strategies and deployment best practices. In this paper, we seek to disentangle the two and focus solely on architectures. We refer to prior surveys for a more focused review of training strategies and related topics (Schütze et al., 2008; Lin et al., 2022; Song et al., 2023).

3 Traditional IR Models

In this section, we briefly review traditional IR models prior to neural methods, with a focus on *boolean model* and *vector space model* which serve as the foundation of later development of IR models (§§ 4 to 7).[1] These models are built upon the basic unit "term" used in the representation (Nie, 2010).

Boolean Model In Boolean Models, a document $\mathcal{D}$ is represented by a set of terms it contains, i.e., $\mathcal{D} = \{t_1, t_2, \ldots, t_n\}$. A query $\mathcal{Q}$ is represented as a similar boolean expression of terms. A document is considered relevant to a query only if a logical implication $\mathcal{D} \rightarrow \mathcal{Q}$ holds, i.e., the document representation logically implies the query expression.

Vector Space Model In Vector Space Models (Salton et al., 1975), the queries and documents are represented by vectors, e.g., $\mathcal{Q} =< q_1, q_2, \ldots, q_n >$ and $\mathcal{D} =< d_1, d_2, \ldots, d_n >$. The vector space $\mathcal{V} =< t_1, t_2, \ldots, t_n >$ is formed by all the terms the system recognizes in the documents and each element (q_i or d_i, $1 \leq i \leq n$) in the vectors represents the weight of the corresponding term in the query or the document. The weights q_i or d_i could be binary, representing presence or absence. Given the vector representations, the relevance score is estimated by a similarity function between the query $\mathcal{Q}$ and the document $\mathcal{D}$.

4 Learning-to-Rank Model Architectures

Different from traditional IR models discussed in § 3, Learning-to-Rank (LTR) leverages the idea of supervised ML on extensively crafted numerical features (Burges et al., 2005; Burges, 2010; Qin and Liu, 2013). For each $(\mathcal{Q}_i, \mathcal{D}_i)$ pair, a k-dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^k$ and a relevance label $\mathbf{y}_i$ is provided to the ranking model f parameterized by θ. Denote the loss function as $l(\cdot)$, the ranking is trained to minimize the empirical loss on labeled training set Ψ: $\mathcal{L} = 1/|\Psi| \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \Psi} l(f_\theta(\mathbf{x}_i), \mathbf{y_i})$.

Explorations in LTR models can be grouped into two directions: *ML-based models* and *neural LTR models*. Under the scope of ML models, RANKSVM (Joachims, 2006) is a pairwise LTR model based on SVM. Burges et al. (2005) studied decision trees and Wu et al. (2010) proposed LAMBDAMART based on Gradient Boosted Decision Trees (GBDT, Friedman, 2001; Ke et al., 2017). Unsurprisingly, early works also explored neural LTR models. RANKNET (Burges et al., 2005) and LAMBDARANK (Burges et al., 2006) parameterize the LTR model with neural networks. Recent works such as GSF (Ai et al., 2019) and APPROXNDCG (Bruch et al., 2019) use multiple fully connected layers. DLCM (Ai et al., 2018a) and SETRANK (Pang et al., 2020) adopt RNN and self-attention for reranking documents. Qin et al. (2021) conducted rigorous study of benchmarking neural ranking models against GBDT-based models. See Table 1 for a list of LTR models and prior surveys on LTR techniques (Liu, 2009; Li, 2011).

LTR techniques use human-crafted numerical features and metadata like PageRank score (Brin

[1]We defer the discussion of *probabilistic model* and *statistical language model* to Appx. A.

and Page, 1998) and click count as features and are still widely used in modern search systems (Google, 2019). However, it lacks the flexibility of being directly used on raw text data, and also cannot overcome the lexical mismatch problem — xthe main focus of neural ranking methods (§ 5).

5 Neural Ranking Models

Different from feature engineering of LTR (§ 4), neural ranking models utilize deep neural networks to learn feature representations directly from raw text and again use neural networks for relevance estimation.[2] Depending on how queries interact with documents during network processing, neural ranking models can be roughly divided into *representation-based models* and *interaction-based models* (Guo et al., 2016a).

Representation-based models can be regarded as extensions of vector space models (§ 3), which independently encode queries and documents into a latent vector space, as illustrated in Fig. 2a. The Deep Structured Semantic Model (DSSM) (Huang et al., 2013) is an early example. It utilizes word hashing and multilayer perceptrons (MLPs) to independently encode term vectors of queries and documents, enabling the computation of ranking scores based on the cosine similarity of their embeddings. Later works modify DSSM's encoder network to better capture richer semantic and contextual information. Convolutional DSSM (Shen et al., 2014) leverages a CNN architecture to project vectors within a context window to a local contextual feature vector. A variation of DSSM replaces MLPs with a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997; Palangi et al., 2016; Wan et al., 2016), utilizing its memory mechanism to capture local and global context information.

Interaction-based models (Fig. 2b), on the other hand, process queries and documents jointly through neural networks. The model's output is typically a scalar relevance score of the input query-document pair. Various network architectures have been proposed under this paradigm. MATCHPYRAMID (Pang et al., 2016) employs CNN over the interaction matrix between query and document terms. The interaction matrix is treated as an image, allowing the CNN to capture local matching patterns through convolution and pooling operations (Hu et al., 2014). Building upon the concept of interaction-focused models, Guo et al. (2016a) highlight the importance of exact term matches in neural ranking models and proposed the Deep Relevance Matching Model (DRMM). The model constructs matching histograms for each query term to capture the distribution of matching signals across document terms. Kernel-Based Neural Ranking Model (K-NRM, Xiong et al., 2017) further advances interaction-based approaches. It employs radial basis function kernels to transform the query-document interaction matrix informative ranking features. CONV-KNRM (Dai et al., 2018) later extends it to convolutional kernels.

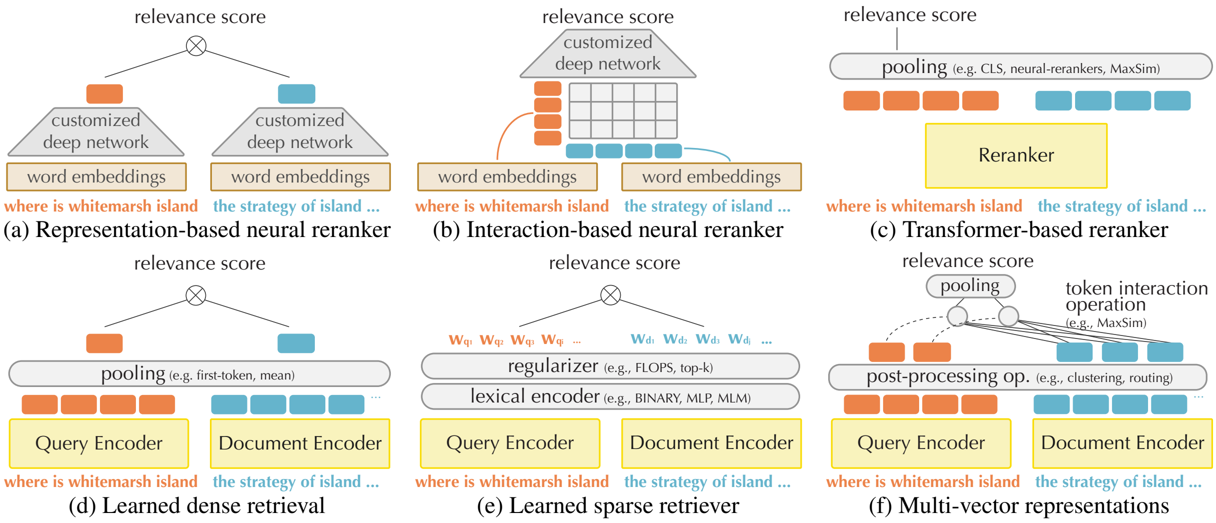
In addition to the development of network architecture, pre-trained embeddings (Mikolov, 2013; Pennington et al., 2014) provide semantic-based term representations to enable neural ranking models to focus on learning relevance matching patterns, improving training convergence and retrieval performance on both representation-based and interaction-based models (Levy et al., 2015).

6 IR with Pre-trained Transformers

BERT (Devlin et al., 2019) changed the research paradigm in both NLP and IR. Its success can be attributed to two factors: (1) the multi-head attention architecture (Vaswani et al., 2017) admits fine-grained, contextualized representations; (2) large-scale pre-training allows BERT to encode both semantics and world knowledge. The expressiveness of BERT has been extensively studied by prior works, e.g., Rogers et al. (2020); Tenney et al. (2019); Clark (2019). This sectino discusses IR architectures based on pre-trained transformers, with a focus on BERT-type encoder models.

Text Reranking Nogueira et al. (2019) first employed BERT model for reranking candidate passages from a first-stage retriever. Their model MONOBERT takes as input the sequence of concatenated $(\mathcal{Q}, \mathcal{D})$ as input, and outputs a relevance score s with a linear layer on top of the BERT model (Fig. 2c). The schema has later been proved to be effective on other pre-trained encoders (Zhang et al., 2021) and encoder-decoder architectures (Nogueira et al., 2020). However, this schema faces two challenges: (1) BERT family models has a limited 512 tokens context length, making reranking long documents challenging; (2) CLS token's single 768-dimensional representation

[2] We define neural information retrieval as retrieval models based on neural networks prior to pre-trained transformers. More models details are deferred to Appx. C and detailed surveys (Onal et al., 2018; Mitra et al., 2018; Xu et al., 2018)



(a) Representation-based neural reranker (b) Interaction-based neural reranker (c) Transformer-based reranker
(d) Learned dense retrieval (e) Learned sparse retriever (f) Multi-vector representations

Figure 2: Illustration on neural ranking models, retriever, and reranker architectures. Brown boxes indicate uncontextualized word embeddings (e.g., Word2vec). Yellow boxes indicate pretrained Transformers (e.g., BERT).

potentially limits the expressiveness of the reranking model. Two directions have been investigated to tackle these two challenges.

In the first direction, one strategy is to segment the long document into shorter passages, score each passage individually, then aggregate the scores to get a document-level relevance score (Yilmaz et al., 2019; Dai and Callan, 2019b). In this case, the underlying model architecture remains unchanged. A different line of work seeks to perform feature-level aggregation. PARADE (Li et al., 2020) uses an additional neural network to aggregate contextualized representations from CLS tokens of passages to get the final document relevance score.

In the second direction, MacAvaney et al. (2019) discovered via CEDR that the effectiveness of reranker could be enhanced when aggregating the contextualized representations with neural ranking models such as K-NRM and DRMM. Zhang et al. (2024a) later observed that integrating token representations with late interactions could also effectively improve the reranking robustness on out-of-domain scenarios, especially for long queries.

Learned Dense Retrieval Here we use the term "bi-encoder" (Humeau et al., 2020) to refer to the model architecture commonly used for dense retrieval.[3] Bi-encoder uses a backbone network to encode query $\mathcal{Q}$ and document $\mathcal{D}$ separately, then uses the encoded dense vector representations to compute the relevance score with similarity functions such as dot product or cosine similarity (Xiong et al., 2020; Karpukhin et al., 2020; Gao et al., 2021c). After training, the model encodes the collection into a dense vector index, and retrieval is performed with fast nearest neighbor search techniques (Johnson et al., 2019; Malkov and Yashunin, 2016). Different from representation-based neural ranking models where distinct architectures are proposed (§ 5), existing dense retrieval methods are mostly based on pre-trained transformer language models, with variance on pooling strategy and training methodologies (detailed in Appx. D).

Learned Sparse Retrieval Similar to learned dense retrieval, learned sparse retrieval (LSR) also uses bi-encoder architecture with language models as the backbone, to transform documents into a static index for later retrieval (Zamani et al., 2018). To use a traditional inverted index for faster retrieval (Bruch et al., 2024), the query $\mathcal{Q}$ and document $\mathcal{D}$ are represented as sparse vectors whose dimensionality typically matches the vocabulary size of the backbone pre-trained transformer model (Yu et al., 2024a). Here sparsity is enforced through regularization (Formal et al., 2021b; Paria et al., 2020) and usually serves as a trade-off between effectiveness and efficiency. At a higher level, LSR can be viewed as a way to learn token importance or "impact" scores from data (Dai and Callan, 2019b; Bai et al., 2020; Mallia et al., 2021), in contrast to static formulas like BM25.

Multi-Vector Representations Learned dense retrieval's bi-encoder architecture encodes queries and documents into single feature vectors sepa-

[3] The term "bi-encoder" is known by many other names, such as two-tower architecture, embedding models. We refer to it as "bi-encoder" in contrast to "cross-encoder", reranker architectures that take concatenated $(\mathcal{Q}, \mathcal{D})$ as input.

rately, and estimate relevance via a similarity function. This enables efficient training, indexing, and inference, but the lack of interactions between query and document terms potentially limits performance. In contrast, the "all-to-all" interaction of cross-encoder models are computationally expensive. Research has explored representing queries and documents using multiple vectors and developing corresponding relevance estimators. POLY-ENCODER (Humeau et al., 2020) computes a fixed number of vectors per query, and aggregate with softmax attention against document vectors. Due to the use of softmax operator, fast nearest neighbor search technique cannot be trivially applied. ME-BERT (Luan et al., 2021) proposes to represent documents with m vectors, and uses the maximum similarity between query vectors and document vectors to estimate relevance. COLBERT (Khattab and Zaharia, 2020) takes the multi-vector representation idea further and represents each token in query and document as a contextualized vector. Each query vector interacts with each document vector via a MaxSim operator, and the relevance score is computed by summing scalar outputs of these operators over query terms. COLBERT is trained end-to-end and achieves strong performance on public retrieval benchmarks. Numerous other studies further investigate token selection and aggregation operations, see Appx. D for details.

7 Large Language Models for IR

LLMs have exhibited proficiency in language understanding and generation, are trained to align with human preferences (OpenAI, 2023; Gemini et al., 2023; Bai et al., 2022) and able to perform complex tasks such as reasoning (Wei et al., 2022; Hurst et al., 2024; Guo et al., 2025) and planning (Song et al., 2023).[4] In this section, we briefly discuss some works that utilize LLMs for IR tasks.

LLM as Retriever Adopting an LLM as the backbone for bi-encoder retrieval model has achieved performance improvement compared to smaller-sized models like BERT. Neelakantan et al. (2022) fine-tuned a series of off-the-shelf GPT models towards text and code representation. They empirically verified that the bi-encoder retriever's performance can benefit from increased backbone language model capacity. Common text retrieval benchmarks like BEIR (Thakur et al., 2021) are currently dominated by LLM-based retrievers. A parallel line of research has explored adapting unidirectional LLM architectures into bidirectional ones to enhance representational power. LLM2VEC (BehnamGhader et al., 2024) enables bidirection and further trains LLAMA-2 (Touvron et al., 2023) with specific adaptive tasks. NV-EMBED (Lee et al., 2025) introduces a new latent attention layer and leads to improve on MTEB benchmark (Muennighoff et al., 2023).

LLM as Reranker Works discussed in § 6 have explored fine-tuning BERT-type encoder models as cross-encoder reranker. Later works further expand this paradigm to encoder-decoder models like T5 (Raffel et al., 2020) and decoder models like LLAMA (Touvron et al., 2023). Nogueira et al. (2020) fine-tuned T5 models with classification loss for passage reranking. Zhuang et al. (2023a) proposed to fine-tune T5 to produce a numerical relevance score, and optimize the model with ranking losses like RANKNET (Burges, 2010). LISTT5 (Yoon et al., 2024) adopts the Fusion-in-Decoder architecture (Izacard and Grave, 2021) to learn a listwise reranker. RANKLLAMA (Ma et al., 2024b) fine-tunes decoder model for pointwise reranking and achieves better performance compared to T5-based rerankers. Leveraging the long-context ability of LLMs, a reranking paradigm is introduced, where LLM-based rerankers directly rerank a list of documents rather than scoring each document individually (Ma et al., 2023; Zhang et al., 2023b; Pradeep et al., 2023b,c). Instead of using raw passages, Liu et al. (2024b) used passage embeddings as input to LLMs and trained corresponding rerankers to achieve improved efficiency. Aforementioned studies still rely on labeled data and gradient updates to backbone LLMs. With the rise of instruction-following LLMs, researchers have explored using LLMs as unsupervised rerankers through prompting techniques. As this line of research does not introduce architectural changes to existing LLMs, we refer to a recent survey (Zhu et al., 2023) for further details.

Generative Retrieval Traditional IR systems follow the "index-retrieval-rank" paradigm (Schütze et al., 2008). Although effective, jointly optimizing the separate index, retrieval, and reranking modules can be challenging. A recent line of research aims to bypass the indexing step by using autoregressive language models to directly generate document

[4] In this work, we use term LLM to denote language models which are trained for text generation, including encoder-decoder models and decoder-only models.

identifiers (DocIDs). DSI (Tay et al., 2022) first constructs semantically structured DocIDs, then fine-tunes T5 models with labeled data. In the decoding phase, DSI uses all DocIDs to construct a trie and performs constrained decoding. Followup works (Wang et al., 2022b; Bevilacqua et al., 2022) further improve upon this paradigm with strategies to construct semantic DocIDs and enable robust training. A significant challenge for generative retrieval is scalability to larger corpus (Pradeep et al., 2023a). Zeng et al. (2024) utilized the residual quantization technique to construct hierarchical DocIDs and achieved comparable performance to dense retrievers on MS MARCO dataset. Generative retrieval is an active research area; see (Li et al., 2025c) for a more comprehensive review.

Remarks We note that the adoption of LLMs in IR model architectures primarily follows two main themes: (1) feature extraction, and (2) relevance estimation, as discussed in § 1. For example, LLMs' semantic knowledge enables their strong performance in being the backbone of a retriever; and instruction-following LLMs can be directly prompted for relevance estimation. Generative retrieval and cross-encoder LLM reranking models are trained end-to-end for both feature extraction and relevance estimation. While LLMs have shown promise, several challenges and open questions remain, which leaves room for discussion (§ 8).

8 Emerging Directions and Challenges

IR systems have become crucial across diverse domains, from retrieval-augmented language modeling (Khandelwal et al., 2020a; Borgeaud et al., 2022) to applications in agents (Wu et al., 2023; Wang et al., 2024a), code generation (Wang et al., 2024c; Zhang et al., 2023a), robotics (Anwar et al., 2024), medicine (Jeong et al., 2024), and protein research (Jumper et al., 2021), *inter alia*. These developments present new challenges for IR research. Drawing from the evolution of IR architectures (§§ 3 to 7), we examine emerging trends, open problems, and potential research directions.[5]

8.1 Better Models for Feature Extraction

Scaling has been a winning recipe for modern neural networks (Kaplan et al., 2020; Hoffmann et al., 2022; Dehghani et al., 2023; Fang et al., 2024; Shao et al., 2024, *inter alia*). As IR moves toward compute-intensive practices, we identify key areas for model improvement:

- **Data & training efficiency** Current transformer-based IR models demand extensive training data (Fang et al., 2024), making them impractical for many real-world applications. Developing architectures that can learn effectively from limited data remains crucial. Additionally, models should support parallel processing and low precision training to reduce costs and accelerate convergence (Nvidia, 2021; Fishman et al., 2024; Liu et al., 2024a).
- **Inference optimization** Real-time applications like conversational search (Mo et al., 2024b) and agent-based systems (Yao et al., 2023) require efficient handling of variable-length queries, necessitating advanced compression and optimization techniques for both retriever backbones and index structures (Dettmers and Zettlemoyer, 2023; Kumar et al., 2024; Warner et al., 2024; Bruch et al., 2024; Xu et al., 2025a, *inter alia*).
- **Multimodality & multilinguality** Future IR systems must handle diverse content types including images (Ma et al., 2024a), audio (Pusateri et al., 2024), structured data (Tan et al., 2024; Edge et al., 2024) as well as multilinguality beyond English (Zhang et al., 2023c; Enevoldsen et al., 2025). Recent advances in multimodal, multilingual retrieval (Ma et al., 2024a; Wei et al., 2025; Yu et al., 2024b; Huang et al., 2024, *inter alia*) and structured data processing (Li et al., 2023d, 2024) have demonstrated promises.
- **Transformer alternatives** While transformers have dominated recent IR research, their quadratic complexity in attention computation remains a significant bottleneck. Recent advances in linear RNNs (Peng et al., 2023, 2024; Qin et al., 2024b), state space models (Gu and Dao, 2024; Dao and Gu, 2024), and linear attention (Katharopoulos et al., 2020; Yang et al., 2024) offer alternatives with theoretical linear complexity. Although preliminary studies (Xu et al., 2025b) show limited gains compared to optimized transformers, developing efficient alternatives architectures for transformers could revolutionize large-scale information retrieval.

Strong foundation models have proven crucial for IR performance (Neelakantan et al., 2022; Ma et al., 2024b). As IR applications expand, developing foundation models that balance computational efficiency with robust performance across tasks and

[5] See Appx. F for an extended discussion *w.r.t.* deployment challenges, robustness, autonomous search agents etc.

modalities emerges as a key research priority.

8.2 Flexible Relevance Estimators

As discussed in § 6, cross-encoders provide complex non-linear relevance estimation but are computationally expensive. In contrast, bi-encoder architectures used in dense and sparse retrieval rely on linear similarity functions (e.g., inner product) to enable fast retrieval through nearest neighbor search and inverted indexing. Balancing complex relevance matching and scalable retrieval remains challenging. COLBERT (Khattab and Zaharia, 2020) addresses this by using document representation matrices with MaxSim operations, while recent work (Killingback et al., 2025) explores Hypernetworks (Ha et al., 2022) to generate query-specific neural networks for relevance estimation. The design of flexible yet scalable relevance estimators remains an active research direction.

8.3 Open Questions

The integration of IR systems into other research domains presents new challenges. We discuss key implications for future IR modeling research.

End "User" of Retrieval While traditional IR systems focus on providing search results to humans, retrieval is increasingly used to support ML models, particularly LLMs, in tasks such as generation (Gao et al., 2023), reasoning (Yao et al., 2024; Islam et al., 2024), and planning (Song et al., 2023). This shifting paradigm raises questions about task formulation, evaluation, and system optimization:

- Current IR research is grounded in human information-seeking behavior (Wilson, 2000). When the end user becomes another ML model, we must reconsider how to define and assess *relevance*. This question suggests a need for flexible, data-efficient models that are adaptable to various downstream tasks.
- Traditional IR metrics, which are designed for human-centric evaluation, may not align with downstream task performance in retrieval-augmented systems. Future IR models should support end-to-end system optimization rather than focusing solely on ranking metrics.

Autonomous Search Agent Complex tasks require retrieving long-tail knowledge using lengthy, complex queries (Soudani et al., 2024; Su et al., 2024), demanding retrieval models capable of instruction following (Weller et al., 2024a; Ravfogel et al., 2024) and reasoning (Su et al., 2024). Existing attempts can be divided into two directions. One line of works focuses on training retrievers and rerankers that are capable of reasoning. They propose data pipelines to synthesize training data (Oh et al., 2024; Weller et al., 2024b; Shao et al., 2025, *inter alia*) and leverage strong backbone language models such as LLAMA (Dubey et al., 2024) and QWEN (Yang et al., 2025). Another line of works treats search/retrieval as an integral component of LLM reasoning chain. They consider search/retrieval system as a static tool that can be called via large reasoning model (LRM), and instead focus on improving LRM's capability to use search tool and synthesize search results. The LRM can decide where, when and how to conduct search, and the search results subsequently influence LRM's further reasoning and decision making (Nakano et al., 2021; Tang et al., 2025; He et al., 2025; Chen et al., 2025; Li et al., 2025a).

Despite the exciting progress, key limitations remain in building instruction following and reasoning-capable retrieval systems. For example, to enable retrievers' reasoning capability requires strong backbone models (e.g., 7B scale), which are often infeasible for production systems. Even larger models (e.g., 32B scale) augmented with retrieval and trained via expensive reinforcement learning (Jin et al., 2025; Chen et al., 2025) still underperform simpler baselines with query decomposition and chain-of-thought prompting (Khot et al., 2023; Trivedi et al., 2023). A key open question is how to endow retrievers with strong reasoning capabilities using lightweight, scalable models. Another open challenge lies in the joint optimization of retrievers and language models within a unified, reasoning-aware framework. Lastly, the human factors of applying such autonomous search agents remains to be studied.

9 Conclusions and Closing Thoughts

Information retrieval modeling has evolved from simple term matching to complex neural networks and LLM-driven approaches, significantly improving search capabilities. Key challenges ahead include balancing computational efficiency with performance, handling diverse data types, maintaining faithfulness and trustworthiness, and integrating with emerging technologies like autonomous agents. These challenges drive opportunities for developing more adaptive, efficient, scalable and intelligent retrieval systems.

Limitations

This survey examines the evolution of IR models, with particular emphasis on challenges arising from LLMs and their implications for future architectures. Due to space constraints, we focus on representative works rather than providing an exhaustive review, with supplementary discussions of interdisciplinary research included in the Appendix.

Ethical Considerations

As this paper solely reviews existing IR developments and future research directions, we believe it presents no direct ethical concerns.

References

Qingyao Ai, Keping Bi, Jiafeng Guo, and W Bruce Croft. 2018a. Learning a deep listwise context model for ranking refinement. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 135–144.

Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W Bruce Croft. 2018b. Unbiased learning to rank with unbiased propensity estimation. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 385–394.

Qingyao Ai, Jiaxin Mao, Yiqun Liu, and W. Bruce Croft. 2018c. Unbiased learning to rank: Theory and practice. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2305–2306, New York, NY, USA. Association for Computing Machinery.

Qingyao Ai, Xuanhui Wang, Sebastian Bruch, Nadav Golbandi, Michael Bendersky, and Marc Najork. 2019. Learning groupwise multivariate scoring functions using deep neural networks. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, pages 85–92.

Bang An, Shiyue Zhang, and Mark Dredze. 2025. RAG LLMs are not safer: A safety analysis of retrieval-augmented generation for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5444–5474, Albuquerque, New Mexico. Association for Computational Linguistics.

Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. 2024. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. *arXiv preprint arXiv:2409.13682*.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675, Toronto, Canada. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024a. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen-tau Yih. 2024b. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Dzmitry Bahdanau. 2014. Neural machine translation by jointly learning to align and translate. *The Twelfth International Conference on Learning Representations*.

Yang Bai, Xiaoguang Li, Gang Wang, Chaoliang Zhang, Lifeng Shang, Jun Xu, Zhaowei Wang, Fangshan Wang, and Qun Liu. 2020. Sparterm: Learning term-based sparse representation for fast text retrieval. *arXiv preprint arXiv:2010.00768*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 222–229.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. In *Advances in Neural Information Processing Systems*.

Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 405–414.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2387–2392.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Leonid Boytsov, Preksha Patel, Vivek Sourabh, Riddhi Nisar, Sayani Kundu, Ramya Ramanathan, and Eric Nyberg. 2024. Inpars-light: Cost-effective unsupervised training of efficient rankers. *Transactions on Machine Learning Research*.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30:107–117.

Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. 2024. Efficient inverted indexes for approximate retrieval over learned sparse representations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 152–162.

Sebastian Bruch, Masrour Zoghi, Michael Bendersky, and Marc Najork. 2019. Revisiting approximate metric optimization in the age of deep neural networks. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1241–1244.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96.

Christopher Burges, Robert Ragno, and Quoc Le. 2006. Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems*, 19.

Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81.

Tianyu Cao, Neel Bhandari, Akhila Yerukola, Akari Asai, and Maarten Sap. 2025. Out of style: Rag's fragility to linguistic variation. *arXiv preprint arXiv:2504.08231*.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, et al. 2024. Main-rag: Multi-agent filtering retrieval-augmented generation. *arXiv preprint arXiv:2501.00332*.

Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin. 2010. Training and testing low-degree polynomial data mappings via linear svm. *Journal of Machine Learning Research*, 11(48):1471–1490.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, et al. 2025. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Kevin Clark. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Charles LA Clarke and Laura Dietz. 2024. Llm-based relevance assessment still can't replace human relevance assessment. *arXiv preprint arXiv:2412.17156*.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024. Neural retrievers are biased towards llm-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 526–537.

Zhuyun Dai and Jamie Callan. 2019a. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv preprint arXiv:1910.10687*.

Zhuyun Dai and Jamie Callan. 2019b. Deeper text understanding for ir with contextual neural language modeling. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 126–134, New York, NY, USA. Association for Computing Machinery.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*.

Jeffrey Dean et al. 2009. Challenges in building large-scale information retrieval systems. In *Keynote of the 2nd ACM international conference on web search and data mining (WSDM)*, volume 10.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. 2023. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR.

Gelei Deng, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, and Yang Liu. 2024. Pandora: Jailbreak gpts by retrieval augmented generation poisoning. *arXiv preprint arXiv:2402.08416*.

Zehang Deng, Yongjian Guo, Changzhou Han, Wanlun Ma, Junwu Xiong, Sheng Wen, and Yang Xiang. 2025. Ai agents under threat: A survey of key security challenges and future pathways. *ACM Computing Surveys*, 57(7):1–36.

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pages 7750–7774. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50.

Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2024. Who determines what is relevant? humans or ai? why not both? *Communications of the ACM*, 67(4):31–34.

Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. 2024. Scaling laws for dense retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1339–1349.

Maxim Fishman, Brian Chmiel, Ron Banner, and Daniel Soudry. 2024. Scaling fp8 training to trillion-token llms. *arXiv preprint arXiv:2409.12517*.

Thibault Formal, Stéphane Clinchant, Hervé Déjean, and Carlos Lassance. 2024. Splate: Sparse late interaction retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2635–2640.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021a. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. *SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking*, page 2288–2292. Association for Computing Machinery, New York, NY, USA.

Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Norbert Fuhr. 1992. Probabilistic models in information retrieval. *The computer journal*, 35(3):243–255.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2020. Understanding bert rankers under distillation. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 149–152.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021a. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3030–3042, Online. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021b. Complement lexical retrieval model with semantic residual embeddings. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*, pages 146–160. Springer.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021c. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Yunfan Gao, Yun Xiong, Yijie Zhong, Yuxi Bi, Ming Xue, and Haofen Wang. 2025. Synergizing rag and reasoning: A systematic review. *arXiv preprint arXiv:2504.15909*.

Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Google. 2024. Grounding with google search.

AI Blog Google. 2019. Understanding searches better than ever before.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *Second Agent Learning in Open-Endedness Workshop*.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. 2025. Deeprag: Thinking to retrieval step by step for large language models. *arXiv preprint arXiv:2502.01142*.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, et al. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016a. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 55–64, New York, NY, USA. Association for Computing Machinery.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016b. Semantic matching by non-linear word transportation for information retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 701–710, New York, NY, USA. Association for Computing Machinery.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Ashim Gupta, Rishanth Rajendhran, Nathan Stringham, Vivek Srikumar, and Ana Marasovic. 2024. Whispers of doubt amidst echoes of triumph in NLP robustness. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5533–5590, Mexico City, Mexico. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

David Ha, Andrew M Dai, and Quoc V Le. 2022. Hypernetworks. In *International Conference on Learning Representations*.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, et al. 2025. Pasa: An llm agent for comprehensive academic paper search. *arXiv preprint arXiv:2501.10120*.

S Hochreiter and J Schmidhuber. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020a. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*.

Sebastian Hofstätter, Omar Khattab, Sophia Althammer, Mete Sertkan, and Allan Hanbury. 2022. Introducing neural bag of whole-words with colberter: Contextualized late interactions using enhanced reduction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 737–747.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.

Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and Allan Hanbury. 2020b. Local self-attention over long text for efficient document retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2021–2024.

Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. 2020c. Interpretable & time-budget-constrained contextualization for re-ranking. In *ECAI 2020*, pages 513–520. IOS Press.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2042–2050, Cambridge, MA, USA. MIT Press.

Yunhai Hu, Yilun Zhao, Chen Zhao, and Arman Cohan. 2025. Mcts-rag: Enhancing retrieval-augmented generation with monte carlo tree search. *arXiv preprint arXiv:2503.20757*.

Ziniu Hu, Yang Wang, Qu Peng, and Hang Li. 2019. Unbiased lambdamart: an unbiased pairwise learning-to-rank algorithm. In *The World Wide Web Conference*, pages 2830–2836.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2333–2338, New York, NY, USA. Association for Computing Machinery.

Zhiqi Huang, Puxuan Yu, Shauli Ravfogel, and James Allan. 2024. Language concept erasure for language-invariant dense retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13261–13273, Miami, Florida, USA. Association for Computational Linguistics.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-RAG: Enhanced retrieval augmented reasoning with open-source large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14231–14244, Miami, Florida, USA. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 217–226, New York, NY, USA. Association for Computing Machinery.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017a. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, volume 51, pages 4–11. Acm New York, NY, USA.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017b. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 781–789.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Armand Joulin, Piotr Bojanowski, Tomáš Mikolov, Hervé Jégou, and Édouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020a. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020b. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Julian Killingback, Hansi Zeng, and Hamed Zamani. 2025. Hypencoder: Hypernetworks for information retrieval. *arXiv preprint arXiv:2502.05364*.

Weize Kong, Jeffrey M Dudek, Cheng Li, Mingyang Zhang, and Michael Bendersky. 2023. Sparseembed: Learning sparse lexical representations with contextual embeddings for retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2399–2403.

Donald H Kraft and Duncan A Buell. 1983. Fuzzy sets and generalized boolean retrieval systems. *International journal of man-machine studies*, 19(1):45–56.

Tanishq Kumar, Zachary Ankner, Benjamin F Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. 2024. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *ArXiv*, abs/1906.00300.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. Parade: Passage representation aggregation fordocument reranking. *ACM Transactions on Information Systems*, 42:1 – 26.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023a. Making large language models a better foundation for dense retrieval. *arXiv preprint arXiv:2312.15503*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Hang Li. 2011. Learning for ranking aggregation. In *Learning to Rank for Information Retrieval and Natural Language Processing*, pages 33–35. Springer.

Millicent Li, Tongfei Chen, Benjamin Van Durme, and Patrick Xia. 2024. Multi-field adaptive retrieval. *arXiv preprint arXiv:2410.20056*.

Minghan Li, Sheng-Chieh Lin, Xueguang Ma, and Jimmy Lin. 2023c. Slim: Sparsified late interaction for multi-vector retrieval with inverted indexes. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1954–1959, New York, NY, USA. Association for Computing Machinery.

Sen Li, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, and Qianli Ma. 2021. Embedding-based product retrieval in taobao search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3181–3189.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025a. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025b. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025c. From matching to generation: A survey on generative information retrieval. *ACM Trans. Inf. Syst.*, 43(3).

Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu, and Ge Yu. 2023d. Structure-aware language model pretraining improves dense retrieval on structured data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11560–11574.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023e. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Jimmy Lin and Xueguang Ma. 2021. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. *arXiv preprint arXiv:2106.14807*.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2022. *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Qi Liu, Bo Wang, Nan Wang, and Jiaxin Mao. 2024b. Leveraging passage embeddings for efficient listwise reranking with large language models. In *THE WEB CONFERENCE 2025*.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, and Maarten de Rijke. 2025. Robust information retrieval. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 1008–1011.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. Association for Computational Linguistics.

Xueguang Ma, Kai Sun, Ronak Pradeep, and Jimmy Lin. 2021. A replication study of dense passage retriever. *arXiv preprint arXiv:2104.05740*.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024b. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *Preprint*, arXiv:2305.02156.

Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. ABNIRML: Analyzing the behavior of neural IR models. *Transactions of the Association for Computational Linguistics*, 10:224–239.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1101–1104.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Yury Malkov and Dmitry A. Yashunin. 2016. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:824–836.

Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. 2021. Learning passage impacts for inverted indexes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1723–1727.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

David RH Miller, Tim Leek, and Richard M Schwartz. 1999. A hidden markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 214–221.

Bhaskar Mitra, Nick Craswell, et al. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126.

Bhaskar Mitra, Sebastian Hofstätter, Hamed Zamani, and Nick Craswell. 2021. Improving transformer-kernel ranking model using conformer and query term independence. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1697–1702.

Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024a. Chiq: Contextual history enhancement for improving query rewriting in conversational search. *arXiv preprint arXiv:2406.05013*.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024b. A survey of conversational search. *arXiv preprint arXiv:2410.15576*.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. Convgqr: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4998–5012.

Fengran Mo, Bole Yi, Kelong Mao, Chen Qu, Kaiyu Huang, and Jian-Yun Nie. 2024c. Convsdg: Session data generation for conversational search. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1634–1642.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 429–438.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Niklas Muennighoff, Hongjin SU, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Haya Nachimovsky, Moshe Tennenholtz, and Oren Kurland. 2025. A multi-agent perspective on modern information retrieval. *arXiv preprint arXiv:2502.14796*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Eric T. Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. 2016. Improving document ranking with dual word embeddings. *Proceedings of the 25th International Conference Companion on World Wide Web*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022a. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022b. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jian-Yun Nie. 2010. *Cross-language information retrieval*. Morgan & Claypool Publishers.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *Preprint*, arXiv:2003.06713.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*.

Cicero Nogueira dos Santos, Xiaofei Ma, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Beyond [CLS] through ranking by generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1722–1727, Online. Association for Computational Linguistics.

Zach Nussbaum, John Xavier Morris, Andriy Mulyar, and Brandon Duderstadt. 2025. Nomic embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*. Reproducibility Certification.

Team Nvidia. 2021. Accelerating inference with sparsity using the nvidia ampere architecture and nvidia tensorrt. In *NVIDIA Technical Blog*.

Hanseok Oh, Hyunji Lee, Seonghyeon Ye, Haebin Shin, Hansol Jang, Changwook Jun, and Minjoon Seo. 2024. Instructir: A benchmark for instruction following of information retrieval models. *arXiv preprint arXiv:2402.14334*.

Kezban Dilek Onal, Ye Zhang, Ismail Sengor Altingovde, Md Mustafizur Rahman, Pinar Karagoz, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, et al. 2018. Neural information retrieval: At the end of the early years. *Information Retrieval Journal*, 21:111–182.

Harrie Oosterhuis. 2021. Computationally efficient optimization of plackett-luce ranking models for relevance and fairness. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1023–1032.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2024. Introducing chatgpt search.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 24(4):694–707.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2793–2799. AAAI Press.

Liang Pang, Jun Xu, Qingyao Ai, Yanyan Lan, Xueqi Cheng, and Jirong Wen. 2020. Setrank: Learning a permutation-invariant ranking model for information retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 499–508.

Biswajit Paria, Chih-Kuan Yeh, Ian E.H. Yen, Ning Xu, Pradeep Ravikumar, and Barnabás Póczos. 2020. Minimizing flops to learn efficient sparse representations. In *International Conference on Learning Representations*.

Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. 2020. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of the web conference 2020*, pages 1194–1204.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, Singapore. Association for Computational Linguistics.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, et al. 2024. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. SIGIR 1998*, pages 275–281.

Ronak Pradeep, Kai Hui, Jai Gupta, Adam Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Tran. 2023a. How does generative retrieval scale to millions of passages? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1305–1321, Singapore. Association for Computational Linguistics.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *Preprint*, arXiv:2309.15088.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023c. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *Preprint*, arXiv:2312.02724.

Ernest Pusateri, Anmol Walia, Anirudh Kashi, Bortik Bandyopadhyay, Nadia Hyder, Sayantan Mahinder, Raviteja Anantha, Daben Liu, and Sashank Gondala. 2024. Retrieval augmented correction of named entity speech recognition errors. *arXiv preprint arXiv:2409.06062*.

Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597.

Tao Qin, Tie-Yan Liu, and Hang Li. 2010. A general approximation framework for direct optimization of information retrieval measures. *Information retrieval*, 13:375–397.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024a. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Zhen Qin, Le Yan, Honglei Zhuang, Yi Tay, Rama Kumar Pasumarthi, Xuanhui Wang, Michael Bendersky, and Marc Najork. 2021. Are neural rankers still outperformed by gradient boosted decision trees? In *International Conference on Learning Representations*.

Zhen Qin, Songlin Yang, and Yiran Zhong. 2024b. Hierarchically gated recurrent neural network for sequence modeling. *Advances in Neural Information Processing Systems*, 36.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *North American Chapter of the Association for Computational Linguistics*.

Tadeusz Radecki. 1979. Fuzzy set theoretical approach to document retrieval. *Information Processing & Management*, 15(5):247–259.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Razieh Rahimi, Youngwoo Kim, Hamed Zamani, and James Allan. 2021. Explaining documents' relevance to search queries. *arXiv preprint arXiv:2111.01314*.

Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2023. What are you token about? dense retrieval as distributions over the vocabulary. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2481–2498, Toronto, Canada. Association for Computational Linguistics.

Shauli Ravfogel, Valentina Pyatkin, Amir David Nissan Cohen, Avshalom Manevich, and Yoav Goldberg. 2024. Description-based text similarity. In *First Conference on Language Modeling*.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Ren, Hongyan Tang, and Siwen Zhu. 2022. Unbiased learning to rank with biased continuous feedback. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1716–1725.

Stephen E Robertson and K Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gerard Salton, Edward A Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. 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*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? *arXiv preprint arXiv:2410.09087*.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. 2024. Scaling retrieval-based language models with a trillion-token datastore. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, et al. 2025. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, page 373–374, New York, NY, USA. Association for Computing Machinery.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2219–2228.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009.

Fei Song and W Bruce Croft. 1999. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, page 12–22, New York, NY, USA. Association for Computing Machinery.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, et al. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.

Jiejun Tan, Zhicheng Dou, Wen Wang, Mang Wang, Weipeng Chen, and Ji-Rong Wen. 2024. Html-rag: Html is better than plain text for modeling retrieved knowledge in rag systems. *arXiv preprint arXiv:2411.02959*.

Jiabin Tang, Tianyu Fan, and Chao Huang. 2025. Autoagent: A fully-automated and zero-code framework for llm agents. *arXiv e-prints*, pages arXiv–2502.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index. In *Advances in Neural Information Processing Systems*.

Michael Taylor, John Guiver, Stephen Robertson, and Tom Minka. 2008. Softrank: optimizing non-smooth rank metrics. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 77–86.

Wilson L. Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism & Mass Communication Quarterly*, 30:415 – 433.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Yu Tian, Xiao Yang, Jingyuan Zhang, Yinpeng Dong, and Hang Su. 2023. Evil geniuses: Delving into the safety of llm-based agents. *arXiv preprint arXiv:2311.11855*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

C Van Rijsbergen. 1979. Information retrieval: theory and practice. In *Proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems*, volume 79, pages 1–14.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. 2004. A primer on kernel methods.

Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. 2016. A deep architecture for semantic matching with multiple positional sentence representations. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2835–2841. AAAI Press.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024b. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. 2025. Chain-of-retrieval augmented generation. *arXiv preprint arXiv:2501.14342*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 115–124.

Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. 2018. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 610–618.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Hao Sun, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, and Mao Yang. 2022b. A neural corpus indexer for document retrieval. In *Advances in Neural Information Processing Systems*.

Zora Zhiruo Wang, Akari Asai, Xinyan Velocity Yu, Frank F Xu, Yiqing Xie, Graham Neubig, and Daniel Fried. 2024c. Coderag-bench: Can retrieval augment code generation? *arXiv preprint arXiv:2406.14497*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Weaviate. 2024. What is agentic rag.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2025. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024a. Followir: Evaluating and teaching information retrieval models to follow instructions. *arXiv preprint arXiv:2403.15246*.

Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. 2024b. Promptriever: Instruction-trained retrievers can be prompted like language models. *arXiv preprint arXiv:2409.11136*.

Thomas D Wilson. 2000. Human information behavior. *Informing science*, 3:49.

SK Michael Wong and YY Yao. 1989. A probability distribution model for information retrieval. *Information Processing & Management*, 25(1):39–53.

Junde Wu, Jiayuan Zhu, and Yuyuan Liu. 2025. Agentic reasoning: Reasoning llms with tools for the deep research. *arXiv preprint arXiv:2502.04644*.

Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval*, 13:254–270.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024a. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, et al. 2024b. Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning. *arXiv preprint arXiv:2406.09187*.

Quanting Xie, So Yeon Min, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Russ Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. 2024. Embodied-RAG: General non-parametric embodied memory for retrieval and generation. In *Language Gamification - NeurIPS 2024 Workshop*.

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 55–64, New York, NY, USA. Association for Computing Machinery.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *ArXiv*, abs/2007.00808.

Jun Xu, Xiangnan He, and Hang Li. 2018. Deep learning for matching in search and recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1365–1368.

Shicheng Xu, Danyang Hou, Liang Pang, Jingcheng Deng, Jun Xu, Huawei Shen, and Xueqi Cheng. 2024a. Invisible relevance bias: Text-image retrieval models prefer ai-generated images. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 208–217, New York, NY, USA. Association for Computing Machinery.

Zhichao Xu. 2024. Rankmamba, benchmarking mamba's document ranking performance in the era of transformers. *arXiv preprint arXiv:2403.18276*.

Zhichao Xu, Aosong Feng, Yijun Tian, Haibo Ding, and Lin Leee Cheong. 2025a. Csplade: Learned sparse retrieval with causal language models. *arXiv preprint arXiv:2504.10816*.

Zhichao Xu, Hemank Lamba, Qingyao Ai, Joel Tetreault, and Alex Jaimes. 2024b. Cfe2: Counterfactual editing for search result explanation. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 145–155.

Zhichao Xu, Jinghua Yan, Ashim Gupta, and Vivek Srikumar. 2025b. State space models are strong text rerankers. In *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pages 152–169, Albuquerque, NM. Association for Computational Linguistics.

Zhichao Xu, Hansi Zeng, Juntao Tan, Zuohui Fu, Yongfeng Zhang, and Qingyao Ai. 2023. A reusable model-agnostic framework for faithfully explainable recommendation and system scrutability. *ACM Transactions on Information Systems*, 42(1):1–29.

Le Yan, Zhen Qin, Honglei Zhuang, Rolf Jagerman, Xuanhui Wang, Michael Bendersky, and Harrie Oosterhuis. 2024. Consolidating ranking and relevance predictions of large language models through post-processing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 410–423, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. 2024. Gated delta networks: Improving mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*.

Tao Yang, Zhichao Xu, and Qingyao Ai. 2023a. Vertical allocation-based fair exposure amortizing in ranking. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 234–244.

Tao Yang, Zhichao Xu, Zhenduo Wang, and Qingyao Ai. 2023b. Fara: Future-aware ranking algorithm for fairness optimization. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2906–2916.

Tao Yang, Zhichao Xu, Zhenduo Wang, Anh Tran, and Qingyao Ai. 2023c. Marginal-certainty-aware fair ranking algorithm. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 24–32.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, Zhichao Yang, and Hong Yu. 2024. Mcqg-srefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. *arXiv preprint arXiv:2410.13191*.

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy J. Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. In *Conference on Empirical Methods in Natural Language Processing*.

Soyoung Yoon, Eunbi Choi, Jiyeon Kim, Hyeongu Yun, Yireun Kim, and Seung-won Hwang. 2024. ListT5: Listwise reranking with fusion-in-decoder improves zero-shot retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2287–2308, Bangkok, Thailand. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Puxuan Yu and James Allan. 2020. A study of neural matching models for cross-lingual ir. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1637–1640.

Puxuan Yu, Antonio Mallia, and Matthias Petri. 2024a. Improved learned sparse retrieval with corpus-specific vocabularies. In *European Conference on Information Retrieval*, pages 181–194. Springer.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024b. Arctic-embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*.

Puxuan Yu, Razieh Rahimi, and James Allan. 2022. Towards explainable search results: a listwise explanation generator. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 669–680.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024c. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14672–14685, Miami, Florida, USA. Association for Computational Linguistics.

Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 497–506.

Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. 2024. Scalable and effective generative information retrieval. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 1441–1452, New York, NY, USA. Association for Computing Machinery.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.

ChengXiang Zhai et al. 2008. Statistical language models for information retrieval a critical review. *Foundations and Trends® in Information Retrieval*, 2(3):137–213.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Repbert: Contextualized text embeddings for first-stage retrieval. *arXiv preprint arXiv:2006.15498*.

Crystina Zhang, Minghan Li, and Jimmy Lin. 2024a. CELI: Simple yet effective approach to enhance out-of-domain generalization of cross-encoders. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 188–196, Mexico City, Mexico. Association for Computational Linguistics.

Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. 2023a. RepoCoder: Repository-level code completion through iterative retrieval and generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2484, Singapore. Association for Computational Linguistics.

Hanqi Zhang, Chong Chen, Lang Mei, Qi Liu, and Jiaxin Mao. 2024b. Mamba retriever: Utilizing mamba for effective and efficient dense retrieval. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4268–4272.

Xinyu Zhang, Sebastian Hofstätter, Patrick Lewis, Raphael Tang, and Jimmy Lin. 2023b. Rank-without-gpt: Building gpt-independent listwise rerankers on open-source large language models. *arXiv preprint arXiv:2312.02969*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023c. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Xinyu Zhang, Andrew Yates, and Jimmy Lin. 2021. Comparing score aggregation approaches for document retrieval with pretrained transformers. In *Advances in Information Retrieval*, pages 150–163, Cham. Springer International Publishing.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. 2024. Sglang: Efficient execution of structured language model programs. *Advances in Neural Information Processing Systems*, 37:62557–62583.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13764–13775, Singapore. Association for Computational Linguistics.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024a. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 358–370, Mexico City, Mexico. Association for Computational Linguistics.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023a. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2308–2313, New York, NY, USA. Association for Computing Machinery.

Shengyao Zhuang, Hang Li, and Guido Zuccon. 2021. Deep query likelihood model for information retrieval. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 463–470. Springer.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023b. Open-source large language models are strong zero-shot query likelihood models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8807–8817, Singapore. Association for Computational Linguistics.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024b. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 38–47, New York, NY, USA. Association for Computing Machinery.

A Supplement Materials on Traditional IR Models

Boolean Models The most basic Boolean model can be extended by incorporating term weighting, allowing both queries and documents to be represented as sets of weighted terms. Then, the logical implication $\mathcal{D} \rightarrow \mathcal{Q}$ is also weighted. The commonly used weighted approaches for the logical implication $\mathcal{D} \rightarrow \mathcal{Q}$ include using a fuzzy set extension of the Boolean logic (Radecki, 1979; Kraft and Buell, 1983) and p-norm (Salton et al., 1983).

Vector Space Model The weights q_i or d_i can be represented by other sophisticated schema, such as TF-IDF (Sparck Jones, 1972) and BM25 (Robertson et al., 1995). The extracted abundant features can improve the capacity and accuracy of the vector space models. Besides, given the vector representations of query $\mathcal{Q}$ and document $\mathcal{D}$, the most commonly used is cosine similarity, defined as

$$\text{sim}(\mathcal{Q}, \mathcal{D}) = \frac{\mathcal{Q} \cdot \mathcal{D}}{|\mathcal{Q}| \times |\mathcal{D}|},$$

where $\mathcal{Q} \cdot \mathcal{D}$ is the dot product and $|\mathcal{Q}|, |\mathcal{D}|$ denotes the length of the vector.

Probabilistic Model In Probabilistic Model, the relevance score of a document $\mathcal{D}$ to a query $\mathcal{Q}$ depends on a set of events $\{x_i\}_1^n$ representing the occurrence of term t_i in this document. The simplest probabilistic model is the binary independence retrieval model (Robertson and Jones, 1976), which assumes terms are independent so only $x_i = 1$ and $x_i = 0$ exist in the representation. Given a set of sample documents whose relevance is judged, the estimation of the relevance score can be derived as $\text{Score}(\mathcal{Q}, \mathcal{D}) \propto \sum_{(x_i=1)\in\mathcal{D}} \log \frac{r_i(T-n_i-R+r_i)}{(R-r_i)(n_i-r_i)}$, where T and R are the total number of sampled judged documents and relevant samples, and n_i and r_i denote the number of samples and relevant samples containing t_i, respectively. The smooth mechanisms (Baeza-Yates et al., 1999) are necessary to deal with zero occurrences of the t_i.

Except for the binary independence retrieval model, more sophisticated probabilistic models are proposed in the literature (Wong and Yao, 1989; Fuhr, 1992), such as the inter-dependency between terms (Van Rijsbergen, 1979).

Statistical Language Model The general idea of a statistical language model is to estimate the relevance score of a document $\mathcal{D}$ to a query $\mathcal{Q}$ via $\mathcal{P}(\mathcal{D}|\mathcal{Q})$ (Ponte and Croft, 1998). Based on Bayes Rule, $\mathcal{P}(\mathcal{D}|\mathcal{Q})$ can be derived as directly proportional to $\mathcal{P}(\mathcal{Q}|\mathcal{D})\mathcal{P}(\mathcal{D})$. For simplification, most studies assume a uniform distribution for $\mathcal{P}(\mathcal{D})$. The main focus is on modeling $\mathcal{P}(\mathcal{Q}|\mathcal{D})$ as a ranking function by treating the query as a set of independent terms as $\mathcal{Q} = \{t_i\}_{i=1}^n$, thus $\mathcal{P}(\mathcal{Q}|\mathcal{D}) = \prod_{t_i\in\mathcal{Q}} \mathcal{P}(t_i|\mathcal{D})$. The probability $\mathcal{P}(t_i|\mathcal{D})$ is determined using a statistical language model θ_D that represents the document, then the relevance is estimated by log-likelihood as $\text{Score}(\mathcal{Q}, \mathcal{D}) = \log \mathcal{P}(\mathcal{Q}|\theta_D) = \sum_{t_i\in\mathcal{Q}} \log \mathcal{P}(t_i|\theta_D)$, where the estimation of the language model θ_D is usually achieved by maximum likelihood.

The statistic language models for IR (Miller et al., 1999; Berger and Lafferty, 1999; Song and Croft, 1999) also encounter the problem of the zero occurrences of the query term t_i, i.e., the probability $\mathcal{P}(\mathcal{Q}|\theta_D)$ becomes zero, if a query term t_i does not appear in the document. This is too restrictive for IR, as a document can still be relevant even if it contains only some of the query terms. To address this zero-probability issue, smoothing techniques are applied, assigning small probabilities to terms that do not appear in the document. The principle behind smoothing is that any text used to model a language captures only a limited subset of its linguistic patterns (or terms, in this case). The commonly used smoothing methods (Zhai and Lafferty, 2004; Zhai et al., 2008) include Jalinek-Mercer smoothing, Dirichlet smoothing, etc.

B Supplement Materials on Learning-to-Rank Architecture and Training Strategy

We present a list of learning-to-rank works and their backbone architectures in Table 1. A significant portion of the literature focuses on loss functions and feature transformers (Qin et al., 2021; Bruch et al.,

Name	Model	Backbone Architecture	Loss Function
MART (Friedman, 2001)	ML	Boosting	Pointwise
RANKBOOST (Freund et al., 2003)	ML	Boosting	Pairwise
RANKNET (Burges et al., 2005)	Neural Nets	DNN	Pairwise
RANKSVM (Joachims, 2006)	ML	SVM	Pairwise
LAMBDARANK (Burges et al., 2006)	Neural Nets	DNN	Pairwise
LISTNET (Cao et al., 2007)	Neural Nets	DNN	Listwise
SOFTRANK (Taylor et al., 2008)	Neural Nets	DNN	Listwise
LISTMLE (Xia et al., 2008)	ML	Linear	Listwise
LAMBDAMART (Burges, 2010)	ML	GBDT	Listwise
APPROXNDCG (Qin et al., 2010)	ML	Linear	Listwise
DLCM (Ai et al., 2018a)	Neural Nets	DNN	Listwise
GSF (Ai et al., 2019)	Neural Nets	DNN	Listwise
APPROXNDCG (Bruch et al., 2019)	Neural Nets	DNN	Listwise
SETRANK (Pang et al., 2020)	Neural Nets	Self Attention Blocks	Listwise

Table 1: A list of learning-to-rank works and their model architectures.

2019; Burges, 2010). Additionally, some studies focus on unbiased relevance estimation using biased feedback (Wang et al., 2016; Joachims et al., 2017b,a; Ai et al., 2018c,b; Wang et al., 2018; Hu et al., 2019; Ren et al., 2022) while other focus on jointly optimizing effectiveness and fairness of the ranking systems (Singh and Joachims, 2018; Biega et al., 2018; Morik et al., 2020; Patro et al., 2020; Oosterhuis, 2021; Yang et al., 2023a,c,b). We omit detailed discussions here and refer readers to the original papers.

C Supplement Materials on Neural Ranking Models

Representation-based Models Representation-based neural ranking models can be regarded as extensions of vector space models (§ 3), which independently encode queries and documents into a latent vector space. The relevance ranking of a document is determined by computing the similarity (e.g., cosine similarity) between the query and document embeddings.

The Deep Structured Semantic Model (DSSM) (Huang et al., 2013) is an early example of a representation-based neural ranking model. It utilizes word hashing and multilayer perceptrons (MLPs) to independently encode term vectors of queries and documents into a shared semantic space, enabling the computation of ranking scores based on the cosine similarity of their embeddings. Research has focused on enhancing DSSM by modifying its encoder network to improve the model's ability to capture richer semantic and contextual information. For instance, Convolutional DSSM (Shen et al., 2014) leverages a CNN architecture to project vectors within a context window to a local contextual feature vector. These local features are then aggregated using a max-pooling layer to produce a representation of the entire query or document. Another variation of DSSM replaces MLPs with a Long Short-Term Memory (LSTM) network (Palangi et al., 2016; Wan et al., 2016) . By leveraging LSTM's memory mechanism, such models can capture both local and global context information without the pooling layer, thus better suited for handling longer documents.

Interaction-based Models Interaction-based models process queries and documents jointly through neural networks. The model's output is typically a score that measures the relevance of the input query-document pair. Various network architectures have been proposed to jointly encode queries and documents. For instance, MATCHPYRAMID (Pang et al., 2016) employs CNN over the interaction matrix between query and document terms. This approach treats the interaction matrix as an image, allowing the CNN to capture local matching patterns. The model then aggregates these patterns through convolution and pooling operations to produce a relevance score, effectively modeling the hierarchical matching structures between queries and documents (Hu et al., 2014). Building upon the concept of interaction-focused models, Guo et al. (2016a) highlighted the importance of exact term matches in neural ranking models and proposed the Deep Relevance Matching Model (DRMM). The model constructs matching histograms for each query term to capture the distribution of matching signals across

Name	Architecture	Backbone	Embeddings
DSSM (Huang et al., 2013)	Representation-based	MLP	Word Hashing
CDSSM (Shen et al., 2014)	Representation-based	CNN	Word Hashing
ARC-I (Hu et al., 2014)	Representation-based	CNN	Word2Vec
ARC-II (Hu et al., 2014)	Interaction-based	CNN	Word2Vec
MatchPyramid (Pang et al., 2016)	Interaction-based	CNN	Randomly Initialized
LSTM-RNN (Palangi et al., 2016)	Representation-based	LSTM	Randomly Initialized
MV-LSTM (Wan et al., 2016)	Representation-based	Bi-LSTM	Word2Vec
DRMM (Guo et al., 2016a)	Interaction-based	MLP	Word2Vec
DESM (Nalisnick et al., 2016)	Interaction-based	MLP	Word2Vec
K-NRM (Xiong et al., 2017)	Interaction-based	MLP + RBF kernels	Word2Vec
Conv-KNRM (Dai et al., 2018)	Interaction-based	CNN	Word2Vec
TK (Hofstätter et al., 2020c)	Interaction-based	Transformer + Kernel	GloVe
TKL (Hofstätter et al., 2020a)	Interaction-based	Transformer + Kernel	GloVe
NDRM (Mitra et al., 2021)	Interaction-based	Conformer + Kernel	BERT

Table 2: A list of neural ranking models and their model architectures.

document terms. These histograms are then processed through a feed-forward neural network to learn hierarchical matching patterns. Xiong et al. (2017) introduced the Kernel-Based Neural Ranking Model (K-NRM), which further advanced interaction-based approaches. K-NRM employs a translation matrix to compute interactions between query and document terms based on their embeddings. It then applies Radial Basis Function (RBF) kernels to transform these word-level interactions into informative ranking features. Later, they extended the RBF kernel approach to a convolutional neural network (Dai et al., 2018).

Word Embeddings In addition to advancements in network architecture, pre-trained textual representations have also contributed to neural ranking models' performance (Guo et al., 2016b). GloVe (Pennington et al., 2014) and Word2Vec (Mikolov, 2013) learn dense vector representations for each vocabulary term from large-scale text corpora. Pre-trained embeddings provide semantic-based term representations to enable neural ranking models to focus on learning relevance matching patterns. Both representation-based and interaction-based models adopt pre-trained word embeddings as input representations to their networks, facilitating training convergence and improved performance (Levy et al., 2015). Interaction-based models with cross-lingual word embeddings (Joulin et al., 2018) for cross-lingual reranking have also been explored (Yu and Allan, 2020).

Table 2 shows a list of neural ranking models and backbone architectures. Researchers have explored different backbone neural network architectures in this era, including Convolutional Neural Network (CNN, LeCun et al., 1989, 1998), Long Short Term Memory (LSTM, Hochreiter and Schmidhuber, 1997) and kernel methods (Vert et al., 2004; Chang et al., 2010; Xiong et al., 2017).

Notably, a line of research explores integrating kernel methods with the TRANSFORMER architecture (Vaswani et al., 2017). The main distinction between this line of research and the models discussed in § 6 is that the transformer modules here are not pre-trained on large-scale corpora like Wikipedia and C4 (Devlin et al., 2019; Raffel et al., 2020). We consider this line of research as an intersection between neural ranking models (§ 5) and retrieval with pre-trained transformers (§ 6). TK (Hofstätter et al., 2020c) uses a shallow transformer neural network (up to 3 layers) to encode the query $\mathcal{Q}$ and document $\mathcal{D}$ separately. After encoding, the contextualized representations are input to one single interaction match matrix, similar to model architecture shown in Fig. 2b. The entire model is trained end-to-end and is able to achieve better performance-efficiency trade-off compared to BERT-based reranker (Nogueira et al., 2019). The main bottleneck of applying transformer architectures to long document reranking is $O(n^2)$ time complexity, where n denotes the document length. TKL (Hofstätter et al., 2020b) further improves upon TK with a local attention mechanism and leads to performance improvement on long document ranking.

D Supplement Materials on Pre-trained Language Models for Information Retrieval

We show a list of models and their corresponding architectures in Table 3, including reranking models, learned dense retrieval, multi-vector representations and learned sparse retrieval. A majority of the models use BERT (Devlin et al., 2019) as the backbone language models, with a few exceptions using DISTILBERT (Sanh, 2019), ROBERTA (Liu, 2019) and encoder part of T5 family models (Raffel et al., 2020; Sanh et al., 2022; Mo et al., 2023; Chung et al., 2024).

One line of work aims to combine the benefits of learned dense retrieval and sparse retrieval. (Gao et al., 2021b,a; Ma et al., 2021; Lin and Ma, 2021; Cormack et al., 2009). Ranklist fusion techniques (e.g., Reciprocal Rank Fusion, Cormack et al., 2009) directs fuses ranklists from different retrievers and has been shown to improve retrieval performance. COIL (Gao et al., 2021a) proposes to enhance traditional bag-of-words retrieval method with semantic embeddings from BERT encoder. UNICOIL (Lin and Ma, 2021) further simplifies reduces the dimension of semantic embeddings to 1 — equivalent to learned term weight in learned sparse retrieval models like SPLADE (Formal et al., 2021b,a).

A few works fall into the intersection of learned sparse retrieval and multi-vector representations. For example, SLIM (Li et al., 2023c) first maps each contextualized token vector to a sparse, high-dimensional lexical space before performing late interaction between these sparse token embeddings. SPLATE (Formal et al., 2024) take an alternative approach to first encodes contextualized token vectors, then maps these token vectors to a sparse vocabulary space with a partially learned SPLADE module. Both models achieve performance improvement compared to learned sparse retrieval baselines such as SPLADE (Formal et al., 2021b,a).

Instead of improving retrieval performance from the modeling perspective, a separate line of works aim to enhance the backbone language models via domain adaptation or continued pre-training, which has been proven successful by prior works in NLP (Howard and Ruder, 2018; Gururangan et al., 2020). Lee et al. (2019) propose to pre-train BERT model with Inverse-Cloze Task (Taylor, 1953) for better text representations. CONDENSER (Gao and Callan, 2021) propose to "condense" text representations into [CLS] token via a dedicated pre-training architecture and corresponding training objective. COCO-DR further extends upon CONDENSER via a technique named implicit Distributionally Robust Optimization to mitigate distribution shift problem in dense retrieval. We refer readers to original papers for details.

As we noted in § 9, one desiderata of future IR models is interpretability and truthfulness. A few works have attempted to interpret transformer-based neural retrieval models' representations, i.e., mechanistic interpretability (Elhage et al., 2021; Saphra and Wiegreffe, 2024). For example, MacAvaney et al. (2022) showed that neural retrieval models rely less on exact match signals and instead encodes rich semantic information. Ram et al. (2023) project dense retrievers' intermediate representations to vocabulary space and show the connection of dense retrieval and traditional bag-of-words sparse retrieval methods. Instead of providing model-intrinsic explanations, a few works design IR systems to provide model-agnostic explanations (Rahimi et al., 2021; Yu et al., 2022; Xu et al., 2024b) in order to meet certain desiderata such as faithfulness (Jacovi and Goldberg, 2020; Xu et al., 2023). As IR systems become an integral part of other applied ML domains, we believe it is important to study and design interpretable, truthful and trustworthiness IR models.

E Supplement Materials on LLM for IR

We summarize a list of works that study LLM for retrieval (Table 4) and reranking (Table 5). For generative retrieval, we point to a dedicated survey (Li et al., 2025c). Another comprehensive survey (Mo et al., 2024b) could be referred to for conversational information retrieval. Modern IR systems require extensive labeled data to achieve good performance. One line of work studies the proposal of using LLMs for synthesizing training data (Bonifacio et al., 2022; Boytsov et al., 2024; Dai et al., 2023; Lee et al., 2024; Mo et al., 2024a,c). From the evaluation perspective, LLMs' superior natural language understanding capability also raise the question of whether they can be used for relevance judgments. A separate line of work tackle the relevance judgments problem (Faggioli et al., 2023, 2024; Clarke and Dietz, 2024). As our focus of this survey is on model architectures, we skip the discussion and point to original papers for further details.

Name	Model	Architecture	Backbone LM	Training strategy
monoBERT (Nogueira et al., 2019)	Reranking	Cross-encoder	BERT	Classification
CEDR (MacAvaney et al., 2019)	Reranking	Cross-encoder	BERT	Contrastive Learning
BERT-MaxP (Dai and Callan, 2019b)	Reranking	Cross-encoder	BERT	Pairwise Loss
Gao et al. (2020)	Reranking	Cross-encoder	BERT	Distillation
TART-full (Asai et al., 2023)	Reranking	Cross-encoder	Flan-T5-Enc	Instruction Tuning
DPR (Karpukhin et al., 2020)	LDR	Bi-encoder	BERT	Contrastive Learning
ANCE (Xiong et al., 2020)	LDR	Bi-encoder	RoBERTa	Contrastive Learning
RepBERT (Zhan et al., 2020)	LDR	Bi-encoder	BERT	In-batch negatives
Margin-MSE (Hofstätter et al., 2020a)	LDR	Bi-encoder	DistilBERT	Distillation
TAS-B (Hofstätter et al., 2021)	LDR	Bi-encoder	BERT	Distillation
RocketQA (Qu et al., 2020)	LDR	Bi-encoder	ERNIE	Contrastive Learning
RocketQA-v2 (Ren et al., 2021)	LDR	Bi-encoder	ERNIE	Distillation
GTR (Ni et al., 2022b)	LDR	Bi-encoder	EncT5	Contrastive Learning
TART-dual (Asai et al., 2023)	LDR	Bi-encoder	Contriever	Instruction Tuning
E5 (Wang et al., 2022a)	LDR	Bi-encoder	BERT	Contrastive Learning
GTE (Li et al., 2023e)	LDR	Bi-encoder	BERT	Contrastive Learning
Poly-encoder (Humeau et al., 2020)	Multi-vector	Misc	BERT	In-batch Negatives
ME-BERT (Luan et al., 2021)	Multi-vector	Bi-encoder	BERT	Contrastive Learning
ColBERT (Khattab and Zaharia, 2020)	Multi-vector	Bi-encoder	BERT	Pairwise Loss
COIL (Gao et al., 2021a)	Multi-vector	Bi-encoder	BERT	Contrastive Learning
ColBERT-v2 (Santhanam et al., 2022)	Multi-vector	Bi-encoder	BERT	Distillation
ColBERTer (Hofstätter et al., 2022)	Multi-vector	Bi-encoder	BERT	Distillation
DeepCT (Dai and Callan, 2019a)	LSR	Bi-encoder	BERT	Unsupervised
SparTerm (Bai et al., 2020)	LSR	Bi-encoder	BERT	Contrastive Learning
SPLADE (Formal et al., 2021b)	LSR	Bi-encoder	BERT	Contrastive Learning
SPLADE-v2 (Formal et al., 2021a)	LSR	Bi-encoder	BERT	Distillation
DeepImpact (Mallia et al., 2021)	LSR	Bi-encoder	BERT	Contrastive Learning
SparseEmbed (Kong et al., 2023)	LSR	Bi-encoder	BERT	Contrastive Learning
SLIM (Li et al., 2023c)	LSR + Multi-vector	Bi-encoder	BERT	Contrastive Learning
SLIM++ (Li et al., 2023c)	LSR + Multi-vector	Bi-encoder	BERT	Distillation
SPLATE (Formal et al., 2024)	LSR + Multi-vector	Bi-encoder	BERT	Distillation

Table 3: Summary of IR model architecture for passage retrieval and passage ranking based on pre-trained transformers. LDR and LSR denote learned dense retrieval and learned sparse retrieval, respectively. DeepCT (Dai and Callan, 2019a) is trained without labeled training set. The "late interaction" mechanism introduced in (Khattab and Zaharia, 2020; Santhanam et al., 2022) can be considered a special case of multi-vector architecture. Contrastive Learning and in-batch negatives means listwise loss function is used.

Name	Architecture	Backbone LM	Training strategy
CPT-Text (Neelakantan et al., 2022)	LLM Encoder	GPT-3	Listwise Loss
SGPT-BE (Muennighoff, 2022)	LLM Encoder	GPT-J & GPT-NeoX	Listwise Loss
GTR (Ni et al., 2022b)	LLM Encoder	T5	Listwise Loss
RepLLaMA (Ma et al., 2024b)	LLM Encoder	LLaMA	Listwise Loss
E5-Mistral (Wang et al., 2023)	LLM Encoder	Mistral	Synthetic Data + Listwise Loss
LLaRA (Li et al., 2023a)	LLM Encoder	LLaMA	Adaptation + Contrastive Training
MambaRetriever (Zhang et al., 2024b)	LLM Encoder	Mamba	Listwise Loss
LLM2Vec (BehnamGhader et al., 2024)	LLM Encoder	LLaMA & Mistral	Adaptation + Contrastive Pre-training
Grit-LM (Muennighoff et al., 2025)	LLM	Mistral & Mixtral 8x7B	Generative/Embedding Joint Training
NVEmbed (Lee et al., 2025)	LLM Encoder	Mistral	Adaptation + Synthetic Data + Listwise Loss
GTE-Qwen2-Instruct (Li et al., 2023e)	LLM Encoder	Qwen	Adaptation + Synthetic Data + Listwise Loss

Table 4: Summary of IR model architecture utilizing large language models as retrieval backbone.

F Extended Discussions on Challenges and New Directions

F.1 Autonomous Search Agents

We discuss recent progress on developing autonomous agents for search and information seeking purposes. While these works do not focus on improving IR models *per se*, we believe it is important for IR researchers to adapt to these new use cases of search/retrieval.

Prior works have studied methods to augment language models with search/retrieval to improve generation quality, which we term as retrieval-augmented generation. Early practices include KNN-

Name	Architecture	Backbone LM	Training / Prompting Strategy
Fine-tune LLM for Reranking			
monoT5 (Nogueira et al., 2020)	LM	T5	Classification
Nogueira dos Santos et al. (2020)	LM	BART	Unlikelihood
QLM-T5 (Zhuang et al., 2021)	LM	T5	Language Modeling
duoT5 (Pradeep et al., 2021)	LM	T5	Pairwise Loss
RankT5 (Zhuang et al., 2023a)	LLM Encoder + Prediction Layer	T5	Listwise Loss
ListT5 (Yoon et al., 2024)	Fusion-in-decoder	T5	Listwise Loss
SGPT-CE (Muennighoff, 2022)	LM	GPT-J & GPT-Neo	Listwise Loss
RankLLaMA (Ma et al., 2024b)	LLM Encoder + Prediction Layer	LLaMA	Listwise Loss
RankMamba (Xu, 2024)	LLM Encoder + Prediction Layer	Mamba	Listwise Loss
RankVicuna (Pradeep et al., 2023b)	LM	Vicuna	Listwise
RankZephyr (Pradeep et al., 2023c)	LM	Zephyr	Listwise
Zhang et al. (2023b)	LM	Code-LLaMA-Instruct	Listwise
Liu et al. (2024b)	Embedding + LM	Mistral	Listwise
Prompt LLM for Reranking			
Zhuang et al. (2023b)	LM	Multiple	Pointwise Prompting
Zhuang et al. (2024a)	LM	Flan-PaLM-S	Pointwise Prompting
UPR (Sachan et al., 2022)	LM	T5 & GPT-Neo	Pointwise Prompting
PRP (Qin et al., 2024a)	LM	Flan-UL2	Pairwise Prompting
Yan et al. (2024)	LM	Flan-UL2	Pairwise Prompting
Zhuang et al. (2024b)	LM	Flan-T5	Pairwise & Setwise Prompting
LRL (Ma et al., 2023)	LM	GPT-3	Listwise Prompting
RankGPT-3.5 (Sun et al., 2023)	LM	GPT-3.5	Listwise Prompting
RankGPT-4 (Sun et al., 2023)	LM	GPT-4	Listwise Prompting

Table 5: Summary of IR model architecture utilizing large language models for reranking. Nogueira dos Santos et al. (2020) and Zhuang et al. (2021) revisit the statistic language model problem with modern transformer-based models, including BART (Lewis et al., 2020a) T5 (Raffel et al., 2020) and GPT-2 (Radford et al., 2019).

LM (Khandelwal et al., 2020b), REALM (Guu et al., 2020), RAG (Lewis et al., 2020b), *inter alia*. With more powerful models such as ChatGPT (OpenAI, 2022), researchers begin to design systems to handle daily tasks autonomously with LLM backbones. We refer to such systems as LLM-based agents (Wang et al., 2024b; Guo et al., 2024). WebGPT (Nakano et al., 2021) leverages reinforcement learning to train GPT-3-based language models for web browsing, which is one of the earliest works along this direction. Due to the instrumental role of retrieval in tasks solving, popular agent frameworks (Wu et al., 2023; Li et al., 2023b) have supported built-in retrieval functionality (commonly referred to as *agentic memory* in agent literature). Most of existing general purpose agent frameworks treat retrieval as one of the available tools, and use LLMs to plan and orchestrate workflows accordingly for task completion, with techniques such as self-refine (Madaan et al., 2023), reflexion (Shinn et al., 2023) and critique (Gou et al., 2023).

Earlier works by the IR and NLP community – such as FLARE (Jiang et al., 2023) and Self-RAG (Asai et al., 2024a) – have proposed methods to build autonomous search systems, i.e., to enable the system to know when, where and how to search. FLARE (Jiang et al., 2023) explored using prompting method while Self-RAG (Asai et al., 2024a) focuses on data synthesis and supervised fine-tuning. Popularized by large reasoning models such as GPT-4o (Hurst et al., 2024) and DeepSeek-R1 (Guo et al., 2025), there is a surge of recent works aiming to incorporate retrieval to augment LLM reasoning, or to train large reasoning models to use search tool for better performance with reinforcement learning techniques (Li et al., 2025a; Jin et al., 2025; Li et al., 2025b; Chen et al., 2025; Zheng et al., 2025; Guan et al., 2025; Song et al., 2025; Wu et al., 2025; Hu et al., 2025; Wang et al., 2025; Gao et al., 2025). As these works mainly focus on optimizing the generator component of RAG systems, we refer the readers to these individual works for further details.

Existing agentic RAG methods, including single agent RAG systems or multi-agent systems (Nachimovsky et al., 2025; Chang et al., 2024; Weaviate, 2024) treat retriever as a static component of the system, and focus on improving the generator via prompting or model optimization. While these methods do not directly propose new IR modeling or training strategies, we believe it is critical for IR researchers to contextualize common agentic use cases and propose new IR model architectures better suited for these application scenarios.

F.2 Deployment of Modern IR Systems

Efficiency and Effectiveness Tradeoff Traditional retrieval systems face significant challenges when scaling to web-scale document corpus, and to deploy such systems requires a blend of science and engineering expertises (Dean et al., 2009; Huang et al., 2020; Li et al., 2021). In recent years, retrieval-augmented generation, conversational search and agentic systems with memory have been widely adopted for information access (Guu et al., 2020; Lewis et al., 2020b; Google, 2019; OpenAI, 2024; Google, 2024, *inter alia*). These applications often require multiple rounds of retrieval and dynamic corpus, urging for efficient and effective retrieval. Mainstream inference optimization frameworks such as vLLM (Kwon et al., 2023) and SGLang (Zheng et al., 2024) have provided support for embedding models. From the modeling perspective, an open question is to design and pre-train models for retrieval purposes (Warner et al., 2024; Nussbaum et al., 2025; Günther et al., 2023).

Robustness in Noisy Environment We discuss a few challenges in IR models' deployment in noisy environment, especially when used in retrieval-augmented generation systems. We should note that while these challenges have been studied by prior works, it remains an open question on how to mitigate these challenges from the perspective of IR modeling and architectures.

- **Robustness to AI generated content.** With the advent of LLMs, the amount of AI-generated content is also increasing. Dai et al. (2024) show that neural retrievers are biased towards AI-generated documents. Xu et al. (2024a) show that similar problems persist in text-image retrieval models. Future IR modeling research should also consider the robustness of models to AI-generated content.

- **Robustness to adversarial attacks.** Recent works on RAG LLM safety have discussed the threat of corpus poisoning where injected harmful documents lead to unsafe outputs (Zhong et al., 2023; Xiang et al., 2024a; Deng et al., 2024, *inter alia*). This topic is also relevant to the safety of LLM agents using tools (Deng et al., 2025; Tian et al., 2023; Xiang et al., 2024b), noting the importance of IR models being robust to adversarial attacks for downstream applications.

- **Robustness to bias and toxicity.** As noted by a recent work (An et al., 2025), documents that contains biases and toxic materials can potentially jailbreak aligned LLMs. This observation highlights the importance for IR models to be robust to bias and toxic contents.

- **Robustness to imperfect retrieval results.** Different works have pointed out that existing RAG systems show performance degradation when the retrieval results contain irrelevant documents (Yoran et al., 2024; Chang et al., 2024; Yu et al., 2024c, *inter alia*). Therefore, the RAG paradigm demands more precise results from the retrieval models.

- **Robustness to out-of-distribution input.** Given the fact that modern neural retrieval models are trained with data-driven approaches, perhaps it is not surprising to find their performance may vary with different linguistic properties of the queries and documents, i.e., out-of-distribution input from the training data. Cao et al. (2025) conduct a rigorous benchmarking, and find formality, readability, politeness and grammatical correctness – fundamental aspects of real-world user-LLM queries – can lead to significant performance variances of retrievers and RAG systems. This observation highlights the importance of retrieval models' robustness to OOD input (Gupta et al., 2024).

We refer readers for more detailed discussions on IR models' robustness to dedicated surveys (Asai et al., 2024b; Liu et al., 2025; Zhou et al., 2024).