A Survey of Model Architectures in Information Retrieval

Zhichao Xu *zhichao.xu@utah.edu*
Kahlert School of Computing
University of Utah

Fengran Mo *fengran.mo@umontreal.ca*
University of Montreal

Zhiqi Huang *zhiqi.huang@capitalone.com*
Capital One

Crystina Zhang *xinyucrystina.zhang@uwaterloo.ca*
University of Waterloo

Puxuan Yu *puxuan.yu@snowflake.com*
Snowflake Inc.

Bei Wang *beiwang@sci.utah.edu*
Kahlert School of Computing
Scientific Computing and Imaging (SCI) Institute
University of Utah

Jimmy Lin *jimmylin@uwaterloo.ca*
University of Waterloo

Vivek Srikumar *svivek@cs.utah.edu*
Kahlert School of Computing
University of Utah

Reviewed on OpenReview: *https://openreview.net/forum?id=xAIbTbHRrX*

Abstract

The period from 2019 to the present marks one of the most significant paradigm shifts in information retrieval (IR) and natural language processing (NLP), culminating in the emergence of powerful large language models (LLMs) from 2022 onward. Methods based on pretrained encoder-only architectures (e.g., BERT) as well as decoder-only generative LLMs have outperformed many earlier approaches, demonstrating particularly strong performance in zero-shot scenarios and complex reasoning tasks. This survey examines the evolution of model architectures in IR, with a focus on two key aspects: backbone models for feature extraction and end-to-end system architectures for relevance estimation. To maintain analytical clarity, we deliberately separate architectural design from training methodologies, enabling a focused examination of structural innovations in IR systems. We trace the progression from traditional term-based retrieval models to modern neural approaches, highlighting the transformative impact of transformer-based architectures and subsequent LLM developments. The survey concludes with a forward-looking discussion of open challenges and emerging research directions, including architectural optimization for efficiency and scalability, robust handling of multimodal and multilingual data, and adaptation to novel application domains such as autonomous search agents, which may represent the next paradigm in IR.

1 Introduction

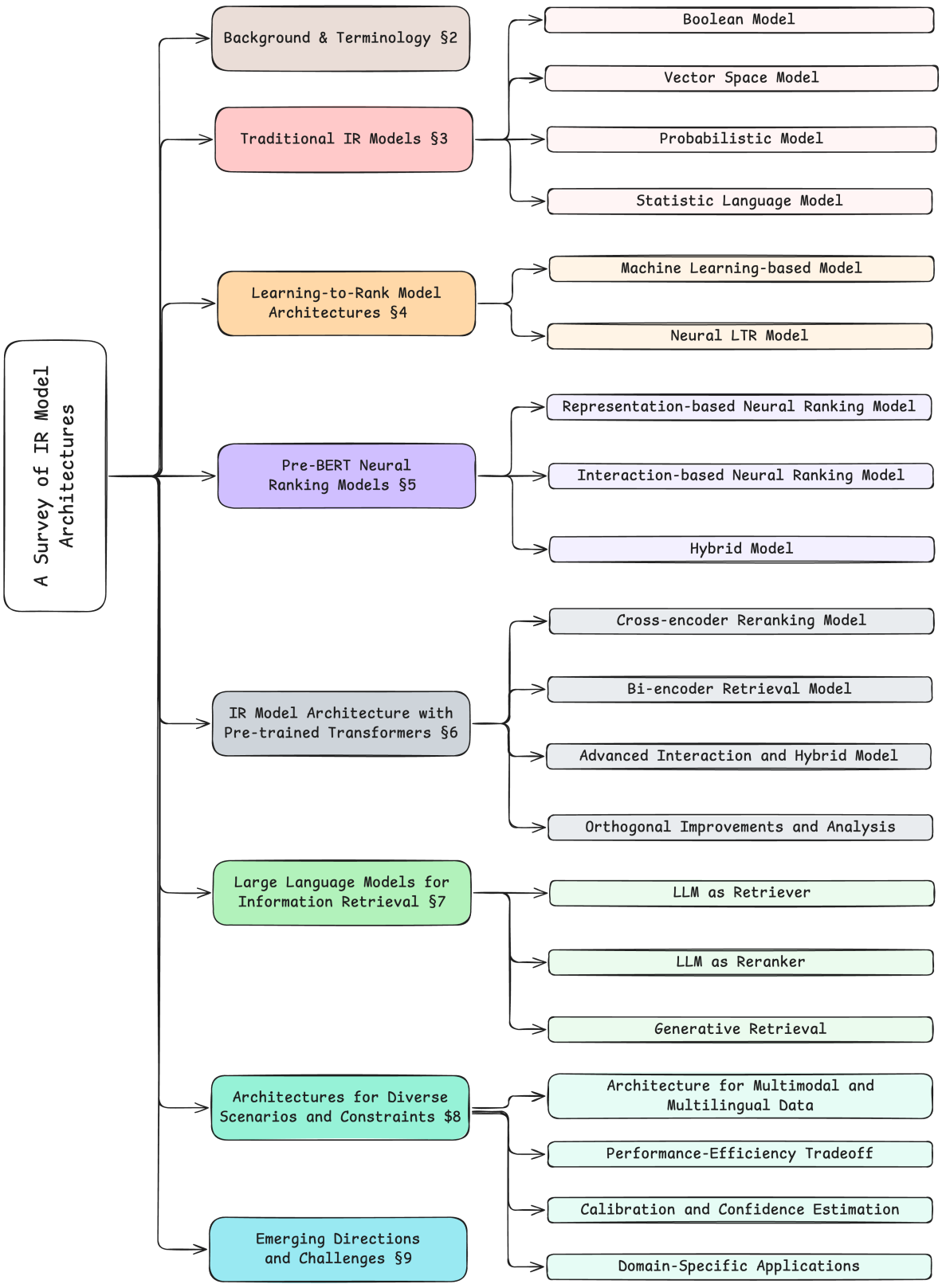


Figure 1: High-level overview of the survey structure and scope.

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 under the generative AI era. Research in IR primarily centers on two key aspects: (1) *extracting better query and document feature representations*, and (2) *developing more accurate relevance estimators*. Extracting better query and document feature representations focuses on modeling textual content so that queries and documents can be compared in a meaningful space, ranging from early term-frequency vectors (e.g., TF–IDF (Sparck Jones, 1972) and BM25 (Robertson & Zaragoza, 2009)) to modern contextual embeddings derived from pre-trained language models. Developing more accurate relevance estimators then builds on these representations to assess how well a document satisfies an information need, using scoring functions or learned ranking models such as BM25, neural interaction models, or later learning-to-rank frameworks that combine multiple signals. The approaches for extracting query and document features have evolved from traditional term-based methods, such as boolean logic (Radecki, 1979; Kraft & Buell, 1983) and vector space models (Salton et al., 1975), to modern solutions such as dense retrieval based on pre-trained language models (Lee et al., 2019; Karpukhin et al., 2020; Logeswaran et al., 2019; Lin et al., 2022, *inter alia*).

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 like support vector machines (Cortes & Vapnik, 1995), boosting methods (Kearns & Valiant, 1994; Freund & Schapire, 1995) as well as 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 & Liu, 2013). In the 2010s, a vast literature explored neural rerankers in different architectures to capture the semantic similarity between queries and documents (Pang et al., 2016; Guo et al., 2016a; Xiong et al., 2017; Dai et al., 2018, *inter alia*). Then pre-trained transformers, represented by BERT (Devlin et al., 2019) and its variants (Liu et al., 2019; Sun et al., 2019; Lan et al., 2020; Beltagy et al., 2020), 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 prediction heads, which operate on learned neural representations (MacAvaney et al., 2019b; Lee et al., 2019; Karpukhin et al., 2020; Nogueira et al., 2020; Lin et al., 2022; Formal et al., 2021b;a).

Recent advancements in LLMs have revolutionized applied machine learning (ML) communities, including IR. One intriguing property of modern instruction-following LLMs (e.g., ChatGPT OpenAI (2022)) 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., 2024b, *inter alia*). The rise of these models builds upon a rich foundation of neural architectures, including the classical Transformer architecture with multi-head attention (MHA, Vaswani et al., 2017), Recurrent Neural Networks (RNN, Elman, 1990), attention mechanisms (Bahdanau et al., 2014), and pre-trained static word representations like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) (Collobert et al., 2011; Le & Mikolov, 2014, *inter alia*).

This work reviews the evolution of model architectures in IR (with an overview in Figure 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 deliver a focused architectural analysis. This analysis centers on the core components of AI infrastructure in the LLM era. The shift towards neural architectures, particularly Transformer-based models, has fundamentally transformed IR by enabling rich, contextualized representations and improved capacity for handling complex queries. While this evolution enhanced retrieval performance, it also presents new challenges, especially with the development of LLMs. These challenges include the need for architectural innovations to optimize performance and scalability, handle multimodal and multilingual data, and understand complex instructions. Moreover, as IR systems are increasingly integrated into diverse applica-

tions — from robotics (Xie et al., 2024b), protein structure discovery (Jumper et al., 2021) to autonomous agents (Wu et al., 2023a; Chen et al., 2025; Hu et al., 2025; OpenAI, 2025; Wu et al., 2025b, *inter alia*) that are capable of reasoning and search — the field must evolve beyond traditional search paradigms. We conclude this survey by examining these challenges and discussing their implications for the future of IR model architectures research.

2 Background and Terminology

We focus on the classical *ad hoc* retrieval task, which forms the foundation for many modern IR applications. In this section, we define the core task, introduce key system architectures and evaluation paradigms, and clarify the scope of our architectural review.

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, list-wise IR metrics. Common metrics include:

- **Mean Reciprocal Rank (MRR):** Measures the rank of the first relevant document. It is particularly useful for tasks where finding one correct answer is the primary goal (e.g., question answering).
- **Recall@k:** Measures the fraction of all relevant documents that are found within the top-k results. It emphasizes the system's ability to find all relevant items.
- **Normalized Discounted Cumulative Gain (nDCG@k):** A sophisticated metric that evaluates the quality of the ranking over the top-k documents. It gives higher scores for ranking highly relevant documents at the top of the list and uses a logarithmic discount to penalize relevant documents that appear lower in the ranking. It is the de facto standard for evaluating ranked lists with graded relevance judgments.

The Multi-Stage "Retrieve-then-Rerank" Architecture. Modern large-scale IR systems almost universally operate on a multi-stage pipeline, commonly known as the "retrieve-then-rerank" architecture. This design balances the tradeoff between efficiency and effectiveness.

1. **Retrieval (or First-Stage Ranking):** In the first stage, a computationally efficient but less precise model scans the entire collection $\mathcal{C}$ (potentially billions of documents) to quickly identify an initial set of several hundred or thousand candidate documents. These models, often called *retrievers*, must be extremely fast. Examples include traditional models like BM25 (Section 3) or modern bi-encoder models (Section 6).
2. **Reranking (or Second-Stage Ranking):** In the second stage, a more powerful but computationally expensive model, known as a *reranker*, is applied only to the small candidate set returned by the retriever. This model can afford to perform deep, fine-grained analysis of the interaction between the query and each candidate document to produce a more accurate final ranking. Examples include Learning-to-Rank models (Section 4) and modern cross-encoder transformers (Section 6).

This two-stage process is a central architectural pattern in IR, and much of the evolution discussed in this survey can be understood as developing more advanced models for each of these stages.

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 a document can also refer to a webpage or an email, depending on the actual IR application of interest.

Disentangling Model Architecture from Training Strategies. Similar to other applied ML domains, the performance of an IR system is a product of its model architecture, its training methodology (e.g., loss functions, data augmentation, optimization algorithms), and deployment best practices (e.g., indexing, quantization, parallelization, algorithm-hardware co-design). In this survey, we intentionally seek to disentangle these aspects to provide a focused analysis on the **evolution of model architecture**. This focus allows for a clearer narrative on how the core components for representation learning and relevance estimation have changed over time, from term-based logic to deep neural networks. We refer readers to dedicated surveys for in-depth reviews of training strategies and other 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 Information Retrieval (IR) models prior to neural methods, with a focus on the **Boolean model**, **vector space model**, **probabilistic model**, and **statistical language model**. These models, which serve as the foundation for later developments in IR (Sections 4 to 7), are built upon the basic unit of a "term" in their representations (Nie, 2010).

Boolean Model. In the Boolean Model, a document $\mathcal{D}$ is represented by a set of terms it contains, i.e., $\mathcal{D} = \{t_1, t_2, \ldots, t_n\}$, and 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. This basic model can be extended by incorporating term weighting, allowing both queries and documents to be represented as sets of weighted terms. Consequently, the logical implication $\mathcal{D} \rightarrow \mathcal{Q}$ is also weighted. Common approaches for this include using a fuzzy set extension of Boolean logic (Radecki, 1979; Kraft & Buell, 1983) and the p-norm (Salton et al., 1983).

Vector Space Model. In Vector Space Models (VSMs, 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 is defined by a vocabulary of terms $\mathcal{V} =< t_1, t_2, \ldots, t_n >$ 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 $\mathcal{D}$. The weights q_i or d_i can be determined by more sophisticated schema (Salton & Buckley, 1988), such as TF-IDF (Sparck Jones, 1972) and BM25 (Robertson & Zaragoza, 2009; Robertson et al., 1995). This allows for more abundant features that can improve the capacity and accuracy of the 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 the 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 & 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.

In contrast, a line of statistical retrieval functions such as TF-IDF (Sparck Jones, 1972) move beyond binary term indicators by incorporating term frequency (TF) and inverse document frequency (IDF), allowing

more nuanced term weighting while still assuming term independence. We illustrate the famous BM25 algorithm (Robertson et al., 1995):

$$\text{BM25}(\mathcal{Q}, \mathcal{D}) = \sum_{t_i \in \mathcal{Q} \cap \mathcal{D}} \text{IDF}(t_i) \cdot \frac{f_i \cdot (k_1 + 1)}{f_i + k_1 \cdot \left(1 - b + b \cdot \frac{|\mathcal{D}|}{\text{avgdl}}\right)},$$

where f_i is the frequency of term t_i in document $\mathcal{D}$, $|\mathcal{D}|$ is the length of the document, avgdl is the average document length in the collection, and k_1 and b are hyperparameters typically set between $[1.2, 2.0]$ and $[0.5, 0.8]$, respectively. The inverse document frequency term is computed as $\text{IDF}(t_i) = \log \frac{N - n_i + 0.5}{n_i + 0.5}$, where N is the total number of documents in the collection and n_i is the number of documents containing term t_i.

The smoothing mechanisms (Baeza-Yates & Ribeiro-Neto, 1999) are necessary to deal with zero occurrences of t_i. Except for the binary independence retrieval model, more sophisticated probabilistic models have been proposed in the literature (Wong & 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 & 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 $\mathcal{Q} = \{t_i\}_{i=1}^n$, we have $\mathcal{P}(\mathcal{Q}|\mathcal{D}) = \prod_{t_i \in \mathcal{Q}} \mathcal{P}(t_i|\mathcal{D})$. The probability $\mathcal{P}(t_i|\mathcal{D})$ is determined using a statistical language model θ_D that represents the document. The relevance is then estimated by log-likelihood: $\text{Score}(\mathcal{Q}, \mathcal{D}) = \log \mathcal{P}(\mathcal{Q}|\theta_D) = \sum_{t_i \in \mathcal{Q}} \log \mathcal{P}(t_i|\theta_D)$, where the estimation of the language model θ_D is usually achieved by maximum likelihood.

The statistical language models for IR (Miller et al., 1999; Berger & Lafferty, 1999; Song & Croft, 1999; Hiemstra & Kraaij, 1999) also encounter the problem of zero occurrences of a 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 & Lafferty, 2004; Zhai et al., 2008) include Jelinek-Mercer smoothing (Jelinek, 1980), Dirichlet smoothing (MacKay & Peto, 1995), etc.

These foundational models, while useful, share a common characteristic: they rely on heuristic-based scoring functions derived from statistical properties of terms. Although their parameters can be tuned (e.g., k_1 and b in BM25), the functional form of the model is fixed. A natural evolution is to frame ranking as a supervised machine learning task, where a model learns to combine various signals of relevance automatically from labeled data. This approach allows for the systematic combination of not only scores from traditional models like BM25 but also a multitude of other features describing the query, the document, and their interaction. This paradigm shift from hand-crafted formulas to learned functions is the core idea behind Learning-to-Rank models, which we discuss next.

4 Learning-to-Rank Model Architectures

Different from traditional IR models that rely on heuristic-based scoring formulas (Section 3), Learning-to-Rank (LTR) frames ranking as a supervised machine learning problem (Liu, 2009). The core idea is to train a model that can optimally combine a wide array of signals, or "features", to predict the relevance of documents to a query. For each query-document pair $(\mathcal{Q}_i, \mathcal{D}_i)$, a k-dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^k$ is extracted, and a relevance label $\mathbf{y}_i$ (e.g., from human judgments) is provided. The goal is to learn a ranking model f parameterized by θ that minimizes an empirical loss $l(\cdot)$ on a labeled training set Ψ:

$$\mathcal{L} = \frac{1}{|\Psi|} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \Psi} l(f_\theta(\mathbf{x}_i), \mathbf{y_i}).$$

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

Name	Backbone Architecture	Loss Function
MART (Friedman, 2001)	Boosting	Pointwise
RankBoost (Freund et al., 2003)	Boosting	Pairwise
RankNet (Burges et al., 2005)	Neural Network	Pairwise
RankSVM (Joachims, 2006)	SVM	Pairwise
LambdaRank (Burges et al., 2006)	Neural Network	Pairwise
ListNet (Cao et al., 2007)	Neural Network	Listwise
SoftRank (Taylor et al., 2008)	Neural Network	Listwise
ListMLE (Xia et al., 2008)	Linear	Listwise
LambdaMART (Burges, 2010)	GBDT	Listwise
ApproxNDCG (Qin et al., 2010)	Linear	Listwise
DLCM (Ai et al., 2018a)	Neural Network	Listwise
GSF (Ai et al., 2019)	Neural Network	Listwise
ApproxNDCG (Bruch et al., 2019)	Neural Network	Listwise
SetRank (Pang et al., 2020)	Self Attention Blocks	Listwise

LTR methods are typically categorized into three main approaches based on their input and loss function: pointwise, pairwise, and listwise.

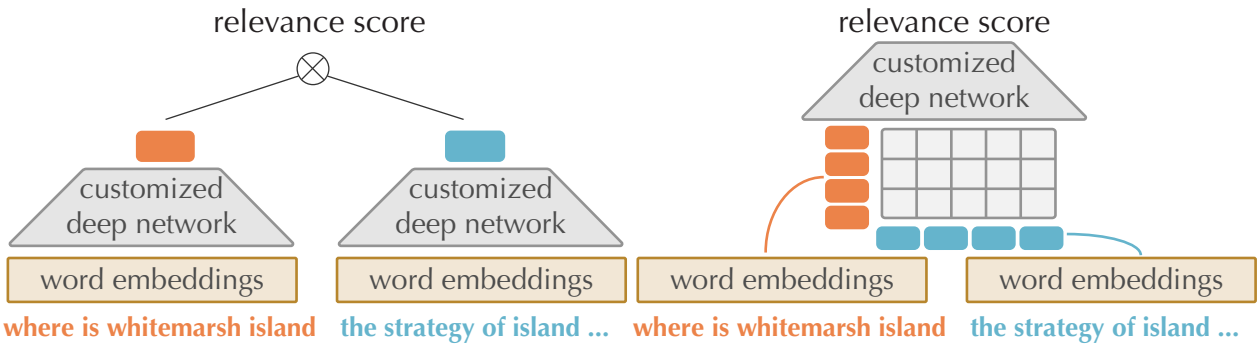
Feature Engineering in LTR. A cornerstone of traditional LTR is the meticulous engineering of the feature vector $\mathbf{x}_i$. These features are designed to capture diverse aspects of relevance and can be grouped into several categories: (1) **Query-based features**, such as the number of terms in the query; (2) **Document-based features**, which are query-independent, such as document length, PageRank (Brin & Page, 1998), or the number of incoming URL links; and (3) **Query-document interaction features**, which form the largest and most critical group. This category includes scores from traditional IR models like BM25 and language models, counts of matching terms, proximity features measuring how close query terms are in the document, and various TF-IDF-related statistics. The power of LTR lies in its ability to learn complex, non-linear combinations of these diverse signals, moving beyond what a single hand-tuned formula could achieve.

4.1 Pointwise, Pairwise, and Listwise Approaches with ML Models

The pointwise approach is the simplest, treating each document independently. It frames the problem as a regression or classification task, where the model $f(\mathbf{x}_i)$ aims to predict the exact relevance label $\mathbf{y}_i$. While straightforward, this approach ignores the crucial fact that ranking is about the relative order of documents, not their absolute scores (Burges, 2010).

The pairwise approach addresses this by focusing on the relative order of document pairs. Given two documents $\mathcal{D}_i$ and $\mathcal{D}_j$ for the same query, the goal is to predict which one is more relevant. This transforms ranking into a binary classification problem. Seminal pairwise models include RankSVM (Joachims, 2006), which adapts the Support Vector Machine framework to maximize the number of correctly ordered pairs, and RankNet (Burges et al., 2005), which uses a neural network and a probabilistic cost function based on pairwise logistic loss. While more aligned with the nature of ranking than the pointwise approach, pairwise methods still do not directly optimize the list-based evaluation metrics (e.g., nDCG, MRR) that are standard in IR evaluation.

The listwise approach directly tackles this issue by defining the loss function over an entire list of documents for a given query. These methods aim to directly optimize ranking metrics. A pivotal line of work began with LambdaRank (Burges et al., 2006), which observed that for gradient-based optimization, one only needs the gradients of the loss function. It introduced "Lambda gradients", which are derived from the change in an IR metric (like nDCG) when two documents in a ranked list are swapped. This technique was then combined with Multiple Additive Regression Trees (MART), a Gradient Boosted Decision Tree (GBDT)



(a) Representation-based neural reranker (b) Interaction-based neural reranker

Figure 2: Illustration on neural ranking models. Brown boxes indicate uncontextualized word embeddings (e.g., Word2vec).

algorithm (Friedman, 2001), to create LambdaMART (Wu et al., 2010). Due to its strong performance and robustness, LambdaMART became a dominant industry standard for many years (Ke et al., 2017).

4.2 Neural LTR Models

While LambdaMART represents a peak for GBDT-based LTR, early works also explored neural networks for the ranking function f_θ. RankNet and LambdaRank both parameterized the LTR model with neural networks. More recent works such as GSF (Ai et al., 2019) and ApproxNDCG (Bruch et al., 2019) have continued this trend, using multiple fully connected layers and designing differentiable approximations of IR metrics. Other architectures like DLCM (Ai et al., 2018a), based on RNNs, and SetRank (Pang et al., 2020), using self-attention, explore ways to model the entire document list jointly. A rigorous benchmark by Qin et al. (2021) compared the performance of these modern neural ranking models against strong GBDT-based baselines. A summary of LTR models and their backbone architectures is provided in Table 1.

4.3 Orthogonal Directions

Beyond core model architectures, LTR research has explored many other important directions. A significant portion of the literature focuses on loss functions and feature transformations (Qin et al., 2021; Bruch et al., 2019; Burges, 2010). Other critical areas include developing methods for unbiased relevance estimation from biased user feedback (e.g., clicks) (Joachims et al., 2017; Ai et al., 2018b; Wang et al., 2018; Hu et al., 2019) and jointly optimizing for both effectiveness and fairness in ranking systems (Singh & Joachims, 2018; Biega et al., 2018; Yang et al., 2023b;a;c). We omit detailed discussions here and refer readers to the original papers and prior surveys (Liu, 2009; Li, 2011).

Despite their success, traditional LTR models have a fundamental ceiling. Their reliance on handcrafted features prevents them from operating in an end-to-end manner and, more importantly, limits their ability to bridge the **lexical gap** — the difference between the words in a query and the semantically related words in a relevant document. Their understanding is based on pre-defined statistical signals, not the underlying meaning of the text. This limitation created a clear need for a new class of models capable of learning semantic representations directly from raw text, setting the stage for the rise of neural ranking (Section 5).

5 Neural Ranking Models

Neural ranking models emerged to directly address the key limitations of LTR (Section 4). By learning semantic representations directly from raw text, they could automatically bridge the lexical gap — for instance, recognizing that a query for "computer" is conceptually related to a document about a "PC" without relying on term overlap. This end-to-end approach simultaneously eliminated the laborious process of manual

feature engineering, shifting the paradigm from engineering statistical signals to learning semantic patterns from data.[1]

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). This division reflects a fundamental tradeoff between efficiency and matching depth. Representation-based models pre-encode documents into vectors offline, enabling highly efficient retrieval suitable for first-pass ranking. In contrast, interaction-based models process the query and document together, allowing for deeper and more precise matching at a higher computational cost, making them ideal for reranking a smaller set of candidates.

5.1 Representation-Based Models

This genre of models can be regarded as an extension of Vector Space Models (Section 3), which independently encode queries and documents into a shared latent vector space, where relevance is determined through simple comparison functions such as cosine similarity or dot product, as illustrated in Figure 2a. This approach maintains clear separation between query and document processing, with no interaction occurring during the encoding procedure. The core architectural challenge is designing an encoder network that transforms a variable-length sequence of term embeddings into a single, fixed-size semantic vector.

The Deep Structured Semantic Model (DSSM, Huang et al., 2013; Gao et al., 2014) is an early example. It utilizes word hashing (a technique to manage large vocabularies by grouping words into a smaller number of hash buckets) 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(C-DSSM, Shen et al., 2014a) leverages a convolutional neural network (CNN) architecture. Specifically, it applies 1D convolutions over the sequence of word embeddings, allowing the model to learn representations for n-grams and local phrases. A max-pooling layer then selects the most salient local features to form the final document vector. Another variant of DSSM replaces MLPs with recurrent layers such as Long Short-Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997; Palangi et al., 2016; Wan et al., 2016; Cohen & Croft, 2016) or tree-structured networks (Tai et al., 2015). The LSTM processes the text sequentially, and its recurrent nature allows it to capture word order and long-range dependencies across the entire text, with the final hidden state often used as the comprehensive representation for the query or document. Based on the assumption of documents' hierarchical structure, Yang et al. (2016b); Song et al. (2018); Zhu et al. (2019) use attention (Bahdanau et al., 2014) to model token, phrase and sentence representations for enhanced document/passage representations.

In line with these works, the NLP community has also extensively investigated passage/document representations. Le & Mikolov (2014) proposed Paragraph Vectors (Doc2Vec), an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, which is based on single-hidden-layer neural network. Kim (2014) studied CNN for sentence representations, while Wieting et al. (2016) proposed to use LSTM network. Arora et al. (2017) reported a weighted average of pre-trained word embeddings fine-tuned with unsupervised random walk algorithm can outperform more complex neural networks on the sentence similarity task. However, weighted averaging word embeddings ignores the word order, which fails to capture the rich, contextual information in longer, more complex documents.

To summarize, representation-based models excel in scenarios requiring global semantic understanding and offer significant computational advantages through their ability to pre-compute document representations offline (Guo et al., 2020). However, these approaches also face inherent limitations due to their reliance on fixed-size embedding vectors, which can struggle to capture all relevant information from the original text and may not effectively handle precise lexical matching requirements. These limitations are the focus of interaction-based models.

[1]For structural clarity, this section focuses on neural information retrieval, covering neural network–based retrieval models developed prior to the advent of pre-trained transformers. For a more comprehensive treatment, we refer readers to the dedicated surveys (Onal et al., 2018; Mitra et al., 2018; Xu et al., 2018).

5.2 Interaction-Based Models

Different from representation-based models, interaction-based models (Figure 2b) process queries and documents jointly through neural networks. Instead of compressing each text into a single vector, they first build a detailed, low-level interaction representation between the query and the document, and then use neural networks to learn hierarchical matching patterns from this representation. 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 with its 2D filters to capture local matching patterns (e.g., phrases or bigrams matching) through convolution and pooling operations (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). Rather than a single interaction matrix, DRMM creates a matching histogram for each query term. This histogram discretizes the similarity scores against all document terms into bins, effectively capturing the distribution of matching signals (e.g., how many terms in the document are an exact match, a strong semantic match, or a weak match to a given query term). An MLP then learns the relevance contribution from these histogram features.

Kernel-Based Neural Ranking Model (K-NRM, Xiong et al., 2017) further advances interaction-based approaches. It employs radial basis function (RBF) kernels to transform the query-document interaction matrix into a more informative feature representation. Each kernel corresponds to a certain similarity level (e.g., "exact match", "strong match", "weak match"). The model uses these kernels to produce "soft-TF" counts for each query term — counting how many document words match the query term at each similarity level. These soft-match features are then aggregated and fed into a simple feed-forward network to compute the final relevance score. This kernel-based mechanism enables models to capture nuanced matching features, enhancing their ability to model complex query-document interactions. Conv-KNRM (Dai et al., 2018) later extends it to convolutional kernels to capture n-gram level soft matches, further improving matching granularity.

In line with these works, the interaction matrix-based approach have been explored for short text matching (Lu & Li, 2013; Yin et al., 2016; Yang et al., 2016a) as well as long document ranking (Mitra et al., 2017; Hui et al., 2017; 2018, *inter alia*). Multi-Perspective CNN approaches compare sentences via diverse pooling functions and filter widths to capture multiple perspectives between texts He et al. (2015). aNMM (Yang et al., 2016a), as an example of attention-based methods, computes passage terms' attention weights over query terms using a query attention network and achieves performance improvement compared to CNN-based baseline (Severyn & Moschitti, 2015). Term adjacency and positional information represent another important dimension of interaction modeling. Models such as MatchPyramid, PACRR, and ConvKNRM capture term adjacency patterns and position-dependent interactions (Pang et al., 2016; Hui et al., 2017; Dai et al., 2018).

The interaction functions in these models can be categorized as either non-parametric (using traditional similarity measures like cosine similarity, dot product, or binary indicators) or parametric (learning similarity functions from data through neural networks) (Dong et al., 2022b). While interaction-focused models require one forward pass through the entire model for each potentially relevant document, making them computationally more expensive than representation-focused approaches, they typically achieve superior ranking quality due to their ability to capture fine-grained matching signals. We list some representative works in Table 2 and direct readers to these works for architectural details.

5.3 Hybrid Models

Recognizing the complementary strengths of representation-focused and interaction-focused architectures, researchers have proposed hybrid models that combine the efficiency of representation-based methods with the effectiveness of interaction-based approaches. These models represent a third category in neural ranking architectures, alongside the two primary approaches.

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

Name	Architecture	Backbone	Embeddings
DSSM (Huang et al., 2013)	Representation-based	MLP	Semantic Hashing
CDSSM (Shen et al., 2014a)	Representation-based	CNN	Semantic Hashing
CLSM (Shen et al., 2014b)	Representation-based	CNN	Semantic Hashing
ARC-I (Hu et al., 2014)	Representation-based	CNN	Word2Vec
Tai et al. (2015)	Representation-based	Tree-structured LSTM	GloVe
LSTM-RNN (Palangi et al., 2016)	Representation-based	LSTM	Randomly Initialized
MV-LSTM (Wan et al., 2016)	Representation-based	Bi-LSTM	Word2Vec
DESM Nalisnick et al. (2016)	Representation-based	MLP	Randomly Initialized
Lu & Li (2013)	Representation-based	MLP	Randomly Initialized
ARC-II (Hu et al., 2014)	Interaction-based	CNN	Word2Vec
MatchPyramid (Pang et al., 2016)	Interaction-based	CNN	Randomly Initialized
DRMM (Guo et al., 2016a)	Interaction-based	MLP	Word2Vec
ABCNN (Yin et al., 2016)	Interaction-based	CNN + Attention	Word2Vec
aNMM (Yang et al., 2016a)	Interaction-based	Attention	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
PACRR (Hui et al., 2017)	Interaction-based	CNN + RNN	Word2Vec
Co-PACRR (Hui 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

The most notable example is DUET (Mitra et al., 2017), which employs two separate deep neural networks operating in parallel. One network performs local interaction-based matching similar to interaction-focused models, while the other learns distributed representations for query and document separately, similar to representation-focused approaches. The term interaction matrix between query and document feeds into the exact matching layers, while term embeddings of the input sequence enter the semantic matching layers. The outputs from both networks are then combined using a fully connected network to produce the final ranking score.

Different from the metric learning theme of representation-based models, a line of works formulates the ranking problem as a classification problem (commonly referred to as **Extreme Label Classification**, or **XMC**), where the input is the query, and the output is a probability distribution over the corpus, where each document is a unique "class" or "label". Instead of optimizing the similarity between query and document representations, XMC models aim to predict the correct subset of relevant document IDs (Prabhu & Varma, 2014; Jain et al., 2016; Liu et al., 2017; Jain et al., 2019). In the inference time, XMC methods use tree-based hierarchies or cluster-based sampling to quickly narrow down the search path to the likely labels without scanning every candidates (Prabhu et al., 2018; You et al., 2019). AttentionXML (You et al., 2019) uses an attention mechanism to focus on specific parts of the input text that are most relevant to the label's semantic meaning, and thus can be considered a hybrid model. We refer readers to (Dasgupta et al., 2023) for a comprehensive review of XMC methods.

This hybrid architecture demonstrates that combining distributed representations with traditional local representations is favorable, with the combined approach significantly outperforming either neural network individually. More recent hybrid approaches have focused on reducing computational costs while maintaining effectiveness, with some models incorporating cached token-level representations to enable faster query-document interactions when document representations are pre-computed (Wrzalik & Krechel, 2021). The success of hybrid models has established that interaction-based and representation-based approaches can be effectively combined for further improvements in ranking performance (Liu et al., 2018).

5.4 Orthogonal Directions

In addition to the development of network architecture, pre-trained embeddings (Salakhutdinov & Hinton, 2009; Mikolov et al., 2013; Pennington et al., 2014; Le & Mikolov, 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). Both GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013) learn dense vector representations for each vocabulary term from large-scale text corpora. By initializing the embedding layer with these pre-trained vectors, models start with a strong semantic foundation, which proved crucial for performance, especially on smaller training datasets (Guo et al., 2016b). Interaction-based models with cross-lingual word embeddings (Joulin et al., 2018) for cross-lingual reranking have also been explored (Yu & 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), Long Short Term Memory (LSTM, Hochreiter & 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 Section 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 (Section 5) and retrieval with pre-trained transformers (Section 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 an interaction module inspired by K-NRM, where RBF kernels are used to create soft-match features from the contextualized embeddings. This fusion of a transformer encoder with a kernel-based interaction mechanism allowed the model to achieve better performance-efficiency tradeoff compared to BERT-based reranker (Nogueira et al., 2019b). 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.

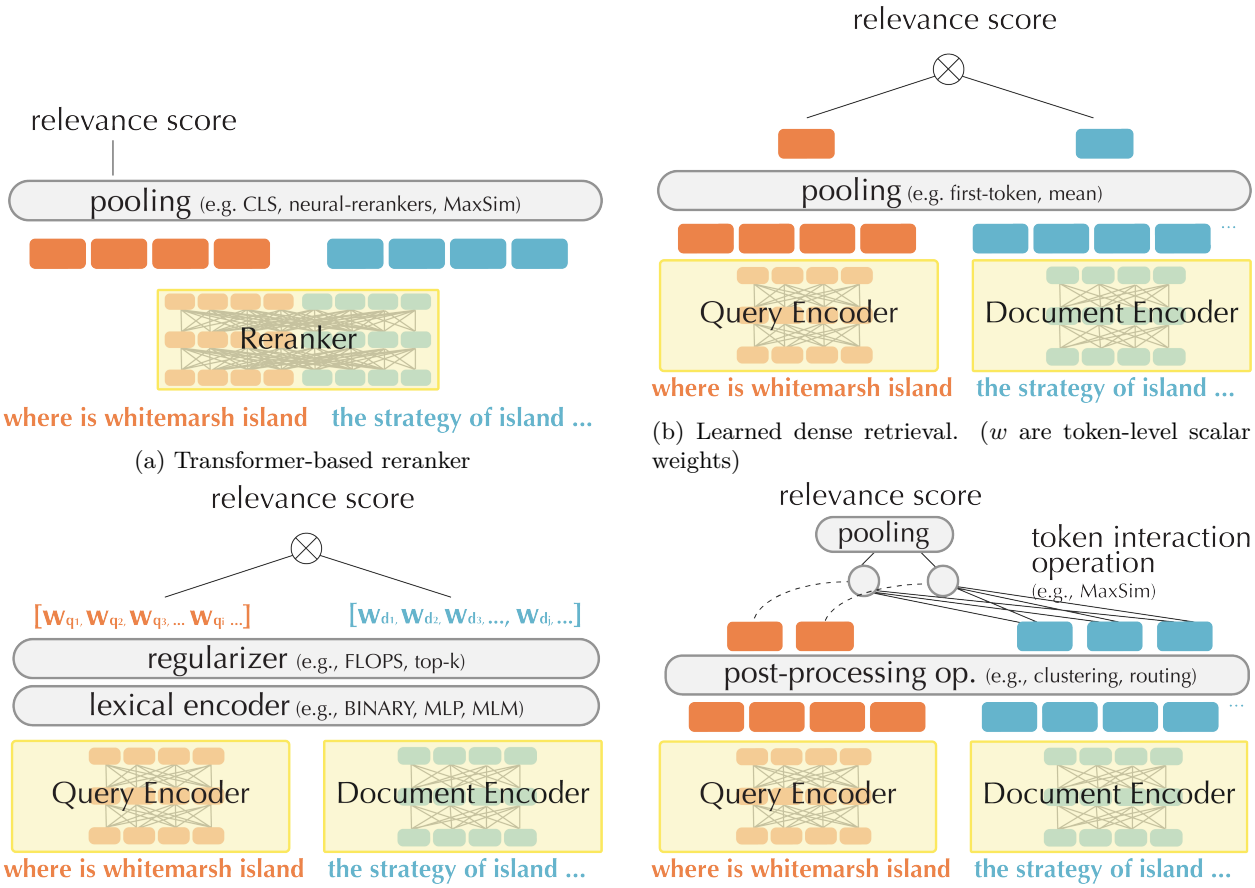
The neural ranking models described above, particularly later developments like TK and TKL, demonstrated the potential of the transformer's attention mechanism for modeling relevance. However, the true paradigm shift occurred when the IR community moved from using these architectures trained from scratch to leveraging massive, pre-trained transformer models like BERT (Devlin et al., 2019) and its variants (Liu et al., 2019; Sun et al., 2019; Lan et al., 2020; Beltagy et al., 2020). This marked a fundamental change in approach: instead of designing novel, task-specific network backbones (e.g., CNNs, LSTMs) on top of static word embeddings, research shifted to fine-tuning a single, powerful, and deeply contextualized architecture for IR tasks. This new foundation did not eliminate the core architectural tradeoffs but rather recast them in a more powerful form, leading to the development of cross-encoder rerankers and bi-encoder retrievers, which we explore next.

6 IR with Pre-Trained Transformers

BERT (Devlin et al., 2019) revolutionized research in both natural language processing (NLP) and information retrieval (IR). Its success is largely attributed to two key factors: (1) the Multi-Head Attention (MHA) architecture (Vaswani et al., 2017), which enables high-dimensional, contextualized token representations; and (2) large-scale pre-training, which equips BERT with the ability to capture rich semantics and world knowledge. The expressive power of BERT has been extensively analyzed in prior work, e.g., (Rogers et al., 2020; Tenney et al., 2019; Clark et al., 2019).

High-level architectural families. Before discussing specific IR modeling architectures, it is useful to understand the architectural families of transformers, as they have distinct implications for IR tasks:

- **Encoder-only models** (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)) use a bidirectional self-attention mechanism, allowing each token's representation to be informed by the



(a) Transformer-based reranker

(b) Learned dense retrieval. (w are token-level scalar weights)

(c) Learned sparse retriever

(d) Multi-vector representations

Figure 3: Illustration on transformer-based retrieval and reranking models. Yellow boxes indicate pretrained Transformers (e.g., BERT). Query text, embeddings, and associated weights are color-coded in orange, whereas document representations are color-coded in blue.

entire input sequence (both left and right context). This deep contextual understanding makes them naturally suited for representation-focused tasks. In IR, they have been the workhorses for building powerful bi-encoder retrievers and cross-encoder rerankers.

- **Decoder-only models** (e.g., proprietary GPT series (Radford et al., 2019; Brown et al., 2020; OpenAI et al., 2024), open-weight Llama (Touvron et al., 2023), Mistral (Jiang et al., 2023a)) use a unidirectional (causal) self-attention mechanism, where each token can only attend to previous tokens in the sequence. This architecture is optimized for next-token prediction and, by extension, text generation. Their application to representation tasks like retrieval is less direct and often requires architectural adaptations to create meaningful summary vectors from their unidirectional hidden states.

- **Encoder-decoder models** (e.g., T5 (Raffel et al., 2020), BART (Lewis et al., 2020a)) combine both architectures. The encoder processes the input sequence bidirectionally to create a rich representation, which then conditions the decoder to generate an output sequence autoregressively. This "sequence-to-sequence" design makes them highly versatile. In IR, they can be framed as rerankers (generating a "relevant" or "irrelevant" token), or as generative retrievers that directly generate document identifiers.

This section discusses IR architectures based on pre-trained transformers, with a focus on BERT-type encoder models, which is to be distinct from encoder-decoder models and decoder-only models covered

in Section 7. We structure our review around the fundamental architectural tradeoff between interaction depth and computational efficiency. These constraints necessitate two primary paradigms:

1. **Cross-Encoder (Deep Interaction):** A single model processes the concatenated query and document as one sequence, allowing every query token to interact deeply with every document token. This provides state-of-the-art ranking quality but is computationally expensive, making it suitable only for reranking.
2. **Bi-Encoder (Separable Pre-computation):** Separate encoders process the query and document independently to create fixed-size vectors. Since document vectors can be pre-computed offline, this architecture enables extremely fast similarity search suitable for first-stage retrieval.

We first discuss the crucial training strategies used to optimize these two architectures. We then detail the cross-encoder models that perform deep, full interaction, followed by the separable bi-encoder architectures that prioritize efficiency, exploring their dense, sparse, and multi-vector variants. Finally, we discuss advanced hybrid models and orthogonal improvements such as continual training and interpretability.

6.1 Training Strategies for Transformer-Based IR

Although we aim to disentangle model architectures from training strategies, the co-evolvement of these two areas is a defining pattern of this era. The architectural dichotomy (cross-encoder vs. bi-encoder) has an impactful influence the training methodology, extending the loss function categories discussed in Section 4 into the deep learning paradigm.

Contrastive and Listwise Objectives. The application of contrastive learning builds on the principle of the InfoNCE loss (Oord et al., 2018), which is derived from Noise-Contrastive Estimation (Gutmann & Hyvärinen, 2012). The general goal is to learn a model that distinguishes a "positive" sample from a set of "negative" samples.

The InfoNCE framework is primarily used to optimize **bi-encoders** in the dense retrieval setting. In this context, the relevance score f is a simple similarity function (e.g., dot product) between the query and document vectors, $f(\mathcal{Q}, \mathcal{D}) = \text{sim}(\mathbf{v}_\mathcal{Q}, \mathbf{v}_\mathcal{D})$. For a query $\mathcal{Q}_i$, a positive document $\mathcal{D}_i^+$, and a set of negative documents $\mathcal{D}_i^-$, the bi-encoder is trained to minimize the negative log probability of correctly classifying the positive document:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{|\mathcal{S}|} \sum_{(\mathcal{Q}_i, \mathcal{D}_i^+) \in \mathcal{S}} \log \frac{\exp f_\theta(\mathcal{Q}_i, \mathcal{D}_i^+)}{\exp f_\theta(\mathcal{Q}_i, \mathcal{D}_i^+) + \sum_{\mathcal{D}_j^- \in \mathcal{D}_i^-} \exp f_\theta(\mathcal{Q}_i, \mathcal{D}_j^-)}$$

where $\mathcal{S}$ is the training set and $\mathcal{D}_i^-$ is the set of sampled negative documents.

For **cross-encoders**, which compute relevance over a concatenated list, the objective is also listwise but often simplified to a standard Negative Log-Likelihood (NLL) loss over the final softmax probabilities of the candidates, minimizing the distance between the predicted distribution and the ground truth relevance distribution for the entire list.

Hard Negative Mining. A critical challenge in training bi-encoders is generating sufficiently difficult negative examples, as random sampling typically yields easy negatives that do not challenge the model effectively. This is where Hard Negative Mining (HNM) becomes essential. HNM strategies ensure that the model is exposed to challenging cases where positive and negative document features are difficult to distinguish. Key strategies include:

1. **In-Batch Negatives (IBN):** Leveraging other queries' positive documents within the same mini-batch as negative examples for the current query. IBN provides a balance of efficiency and difficulty.
2. **Lexical Negatives:** Using documents highly ranked by a traditional sparse model (e.g., BM25) but not labeled as relevant.

3. **Iterative Hard Negative Mining:** Employing an existing dense retriever to periodically mine difficult negatives from the collection (i.e., documents that the current model mistakenly ranks highly). Seminal methods like ANCE (Xiong et al., 2020) and ADORE (Zhan et al., 2021) use this iterative approach to continually feed the model better training data.

Knowledge Distillation. To further narrow the effectiveness gap between efficient bi-encoders and powerful cross-encoders, the community widely adopted Knowledge Distillation (KD) (Buciluǎ et al., 2006; Hinton et al., 2015). KD is a technique for training a smaller, efficient "student" model by transferring knowledge from a larger, more capable "teacher" model (Hinton et al., 2015; Gou et al., 2021). In IR, KD is used to create fast rerankers or retrievers that approximate the performance of slower, larger models, which is critical for production systems (Hofstätter et al., 2020a; 2021a; Chen et al., 2021; Kim et al., 2023; Xu et al., 2025c; Zhang et al., 2025b; Vera et al., 2025).

Let f_t denote the teacher model (typically a cross-encoder) and f_s denote the student model (typically a bi-encoder or a smaller cross-encoder). For a given query $\mathcal{Q}$ and a list of candidate texts $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k\}$, we first compute relevance scores (logits) from both models:

$$\mathbf{z}_t = [f_t(\mathcal{Q}, \mathcal{D}_1), f_t(\mathcal{Q}, \mathcal{D}_2), \ldots, f_t(\mathcal{Q}, \mathcal{D}_k)]$$
$$\mathbf{z}_s = [f_s(\mathcal{Q}, \mathcal{D}_1), f_s(\mathcal{Q}, \mathcal{D}_2), \ldots, f_s(\mathcal{Q}, \mathcal{D}_k)]$$

The student model f_s is trained to mimic the teacher's output distribution over the candidate texts by minimizing the Kullback-Leibler (KL) divergence between the two softened probability distributions:

$$\mathcal{L}_{\text{KD}} = D_{\text{KL}}\left(\text{softmax}\left(\frac{\mathbf{z}_t}{T}\right) \middle| \text{softmax}\left(\frac{\mathbf{z}_s}{T}\right)\right)$$

where T is the temperature hyperparameter. A higher temperature creates a softer probability distribution, which can help in transferring more nuanced information from the teacher. KD thus provides an essential bridge, allowing the efficiency of bi-encoders to approach the effectiveness ceiling set by cross-encoders.

6.2 Deep Interaction Models: The Cross-Encoder for Reranking

The most effective application of transformers in IR involves full, deep interaction between query and document tokens. This architecture, known as a cross-encoder (Humeau et al., 2020), takes the concatenated sequence of $(\mathcal{Q}, \mathcal{D})$ as a single input. Nogueira et al. (2019b) first employed this approach in their MONOBERT model for reranking candidate passages from a first-stage retriever. The model outputs a relevance score s via a linear layer on top of the final BERT representation, typically from a linear layer using the `[CLS]` token's representation (Figure 3a).

Conceptually similar to pre-Transformer interaction-based neural ranking models, this schema has proven effective across various pre-trained encoders (Zhang et al., 2021b), as well as other transformer architectures (Section 7). However, this cross-encoder schema faces two primary challenges: (1) the fixed context length of models like BERT (e.g., 512 tokens) makes processing long documents difficult, and (2) relying on a single token's fixed-dimensional representation (e.g., 768-dimensional representation for BERT) may limit the model's expressive power.

Handling Long Documents. Corresponding mitigations for these two challenges have been extensively investigated in literature. **Chunk-and-aggregate** approaches represent a practical solution to handle long documents that exceed BERT's input constraints by decomposing the ranking problem into passage-level scoring followed by aggregation (Gao & Callan, 2022). The fundamental strategy involves splitting documents into fixed-length passages or sentences, applying BERT-based cross-encoders to score each query-passage pair independently, then combining these scores to produce a final document-level relevance score. Early work in this direction explored sentence-level aggregation, where BERT scores computed at the sentence level were shown to be effective for document ranking (MacAvaney et al., 2020; Yilmaz et al., 2019). The BERT-MaxP (Dai & Callan, 2019a) approach became particularly influential, where documents are split into fixed-length passages and the maximum passage score serves as the document score.

Two primary aggregation strategies have emerged: (1) score-pooling and (2) representation aggregation. Score-pooling methods apply simple operations like maximum, sum, or first passage scores to combine passage-level relevance scores (Dai & Callan, 2019a). In contrast, representation aggregation methods address both the long-document problem and the single-vector expressiveness limitation. Instead of collapsing each passage's signal into a single scalar score, these approaches gather the rich, low-dimensional `[CLS]` token representations from each passage. This collection of vectors forms a more comprehensive and expressive document-level feature set, which is then processed by additional neural networks (MacAvaney et al., 2020). Notable systems like PARADE (Li et al., 2023a) employ CNNs and transformers for aggregation, while CEDR (MacAvaney et al., 2019b) pioneered joint approaches that combine BERT outputs with existing neural IR models through averaging.

While chunk-and-aggregate approaches successfully handle long documents, they fundamentally limit query-document interactions to the passage level, creating information bottlenecks where passage scores or low-dimensional representations constrain the model's ability to capture document-wide relevance patterns. Hofstätter et al. (2021b) argue that this tradeoff between scalability and interaction richness remains a defining characteristic of this approach.

6.3 Efficient Pre-Computation Models: The Bi-Encoder Architecture

While cross-encoders offer state-of-the-art effectiveness, their computational cost — requiring a full transformer pass for every $(\mathcal{Q}, \mathcal{D})$ pair — makes them infeasible for retrieval over large collections. This limitation motivated the development of bi-encoder architectures, which are conceptually similar to representation-based neural ranking models (Section 5).[2]

A bi-encoder uses a backbone network (typically a transformer) to encode the query $\mathcal{Q}$ and document $\mathcal{D}$ separately. The resulting dense vector representations are then used to compute a relevance score with a simple similarity function like dot product or cosine similarity (Xiong et al., 2020; Karpukhin et al., 2020; Gao et al., 2021b). The key advantage of this separation is efficiency: the entire document collection can be encoded into a vector index offline. At query time, retrieval becomes a fast approximate nearest neighbor (ANN) search problem (Johnson et al., 2019; Malkov & Yashunin, 2018) or search with an inverted index data structure (Zobel et al., 1998), avoiding costly neural network inference.

A notable insight from Lin (2021) is that the bi-encoder framework provides a unifying lens for understanding diverse retrieval approaches. Dense retrieval models, learned sparse retrieval models, and traditional bag-of-words approaches such as BM25 can all be interpreted as parametric variations of this architecture. They differ primarily along two axes: (1) the representational basis (dense semantic embeddings versus sparse lexical vectors), and (2) whether those representations are learned from data or manually engineered. Existing methods based on this bi-encoder architecture vary primarily in their representation format (dense versus sparse), pooling strategies, and training methodologies.

6.3.1 Learned Dense Retrieval

Dense retrieval models typically adopt a standardized dual-encoder architecture built on pre-trained transformer models, most commonly BERT in this context. The standard formulation uses separate BERT encoders for queries and documents, with a layer-normalized linear projection applied to the token representation: $\text{Encoder}(\cdot) = \text{Linear}\big(\text{BERT}(\cdot)\big)$. The encoder weights can be separate or shared between query and document sides.

The core architectural principle involves encoding queries and documents into low-dimensional dense vectors, typically 768 dimensions matching BERT's hidden size. Rather than using all hidden representations, most models compress the sequence information using a reduction function, usually the token representation or mean pooling of the final transformer layer outputs. This creates a single dense vector representation per text sequence that captures semantic information beyond simple lexical matching.

[2]The term "bi-encoder" is also known as a two-tower architecture or an embedding model. We use "bi-encoder" to contrast with "cross-encoder", which takes concatenated input.

Relevance scoring in dense retrieval is performed through simple similarity functions, most commonly dot product or cosine similarity between query and document vectors. This design enables efficient ANN search over pre-computed document representations (Johnson et al., 2019), making dense retrieval practical for large-scale collections while maintaining the semantic understanding capabilities of transformer models. The success of this architecture stems from its ability to learn semantic representations that address the vocabulary mismatch problem inherent in traditional sparse retrieval methods (Lee et al., 2019; Karpukhin et al., 2020; Xiong et al., 2020; Reimers & Gurevych, 2019). Dense retrieval models have demonstrated notable effectiveness improvements over BM25 baselines across various tasks including open-domain question answering and web search.

6.3.2 Learned Sparse Retrieval

Learned sparse retrieval (LSR, Figure 3c) employs the same bi-encoder architecture as dense retrieval but produces fundamentally different representations.[3] While sharing the transformer backbone, sparse retrieval models encode queries and documents into high-dimensional sparse vectors whose dimensionality typically matches the vocabulary size of the underlying pre-trained model, often containing tens of thousands of dimensions. Each dimension corresponds to a specific vocabulary term, creating an interpretable representation where non-zero weights indicate term importance (Formal et al., 2021b;a; Nguyen et al., 2023).

Three key architectural constraints distinguish sparse encoders from their dense counterparts. First, sparsity is enforced through explicit regularization techniques, ensuring most term weights remain zero to maintain efficiency (Formal et al., 2021b; Xu et al., 2025b; 2026). Second, all weights must be non-negative to maintain compatibility with traditional inverted index software designed for lexical search systems like Lucene. Third, the high-dimensional vocabulary-aligned vectors enable integration with existing inverted index infrastructure and optimization algorithms (Turtle & Flood, 1995; Broder et al., 2003; Bruch et al., 2024). Notably, inference-free learned sparse retrieval methods such as SPLADE-Doc (Formal et al., 2021a; Shen et al., 2025) eliminates requirement of specialized accelerators such as GPUs, making them highly efficient for inference on multi-core CPU machines.

At a conceptual level, learned sparse retrieval can be viewed as a sophisticated evolution of traditional term weighting schemes, learning context-aware token importance scores from data rather than relying on heuristic formulas (Zamani et al., 2018; Dai & Callan, 2019b; Mallia et al., 2021; Yu et al., 2024a; Xu et al., 2025b; 2026). This approach inherits desirable properties from bag-of-words models such as exact term matching while leveraging the semantic understanding capabilities of pretrained transformers. Notable implementations include SPLADE (Formal et al., 2021a), DeepImpact (Mallia et al., 2021), and uniCOIL (Lin & Ma, 2021), which demonstrate that transformer-based sparse representations can achieve effectiveness comparable to dense retrieval while maintaining the efficiency benefits of inverted indexes.

6.4 Bridging the Gap: Advanced Interaction and Hybrid Models

The standard bi-encoder's lack of term-level interaction is a performance bottleneck compared to cross-encoders. Several lines of research aim to bridge this gap by introducing more granular representations or by combining different retrieval paradigms.

6.4.1 Multi-Vector Representations

To re-introduce query-document interaction without the full cost of a cross-encoder, multi-vector models represent queries and documents using multiple vectors. Poly-Encoder (Humeau et al., 2020) computes a fixed number of vectors per query and aggregates them with softmax attention over document vectors. ME-BERT (Luan et al., 2021) represents documents with m vectors and uses the maximum similarity between any query and document vector to estimate relevance.

[3] We focus on learned sparse retrieval within the bi-encoder formulation, thereby excluding approaches based on learned document expansion (e.g., Nogueira et al. (2019a); Zbib et al. (2019)) and document term reweighting (e.g., DeepCT (Dai & Callan, 2019b)). For a unified treatment of learned sparse retrieval that encompasses these lines of work, see Mallia et al. (2021); Basnet et al. (2024).

In line with this idea, ColBERT (Khattab & Zaharia, 2020; Santhanam et al., 2022; Hofstätter et al., 2022) represents each token in the query and document as a contextualized vector. It then performs a "late interaction" step where each query vector is compared against all document vectors via a MaxSim operator, and the final score is the sum of these maximum similarities. This late interaction scheme (Figure 3d) allows ColBERT for end-to-end training to achieve strong performance while still achieving efficient retrieval through a dedicated index structure. On the other hand, it also leads to drastically increased index size, which has been the focus in later studies (Santhanam et al., 2022; Hofstätter et al., 2022, *inter alia*).

We should also note that multi-vector retrieval can be viewed as a special case of dense retrieval where the learned feature representation is a matrix of size $n \times h$, with n vectors of hidden dimension h. This matrix can be conceptually flattened into a single dense vector, showing its connection to the vanilla single-vector retrieval. The key difference lies not in the representation itself but in the richer relevance estimation strategy: instead of applying a simple linear relevance like dot product, models like ColBERT aggregate fine-grained token-level interactions to compute a relevance score.

6.4.2 Hybrid Retrieval

Another direction combines the strengths of different retrieval systems. A simple yet effective approach is ranklist fusion (e.g., Reciprocal Rank Fusion, Cormack et al., 2009), which merges ranked lists from sparse (e.g., BM25) and dense retrievers post-retrieval without architectural changes. More integrated models combine signals at a deeper level. COIL (Gao et al., 2021a) enhances traditional bag-of-words retrieval with semantic embeddings from a BERT encoder. uniCOIL (Lin & Ma, 2021) simplifies this by reducing the semantic embedding to a single dimension, effectively learning a term weight akin to LSR models like SPLADE (Formal et al., 2021a;b). 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) takes an alternative approach to first encode contextualized token vectors, then map 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.

6.5 Orthogonal Improvements and Analysis

Beyond architectural innovations, performance can be enhanced through improvements to the underlying models and a deeper analysis of their behavior. We show a list of models and their corresponding architectures in Table 3, a majority of which use BERT (Devlin et al., 2019) as the backbone, with exceptions using DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and T5 (Raffel et al., 2020; Sanh et al., 2022; Mo et al., 2023; Chung et al., 2024).

Continual Training and Adaptation. Instead of modifying the retrieval architecture, this line of work enhances the backbone language model itself through domain adaptation or continued pre-training, a proven strategy in NLP (Howard & Ruder, 2018; Gururangan et al., 2020). For instance, Lee et al. (2019) pre-train BERT with an Inverse-Cloze Task (Taylor, 1953) for better text representations. Condenser (Gao & Callan, 2021) proposes a dedicated pre-training architecture to "condense" text representations into the `[CLS]` token. COCO-DR (Yu et al., 2022c) extends Condenser by using a technique named Distributionally Robust Optimization to mitigate distribution shifts in dense retrieval. A line of works have also explored other pre-training objectives such as masked auto-encoders (Xiao et al., 2022; Wu et al., 2023b) and bag-of-words prediction (Ma et al., 2024a). Recent works (Wang et al., 2023a; Nussbaum et al., 2025; Yu et al., 2024b, *inter alia*) have employed a "middle-stage" training on large-scale unlabeled text pairs, between pretrained encoder models and supervised fine-tuning on labeled text pairs, and have demonstrated the corresponding performance improvement compared to the traditional two-stage pipeline. We refer readers to the original papers for details.

Interpretability and Explainability. A few works have attempted to interpret what transformer-based models learn, i.e., mechanistic interpretability (Elhage et al., 2021; Saphra & Wiegreffe, 2024). MacAvaney

Table 3: Summary of IR model architecture for passage retrieval and passage ranking based on pre-trained transformers. Dense Retrieval and LSR denote learned dense retrieval and learned sparse retrieval, respectively. DeepCT (Dai & Callan, 2019b) is trained without labeled training set. Contrastive Learning and in-batch negatives means listwise loss function is used. SentenceBERT (Reimers & Gurevych, 2019) is originally designed for the symmetrical sentence similarity tasks, but is quickly expanded to asymmetrical retrieval tasks.

Name	Model	Architecture	Backbone LM	Training strategy
monoBERT (Nogueira et al., 2019b)	Reranking	Cross-encoder	BERT	Classification
CEDR (MacAvaney et al., 2019b)	Reranking	Cross-encoder	BERT	Contrastive Learning
BERT-MaxP (Dai & Callan, 2019a)	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
ODQA (Lee et al., 2019)	Dense Retrieval	Bi-encoder	BERT	Unsupervised
SentenceBERT (Reimers & Gurevych, 2019)	Dense Retrieval	Bi-encoder	BERT	Triplet
DPR (Karpukhin et al., 2020)	Dense Retrieval	Bi-encoder	BERT	Contrastive Learning
ANCE (Xiong et al., 2020)	Dense Retrieval	Bi-encoder	RoBERTa	Contrastive Learning
RepBERT (Zhan et al., 2020)	Dense Retrieval	Bi-encoder	BERT	In-batch negatives
Margin-MSE (Hofstätter et al., 2020a)	Dense Retrieval	Bi-encoder	DistilBERT	Distillation
TAS-B (Hofstätter et al., 2021a)	Dense Retrieval	Bi-encoder	BERT	Distillation
RocketQA (Qu et al., 2020)	Dense Retrieval	Bi-encoder	ERNIE	Contrastive Learning
RocketQA-v2 (Ren et al., 2021)	Dense Retrieval	Bi-encoder	ERNIE	Distillation
GTR (Ni et al., 2022b)	Dense Retrieval	Bi-encoder	EncT5	Contrastive Learning
TART-dual (Asai et al., 2023)	Dense Retrieval	Bi-encoder	Contriever	Instruction Tuning
E5 (Wang et al., 2022a)	Dense Retrieval	Bi-encoder	BERT	Contrastive Learning
GTE (Li et al., 2023d)	Dense Retrieval	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 & 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 & Callan, 2019b)	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
uniCOIL Lin & Ma (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

et al. (2022) showed that neural models rely less on exact match signals and instead encode rich semantic information. Ram et al. (2023a) connected dense and sparse retrieval by projecting a dense retriever's intermediate representations into the vocabulary space. Separately, other work focuses on designing systems that provide model-agnostic explanations (Rahimi et al., 2021; Yu et al., 2022b; Xu et al., 2024c) to satisfy desiderata like faithfulness (Jacovi & Goldberg, 2020; Xu et al., 2023). As IR systems become integral to other applied ML domains, we believe it is important to study and design interpretable, truthful, and trustworthy IR models.

The architectural innovations discussed in this section highlight a mature research field dedicated to harnessing pre-trained transformers for information retrieval. The central theme has been the architectural tradeoff between interaction depth and computational cost, giving rise to a spectrum of models from highly effective cross-encoder rerankers to efficient bi-encoder retrievers. By developing advanced representations — whether dense, sparse, or multi-vector — and hybridizing these approaches, the community has pushed the boundaries of the classic "retrieve-then-rank" paradigm. However, these models still primarily function as specialized components for representation and scoring. The next wave of innovation emerged from models capable not only of understanding text but also of generating it, marking the beginning of the era of Large Language Models.

7 Large Language Models for IR

The natural evolution from pre-trained encoders is the recent ascendance of Large Language Models (LLMs).[4] While building on the same transformer principles discussed in Section 6, the sheer scale and generative capabilities of modern instruction-following LLMs are reshaping the architectural landscape of IR. These models are not just larger backbones for feature extraction; their proficiency in language understanding, generation, and instruction-following allows them to take on entirely new roles. Trained to align with human preferences (OpenAI et al., 2023; Gemini Team Google et al., 2023; Bai et al., 2022), LLMs can perform complex tasks such as reasoning (Wei et al., 2022; OpenAI et al., 2024; Guo et al., 2025), tool usage (Schick et al., 2023; Patil et al., 2024b; Qin et al., 2024a; Patil et al., 2024a) and planning (Song et al., 2023; Huang et al., 2024a). In this section, we review how these powerful models — spanning *decoder-only* and *encoder-decoder* architectures — are being adapted for IR tasks, moving beyond established paradigms into new frontiers of retrieval, reranking, and direct generation.

7.1 LLM as Retriever

A straightforward yet highly effective application of LLMs is to serve as the backbone for bi-encoder retrieval models. We categorize these developments into backbone scaling, architectural adaptation, and unified modeling. We show a shortlist of works in Table 5.

Scaling Bi-Encoders. The dramatic increase in parameter count and training data provides LLMs with richer world knowledge and a more nuanced understanding of semantics compared to their smaller BERT-sized predecessors. This directly translates to performance improvements. 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. Muennighoff (2022); Ma et al. (2024b) empirically verified the effectiveness of scaling the size of dense retrievers with open-weight models such as GPT-J (Wang & Komatsuzaki, 2021) and Llama-2 (Touvron et al., 2023). Today, common text retrieval benchmarks like BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) are dominated by proprietary and open-source LLM-based embedding models (Wang et al., 2023a; Li et al., 2023d; Lee et al., 2025; Zhang et al., 2025b; Muennighoff et al., 2025, *inter alia*).

Decoder-Only Adaptation. A parallel line of research focuses on adapting the unidirectional architecture of decoder-only LLMs (e.g., Llama) to better suit the needs of bidirectional text representation. Standard decoder-only models are optimized for next-token prediction, which may not be ideal for creating a single summary vector for a whole text. To address this, LLM2Vec (BehnamGhader et al., 2024) enables bidirectional attention and further trains Llama-2 (Touvron et al., 2023) with specific unsupervised and supervised adaptive tasks. Similarly, NV-Embed (Lee et al., 2025) introduces a new latent attention mechanism to produce improved representations, leading to improved performance on the MTEB benchmark compared to directly enabling bi-directional attention.

Unified Modeling. GritLM (Muennighoff et al., 2025) fine-tunes Mistral family models with both dense retrieval task and text generation task with different attention mechanisms and demonstrate the potential of unifying retrieval and generation with one single foundation model.

7.2 LLM as Reranker

The reranking task has also been significantly advanced by LLMs, which introduce new capabilities beyond the cross-encoder architecture discussed in Section 6. This evolution can be categorized into two main architectural approaches: fine-tuning and zero-shot prompting. We show a shortlist of works that use LLM as rerankers in Table 6.

[4]The term "Large Language Model" lacks a precise, universally accepted definition in the literature. In this survey, we use the term to refer to models with over one billion parameters that are pre-trained with a text generation objective, such as text infilling (e.g., T5) or causal language modeling (e.g., the GPT series), and are often optionally post-trained for instruction following and human preference alignment.

Fine-Tuned Rerankers. First, LLMs can be fine-tuned as powerful rerankers. Extending earlier work on BERT-type models, researchers have applied similar techniques to larger encoder-decoder models like T5 (Raffel et al., 2020) and decoder-only models like Llama (Touvron et al., 2023).

- **Pointwise and Pairwise:** Nogueira et al. (2020) fine-tuned T5 with a classification loss, treating reranking as a binary relevance decision. RankT5 (Zhuang et al., 2023a) took a more direct approach by fine-tuning T5 to output a numerical relevance score, optimizing the model with established ranking losses like RankNet (Burges, 2010). Further, Zhuang et al. (2023a) also investigated the impact of language model architectures (T5 encoder-decoder versus T5 encoder), loss functions (pointwise, pairwise, listwise), and pooling strategies to ranking performance.
- **Listwise:** Instead of scoring documents individually, ListT5 (Yoon et al., 2024) adopts a Fusion-in-Decoder architecture (Izacard & Grave, 2021) to process and rank an entire list of candidate documents in a single forward pass. More specifically, the architecture consists of an encoder and decoder, where the encoder takes a query and multiple passages as input in parallel, and the decoder outputs a sorted list of input passages in the decreasing order of relevance, achieving better reranking efficiency compared to pointwise methods. Yang et al. (2025b) fine-tuned QwQ-32B (Qwen Team, 2025) for listwise reranking where multiple passages are concatenated together as input and achieved better performance than pointwise reranking, leveraging the reasoning and long context capability of the strong base model.
- **Long-Context Reranking:** The long-context capabilities of modern LLMs have also enabled a new reranking paradigm. RankLlama (Ma et al., 2024b) demonstrated superior pointwise reranking performance compared to BERT and T5-based rerankers for long document reranking where the input is truncated at 4,096 tokens.

Zero-Shot / Few-Shot Prompting. Second, the instruction-following ability of modern LLMs has unlocked zero-shot and few-shot reranking via prompting. This paradigm requires no task-specific fine-tuning. Instead, the LLM is prompted with a query and a list of candidate documents and asked to identify the most relevant ones (Ma et al., 2023; Zhang et al., 2023c; Pradeep et al., 2023b;c; Sun et al., 2023). This listwise approach is a natural fit for the long context windows of models like GPT-4. To mitigate the high computational cost and context length limitations of processing full documents, Liu et al. (2024b) proposed using passage embeddings as compact document representations for the LLM, training a specialized reranker that operates on these embeddings to improve efficiency. As this line of research primarily involves prompt engineering rather than architectural changes, we refer readers to a recent survey (Zhu et al., 2023) for further details.

7.3 Generative Retrieval

Perhaps the most radical architectural shift enabled by LLMs is generative retrieval. Traditional IR systems follow the "retrieve-then-rerank" paradigm (Schütze et al., 2008; Asai et al., 2024a; Xu et al., 2025c). Generative retrieval fundamentally challenges this by reframing retrieval as a sequence-to-sequence task. Instead of searching an index, an autoregressive language model is trained to directly generate the unique document identifiers (DocIDs) of relevant documents in response to a query.

Evolution of the Paradigm. The foundational work in generative retrieval emerged from the entity linking domain with Generation of ENtity REtrieval (GENRE, De Cao et al., 2021). Rather than treating entity retrieval as a classification problem over atomic labels with dense representations, GENRE reframed it as a generative task where an encoder-decoder model produces entity names autoregressively, token-by-token, conditioned on the input context. Building on GENRE's success, DSI introduced generative retrieval to the broader document retrieval domain (Tay et al., 2022). The core innovation was fully parameterizing traditional "retrieve-then-rerank" pipelines within a single neural model, where all corpus information is encoded in the model parameters rather than external indices (Tay et al., 2022; Pradeep et al., 2023a). DSI operates through two fundamental sequence-to-sequence tasks that can be trained jointly or sequentially (He et al., 2024): the indexing task (Learn to Index) and the retrieval task (Learn to Retrieve).

Table 4: Taxonomy of identifier types in Generative Retrieval. The choice of identifier is a key architectural distinction.

Identifier Type	Description
Atomic Identifiers	Unique integers assigned to documents. Simple but lacks semantic generalization.
String Identifiers	Natural language strings (titles, URLs). Leverages pre-trained knowledge but can be ambiguous.
Semantic Identifiers	Structured IDs derived from clustering embeddings. Enables semantic generalization in the ID space.

Table 5: Summary of IR model architecture utilizing large language models as retrieval backbone.

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., 2023a)	LLM Encoder	Mistral	Synthetic Data + Listwise Loss
LLaRA (Li et al., 2023b)	LLM Encoder	Llama	Adaptation + Contrastive Training
MambaRetriever (Zhang et al., 2024)	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., 2023d)	LLM Encoder	Qwen	Adaptation + Synthetic Data + Listwise Loss
Qwen3-Retriever (Zhang et al., 2025b)	LLM Encoder	Qwen3	Synthetic Data + Listwise Loss

Taxonomy of Identifiers. A critical design choice in DSI involves document identifier representation (summarized in Table 4).

- **Atomic and String Identifiers:** Early work explored atomic identifiers (unique integers) and simple string identifiers (titles, URLs) (Chen et al., 2023a). DSI can be implemented in two variants: classification-based approaches that use a classification layer to output atomic document identifiers, and generative approaches that autoregressively generate identifier strings (Mehta et al., 2023). More sophisticated methods have introduced n-gram-based identifiers (Bevilacqua et al., 2022).

- **Semantic Identifiers:** Semantically structured identifiers created through clustering algorithms proved most effective (Zhu et al., 2023). This often involves hierarchical representations using techniques like residual quantization (Zeng et al., 2024), where the model learns to associate document content with corresponding semantically meaningful document identifiers (Kishore et al., 2023).

Inference and Constraints. The technical implementation of generative retrieval systems centers on sequence-to-sequence modeling. A critical technical requirement is ensuring only valid document identifiers are generated during inference. This is typically achieved through constrained beam search over a prefix tree (trie) constructed from all valid document identifiers (Tang et al., 2023). Alternative approaches include constrained greedy search and FM-index structures (Tang et al., 2023). The T5 model backbone (Raffel et al., 2020) serves as the foundation for most DSI implementations, trained with cross-entropy loss on both indexing and retrieval objectives.

Challenges. Generative retrieval faces significant challenges in dynamic environments. The tight coupling between index and retrieval modules makes updating the corpus computationally expensive, typically requiring full model retraining (Mehta et al., 2023). Scalability poses a major challenge; most research has focused on relatively small collections, as the memory and computational requirements grow substantially as corpus size increases. Generative retrieval is an active and rapidly evolving research area; we direct interested readers to a dedicated survey (Li et al., 2025d) for a comprehensive review.

Table 6: 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). We use Seq2Seq LLM to refer to encoder-decoder architecture language models such as T5 and BART, and Casual LLM to refer to modern LLMs with causal language model architecture like GPT family models.

Name	Architecture	Backbone LM	Training / Prompting Strategy
Fine-tune LLM for Reranking			
monoT5 (Nogueira et al., 2020)	Seq2Seq LM	T5	Classification
Nogueira dos Santos et al. (2020)	Seq2Seq LLM	BART	Unlikelihood
QLM-T5 (Zhuang et al., 2021)	Seq2Seq LLM	T5	Language Modeling
duoT5 (Pradeep et al., 2021)	Seq2Seq LLM	T5	Pairwise Loss
RankT5 (Zhuang et al., 2023a)	Seq2Seq LLM Encoder + Prediction Layer	T5	Listwise Loss
ListT5 (Yoon et al., 2024)	Fusion-in-decoder	T5	Listwise Loss
SGPT-CE (Muennighoff, 2022)	Causal LLM	GPT-J & GPT-Neo	Listwise Loss
RankLlama (Ma et al., 2024b)	Casual LLM Encoder + Prediction Layer	Llama	Listwise Loss
RankMamba (Xu, 2024)	Causal LLM Encoder + Prediction Layer	Mamba	Listwise Loss
RankVicuna (Pradeep et al., 2023b)	Causal LLM	Vicuna	Listwise
RankZephyr (Pradeep et al., 2023c)	Causal LLM	Zephyr	Listwise
Zhang et al. (2023c)	Causal LLM	Code-LLaMA-Instruct	Listwise
Liu et al. (2024b)	Embedding + Causal LLM	Mistral	Listwise
Qwen3-Reranker (Zhang et al., 2025b)	Causal LLM	Qwen3	Synthetic Data + Pairwise Loss
Prompt LLM for Reranking			
Zhuang et al. (2023b)	Causal LLM	Multiple	Pointwise Prompting
Zhuang et al. (2024a)	Causal LLM	Flan-PaLM-S	Pointwise Prompting
UPR (Sachan et al., 2022)	Seq2Seq LM & Causal LLM	T5 & GPT-Neo	Pointwise Prompting
PRP (Qin et al., 2024b)	Seq2Seq LM	Flan-UL2	Pairwise Prompting
Yan et al. (2024)	Seq2Seq LM	Flan-UL2	Pairwise Prompting
Zhuang et al. (2024b)	Seq2Seq LM	Flan-T5	Pairwise & Setwise Prompting
LRL (Ma et al., 2023)	Casual LLM	GPT-3	Listwise Prompting
RankGPT-3.5 (Sun et al., 2023)	Causal LLM	GPT-3.5	Listwise Prompting
RankGPT-4 (Sun et al., 2023)	Causal LLM	GPT-4	Listwise Prompting

7.4 Broader Ecosystem and Concluding Remarks

Beyond core architectural changes, LLMs are influencing the entire IR ecosystem. Their advanced generative and understanding capabilities are being harnessed for crucial supporting tasks:

- **Data Synthesis:** Modern IR systems require extensive labeled data for training, which is expensive to create. A promising line of work is to use LLMs to synthesize high-quality training data (e.g., queries, relevant passages, and hard negatives) (Bonifacio et al., 2022; Boytsov et al., 2024; Dai et al., 2023; Lee et al., 2024; Mo et al., 2024a;c; Zhang et al., 2025b).

- **Evaluation:** From an evaluation perspective, LLMs' language understanding has led to research on using them as proxies for human relevance judges, which could dramatically accelerate the evaluation cycle (Faggioli et al., 2023; 2024; Clarke & Dietz, 2024).

We also point readers to a comprehensive survey on conversational information retrieval (Mo et al., 2024b), another area being reshaped by LLMs.

In conclusion, the adoption of LLMs in IR represents more than a simple increase in model scale. While they certainly serve as more powerful feature extractors within existing bi-encoder and cross-encoder frameworks, their unique generative and instruction-following abilities are forging entirely new architectural paradigms like generative retrieval and zero-shot listwise reranking. However, the advancement of IR architecture is not driven solely by the pursuit of superior semantic matching capabilities. The practical deployment of these systems — ranging from lightweight encoders to massive LLMs — necessitates architectures that can withstand rigorous efficiency constraints, handle diverse data modalities, and ensure reliability. We examine these cross-cutting architectural adaptations in the following section.

8 Architectures for Diverse Scenarios and Constraints

The evolution of IR models described in previous sections — from vector space models to LLMs — primarily traces the pursuit of better semantic matching for English text. However, deploying these models in real-world environments requires navigating complex scenarios and constraints beyond pure textual relevance. These include handling diverse data modalities and languages, balancing the inherent tradeoff between accuracy and latency, and ensuring model reliability through calibration. In this section, we review how IR architectures are adapted to meet these specific requirements. Across these settings, architectural choices such as representation granularity and modularity, serve as the primary mechanisms to balance task-specific constraints with scalable retrieval.

8.1 Architectures for Multimodal and Multilingual Data

8.1.1 Multilingual and Cross-Lingual Architectures

Problem Definitions and Retrieval Settings. Although often used interchangeably, Multilingual Information Retrieval (MLIR) and Cross-Lingual Information Retrieval (CLIR) correspond to distinct retrieval scenarios. CLIR refers to the setting in which a user issues a query in a source language and retrieves documents written in a different target language (Oard & Dorr, 1998; Fluhr et al., 1999). This distinction has direct architectural implications: CLIR requires explicit cross-language alignment at query time, whereas MLIR emphasizes building shared or interoperable representations during indexing. The central architectural challenge in CLIR is bridging the lexical and semantic gap between the query and document language spaces (Goworek et al., 2025). In contrast, MLIR is a broader paradigm in which a system indexes and searches over document collections containing multiple languages, potentially serving queries in any of them (Oard & Dorr, 1996; Oard et al., 1999; Fluhr et al., 1999). Architecturally, CLIR can be viewed as a special case of MLIR that explicitly requires cross-language alignment at retrieval time.

Early Translation-Centric Architectures. The architectural foundations of cross-lingual retrieval date back to early work on the Vector Space Model (Salton et al., 1975), where cross-language retrieval was framed as a synonymy problem. Early systems relied on bilingual thesauri to map query terms into a shared conceptual space prior to retrieval, treating translation as a distinct pre-processing step (Salton, 1969; 1970). Architecturally, these systems isolated linguistic complexity into a separate translation component, leaving the retrieval engine itself unchanged and monolingual. When high-quality thesauri were available, such architectures could approach monolingual retrieval performance; however, they were brittle due to limited vocabulary coverage, poor handling of polysemy, and an inability to model multi-word expressions or contextual meaning.

Statistical and Representation-Based Cross-Lingual Models. The availability of large parallel corpora in the 1990s, such as the Canadian Hansards[5], enabled a shift from static dictionaries to statistically grounded architectures. A major departure from direct translation was introduced by Cross-Language Latent Semantic Indexing (CL-LSI), which projected documents and queries from different languages into a shared, language-independent latent space using Singular Value Decomposition over parallel data (Dumais et al., 1997). In parallel, Statistical Machine Translation (SMT) became a dominant architectural component in CLIR systems, with retrieval pipelines adopting either query translation or document translation using probabilistic alignment models such as the IBM Models (Brown et al., 1993). Both CL-LSI and SMT-based pipelines preserved monolingual retrieval backends, differing primarily in whether cross-lingual alignment was achieved through latent semantic projection or probabilistic translation. These systems established the translation-based retrieval paradigm, where the retrieval engine itself remained monolingual and linguistic complexity was isolated within the translation module.

Multilingual Transformers and End-to-End Retrieval. The introduction of multilingual pre-trained transformers, including mBERT and XLM-R, marked a fundamental architectural shift away from explicit

[5] https://en.wikipedia.org/wiki/Hansard

translation toward shared semantic representation learning (Devlin et al., 2019; Conneau et al., 2020a; MacAvaney et al., 2019a). Trained on large-scale multilingual corpora, these models enabled a single encoder to represent queries and documents across languages, substantially reducing dependence on parallel data (Conneau et al., 2020b; Feng et al., 2022; Shi et al., 2020; Goswami et al., 2021). However, general-purpose multilingual encoders often underperform in retrieval settings due to insufficient cross-lingual alignment induced during pretraining (Zhang et al., 2022; Elmahdy et al., 2024), which has been the focus of subsequent research works.

Per-Language Modules. To improve the performance of multilingual pre-trained transformers on low-resource languages, a line of works proposed to add language-specific adapters to enable zero-shot or few-shot cross-lingual transfer. Early explorations in NLP community focus on classical NLP tasks such as dependency parsing and named entity recognition (Üstün et al., 2020; Pfeiffer et al., 2020; Artetxe et al., 2020; Pfeiffer et al., 2021). In the context of IR, Litschko et al. (2022) compared the performance of adapter-based approaches and Sparse-Fine-Tuning Masks (Ansell et al., 2022) to NMT-based approaches and reported their efficacy in both performance and training efficiency.

Alignment-Focused and Modern Architectures. Tackling the same challenge of per-language modules approaches, specialized multilingual retrieval models such as LaREQA, InfoXLM, and LaBSE enforce tighter alignment between semantically equivalent cross-lingual pairs using parallel corpora and contrastive objectives (Roy et al., 2020; Chi et al., 2021; Feng et al., 2022; Huang et al., 2024b). Notably, these approaches typically preserve the standard transformer backbone, focusing architectural innovation on representation learning objectives rather than structural redesign. Recent work further refines end-to-end multilingual retrieval through improved data curation, supervision, and knowledge distillation, continuing the shift away from multi-stage translation pipelines toward unified models that directly match meaning across languages (Zhang et al., 2023d; Yang et al., 2024).

8.1.2 Multimodal Architectures

From Shallow Fusion to Deep Joint Embeddings. Early multimodal retrieval architectures relied on modality-specific feature extraction followed by shallow fusion, with text typically represented using bag-of-words models and images encoded via hand-crafted descriptors such as scale-invariant feature transform (SIFT) (Lowe, 1999). To enable cross-modal comparison, projection-based methods such as canonical correlation analysis (CCA) (Hotelling, 1992) mapped heterogeneous features into a shared embedding space, establishing the foundational principle that multimodal retrieval requires learning semantic correspondences across modalities rather than treating them independently (Rasiwasia et al., 2010; Sharma et al., 2012; Gong et al., 2014; Ranjan et al., 2015). The transition to deep learning replaced manual feature engineering with end-to-end representation learning, exemplified by early visual-semantic embedding models such as DeViSE (Frome et al., 2013; Wang et al., 2016). Sentence encoders incorporating syntactic structure, including Dependency Tree Recursive Neural Networks (DT-RNNs), further improved alignment by modeling relational semantics (Socher et al., 2014). Fragment-level architectures extended this paradigm by decomposing images and sentences into finer-grained units and aligning them with structured max-margin objectives, enabling both global and local cross-modal reasoning (Karpathy et al., 2014). Adversarial frameworks such as ACMR subsequently refined joint embeddings by introducing nonlinear projections and modality-invariant regularization (Wang et al., 2017). We refer to Wang et al. (2024b) for a comprehensive survey.

Dual-Encoder Contrastive Models and Controlled Interaction. Large-scale vision–language datasets enabled a paradigm shift toward bi-encoder architectures trained with contrastive objectives. CLIP and ALIGN independently encoded images and text and aligned them through contrastive learning on hundreds of millions to billions of image–text pairs, establishing scalable and efficient retrieval backbones (Radford et al., 2021; Jia et al., 2021; Wei et al., 2025). This architectural separation facilitated efficient indexing and retrieval and rapidly generalized to additional modalities including video, audio, depth, and sensor data (Girdhar et al., 2023; Chen et al., 2023b; Kong et al., 2025). Subsequent refinements explored the efficiency–expressivity tradeoff within this paradigm, most notably through late-interaction architectures such as ColBERT, which replaced single-vector embeddings with multi-vector representations to enable

token-level matching while preserving much of the efficiency of bi-encoders (Khattab & Zaharia, 2020; Faysse et al., 2025; Wan et al., 2025). Together, these models established a dominant architectural family centered on modality-separated encoding with limited but scalable interaction.

Rich Fusion, Hierarchical Interaction, and Unified Architectures. Beyond bi-encoders, multimodal retrieval architectures increasingly incorporated sophisticated mechanisms to support complex reasoning, particularly in video–text retrieval (Chen et al., 2020a). Transformer-based and hierarchical models decomposed alignment into global-to-local or multi-granular stages, combining cross-modal attention with temporal and semantic structure (Gabeur et al., 2020; Liu et al., 2021; Zhang et al., 2023b; Gorti et al., 2022). Hybrid designs integrated CLIP-style encoders with cross-modal fusion or temporal alignment modules, balancing pretrained representations with task-specific interaction (Portillo-Quintero et al., 2021; Fang et al., 2021). These architectures substantially increase computational cost, often restricting their use to reranking or small candidate sets.

Recent architectures further differentiate between multi-encoder designs that preserve modality-specific encoders and single-encoder models employing full cross-attention at higher computational cost (Li et al., 2019; 2022b; Zhai et al., 2023; Kim et al., 2025). At the same time, multimodal large language models (MLLMs) and multi-agent retrieval frameworks unify retrieval, reasoning, and generation within a single system, while modality-preserving and any-to-any architectures emphasize flexible interaction without collapsing modality structure (Liu et al., 2023; Xie et al., 2024a; Liu et al., 2025b; Xu et al., 2025a; Ju et al., 2025).

Remarks. From a structural perspective, multimodal retrieval architectures can be broadly grouped into: (1) shallow fusion and projection-based models, (2) deep joint-embedding architectures with global or fragment-level alignment, (3) contrastively trained dual-encoder models emphasizing scalability, (4) late-interaction hybrids balancing efficiency and expressivity, and (5) cross-attention and MLLM-centric architectures enabling deep fusion and unified multimodal reasoning. This progression reflects a recurring architectural tension between interaction richness and computational efficiency.

8.2 Performance-Efficiency Tradeoffs and the Role of Architectural Choices

Retrieval systems face a fundamental tradeoff between effectiveness (e.g., recall, MRR, nDCG) and resource efficiency (latency, throughput, memory footprint, and index/update cost). Architectural choices — how queries and documents are represented, how similarity is computed, and how candidates are staged and scored — drive where a method falls on this tradeoff curve. Modern neural retrievers demonstrated that learned dense embeddings can substantially improve effectiveness over classic lexical baselines, but achieving that gain requires extra compute and index engineering compared to lightweight lexical methods.

A common high-performance design is the bi-encoder (single-vector) architecture (Sections 6 and 7): queries and documents are encoded independently into compact vectors and approximate nearest-neighbor search (ANN) retrieves candidates quickly. This architecture is attractive for strict latency budgets and large corpora because it enables highly optimized ANN indexes (e.g., HNSW (Malkov & Yashunin, 2018)) that deliver millisecond-scale queries at large scale; however, single-vector representations can miss fine-grained token-level matches that matter for some queries.

To improve effectiveness, researchers have pursued richer interaction patterns. Late-interaction or multi-vector models (e.g., ColBERT and ColPali (Khattab & Zaharia, 2020; Faysse et al., 2025)) keep token-granular signals and aggregate local token similarities, which raises effectiveness but increases index size and per-query compute; compression and residual quantization techniques can reduce the space/latency penalty but do not eliminate it entirely. Similarly, learned sparse/lexical models (e.g., SPLADE (Formal et al., 2021a;b)) produce high-dimensional but sparse representations that recover lexical signals and interpretability while trading off somewhat higher compute or indexing complexity versus classical BM25. These architectural variants illustrate the spectrum: more expressive interactions often translates to better quality, typically at higher memory and latency cost (unless mitigated by compression).

A pragmatic pattern in production is staged (two- or multi-stage) retrieval: a fast, coarse first-stage (BM25 or compact dense ANN) produces a small candidate set, and a more expensive cross-encoder or interaction-

based reranker refines the top results (Huang et al., 2020; Su et al., 2025). This cascade yields most of the accuracy of expensive models while preserving throughput, but it requires careful budgeting (how many candidates to pass) and engineering (batching, caching, and efficient GPU/CPU placement). Indexing and compression techniques (e.g., product quantization, pruning, residual quantization), along with hybrid lexical–dense pipelines, are commonly employed to navigate the tradeoff curve as operational constraints evolve.

Architectural Design Heuristics. Adopt a single-vector dual-encoder with efficient ANN search when low latency, low cost, and frequent updates are the primary constraints. Favor multi-vector or interaction-intensive models when top-tier ranking quality is required and computational budget permits. To balance these objectives, employ a two-stage pipeline comprising coarse retrieval followed by a specialized reranker. In large-scale deployments, the effectiveness-efficiency operating point is chiefly adjusted through index compression techniques, hybrid lexical-dense signals, and careful calibration of candidate set size.

8.3 Calibration and Confidence Estimation

Calibration in IR aims to align model scores with true probabilities of relevance, such that confidence faithfully reflects correctness. Unlike traditional deterministic rankers that output single relevance scores, calibrated IR models explicitly represent uncertainty, often as a distribution over scores whose mean encodes relevance belief and whose variance captures uncertainty (Cohen et al., 2021). This distinction is architecturally significant: uncertainty-aware scoring exposes information that is otherwise hidden in standard ranking pipelines. The importance of calibration is formalized by the Probability Ranking Principle (PRP), which guarantees optimal ranking only when relevance probabilities are well calibrated and reported with certainty (Penha & Hauff, 2021). However, modern neural rankers frequently violate these assumptions, motivating calibration-aware architectural design.

Neural Architectures and Uncertainty Modeling. Calibration properties in IR are strongly dependent on architectural choices, with empirical studies showing mixed calibration behavior across neural model families (Guo et al., 2017; Minderer et al., 2021). Transformer-based rankers, including BERT variants, are often poorly calibrated, with calibration quality varying by model scale and design (Dan & Roth, 2021; Li et al., 2022a). Introducing stochasticity at the architectural level — through stochastic inference or approximate Bayesian formulations — consistently improves calibration compared to deterministic counterparts, while adding only modest computational overhead (Penha & Hauff, 2021; Cohen et al., 2021). These findings position stochastic and Bayesian architectures as principled mechanisms for embedding uncertainty directly into relevance estimation.

Calibration in Multi-Component Retrieval Systems. Calibration challenges are amplified in multi-stage retrieval architectures, such as the "retrieve-then-rerank" pipeline, where hard top-k retrieval steps break differentiability and preclude end-to-end calibration. Architectural interventions, including differentiable sampling mechanisms based on Gumbel approximations, restore gradient flow and enable joint calibration of retriever and reader components (Dhuliawala et al., 2022). Jointly calibrated systems produce more reliable confidence estimates than calibrating individual modules in isolation, underscoring how calibration requirements can directly shape architectural design in complex IR pipelines.

Modularity and Calibration-Aware Computation. Beyond end-to-end design, modular architectures enable calibration to be treated as an attachable component rather than an intrinsic model property. Universal post-hoc calibrators can be applied across heterogeneous retrieval architectures without modifying their internals, offering scalable and architecture-agnostic improvements (Zhang et al., 2021a). When uncertainty is explicitly modeled, calibration becomes operational rather than merely diagnostic: confidence estimates can guide adaptive computation, such as selective reranking or deferred inference for ambiguous queries, improving both efficiency and robustness (Cohen et al., 2021; Yoon et al., 2025). Recent evidence further suggests that architectural specialization — e.g., modeling relevance across multiple criteria instead of a single scalar score — can inherently reduce calibration error, reinforcing calibration as a first-class archi-

tectural consideration in IR system design (Penha & Hauff, 2021; Javdan et al., 2025). To summarize, how to incorporate calibration into modern retrieval systems is still an open question in the IR community.

8.4 Domain-Specific Applications

General Web Search. General-purpose web search has historically been dominated by the "retrieve-then-rerank" paradigm, built around inverted indexes and multi-stage cascade ranking to ensure efficiency at web scale (Shen et al., 2014a; Mitra et al., 2016; Qin et al., 2022; Zhang et al., 2025a). These architectures decompose retrieval into heterogeneous components for query understanding, candidate generation, ranking, and re-ranking, each optimized independently (Wang & Na, 2023). As web content diversified, search engines incorporated additional signals such as anchor text, hyperlinks, and layout-based features to enrich document representations beyond plain text (Oliveira & Teixeira Lopes, 2023). While highly effective, these pipelines impose rigid, predefined information flows that limit adaptability. The recent generative retrieval (Section 7.3) thus explores replacing this modular pipeline with model-centric architectures, where a single LLM performs indexing, retrieval, and ranking end-to-end.

Domain-Specific and Thematic Retrieval. Domain-specific search engines target focused corpora such as scientific literature, medicinal chemistry databases, and job postings, leveraging domain knowledge to improve relevance beyond general web search. These applications share architectural challenges arising from specialized terminology, underspecified expert queries, and narrowly scoped user intents (Wang & Na, 2023; Kang et al., 2024). Generic pretrained models often fail to capture domain-specific semantics without architectural support for structured knowledge. Traditional solutions rely on query expansion, lexical analysis, and large task-specific training sets, but suffer from scalability and complexity limitations (Xu et al., 2025f). In response, modern architectures increasingly integrate entity- and relation-centric representations to better align retrieval with domain semantics (Dong et al., 2022a). As a result, domain-specific IR architectures emphasize knowledge-aware representations and domain adaptation over purely generic pretrained models.

Medical and Legal Information Retrieval. Medical and legal search systems represent high-stakes instantiations of domain-specific retrieval, imposing stricter architectural constraints. Medical IR must operate over long temporal records that combine unstructured clinical notes with structured diagnoses, laboratory values, and medications, motivating architectures that integrate information extraction, faceted search, and structured database querying (Sonntag & Profitlich, 2018). These systems must additionally handle specialized medical terminology, clinical documentation standards, and stringent privacy requirements. Legal IR faces analogous challenges due to highly specialized language, complex document structures, and limited labeled relevance data (Althammer et al., 2020). Despite recent efforts to adapt pretrained language models such as BERT, neural approaches often fail to consistently outperform strong lexical baselines like BM25, reinforcing the continued relevance of hybrid and lexically grounded architectures in legally specialized domains (de Araujo et al., 2014; Althammer et al., 2020).

E-commerce Search. E-commerce search systems operate over structured product catalogs rather than unstructured web pages, fundamentally shaping their architectures (Wang & Na, 2023; Ren et al., 2024). The dominant approach employs embedding-based retrieval using the bi-encoder architecture combined with approximate nearest neighbor search (Nigam et al., 2019; Lin et al., 2024). These systems must address severe vocabulary mismatch between customer queries and product descriptions, motivating query rewriting and semantic bridging components, often implemented using LLMs trained with contrastive or instruction-based objectives (Peng et al., 2024b).

Ads Ranking. Ads ranking is another important industrial IR application scenario. Modern advertising recommendation systems face the fundamental challenge of processing massive candidate sets while maintaining strict latency requirements and ranking quality. The traditional approach employs multi-stage cascading architectures that decompose the Ad ranking problem into sequential stages: retrieval, pre-ranking (or lightweight ranking), ranking (or heavyweight ranking), and auction (Gallagher et al., 2019; Wang et al., 2023b; Zheng et al., 2024a; Yang et al., 2025d), where the research focus is similar to that of ad hoc retrieval.

In the context of Ads ranking, the XMC formulation described in Section 5 has been shown to improve ranking quality of tail items/Ads (Jain et al., 2019; Dahiya et al., 2021; Yu et al., 2022a; Dahiya et al., 2023; Gupta et al., 2024b), which is often critical for revenue maximization. However, it also comes with higher engineering complexity and often requires retraining of the models when new items/Ads are added.

The recent breakthrough of LLMs has catalyzed a paradigm shift toward generative recommendation frameworks that can directly generate personalized item sequences from user interaction histories. These approaches index items with meaningful IDs using vector quantization algorithms and generate items from the entire item set for recommendation, conceptualizing online advertising systems as a unified generative process that eliminates inherent goal conflicts between different pipeline stages (Rajput et al., 2023; Zheng et al., 2023; Zhai et al., 2024). Given space constraints, we refer readers to the cited works for a more in-depth treatment of the field.

Cross-Application IR. Across applications, several architectural trends are reshaping IR system design. Retrieval-augmented generation (RAG) architectures combine dense vector search with LLMs to support search, reasoning, and synthesis over structured and unstructured data (Asai et al., 2024c). Modular and multi-agent retrieval frameworks further decouple retrieval functions into interoperable components, enabling dynamic adaptation to application-specific requirements (Chang et al., 2025; Zhang et al., 2025a). Finally, contextual personalization and the convergence of search and recommendation architectures reflect a broader shift toward unified representations of users, documents, and intent (Zamani & Croft, 2018).

Taken together, the diversity of application-driven architectures underscores that modern IR systems are no longer optimized solely for static relevance ranking, but must increasingly adapt to new roles, constraints, and integration patterns that give rise to emerging architectural directions and open challenges.

9 Emerging Directions and Challenges

IR systems have become crucial across diverse domains, from retrieval-augmented language modeling (Khandelwal et al., 2020; Borgeaud et al., 2022) to applications in agents (Wu et al., 2023a; Wang et al., 2024a; 2025a), code generation (Wang et al., 2025b; 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 (Sections 3 to 8), we examine emerging trends, open problems, and potential research directions. Our discussion is structured around three key areas: advancing the core components of IR models, adapting to the new paradigm of retrieval for AI, and tackling the pragmatic challenges of real-world deployment.

9.1 Advancing the Core Components of IR Models

At the heart of any IR system are the models that extract features and estimate relevance. As IR moves toward more compute-intensive practices, we identify key areas for improving these fundamental components.

More Powerful and Efficient Foundation Models. 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*). However, for IR to leverage this trend sustainably, several challenges in model design and training must be addressed:

- **Data and 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 or in a few-shot/zero-shot setting remains crucial. Additionally, models should support parallel processing and low-precision training to reduce costs and accelerate convergence (Pool et al., 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 & Zettlemoyer, 2023; Kumar et al., 2024; Xu et al., 2024b; Warner et al., 2025; Bruch et al., 2024; Xu et al., 2025b, *inter alia*).

- **Better lite foundation models.** IR models need to process queries in real time, and using compact-sized foundation models is often a practical solution. Warner et al. (2025) presented an interesting study on wide-and-shallow versus deep-and-narrow architecture in the context of training a "modern" BERT model. Recent works (Günther et al., 2024; Fu et al., 2023; Portes et al., 2023) studied training efficient BERT model that supports longer context length, while Nussbaum & Duderstadt (2025) investigated the efficacy of a mixture-of-expert BERT-style encoder model. The best recipe for lite foundational models for IR applications remains an open question.

- **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; 2024a; Qin et al., 2024c), state space models (Gu & Dao, 2024; Dao & Gu, 2024), and linear attention (Katharopoulos et al., 2020; Yang et al., 2025c) offer alternatives with theoretical linear complexity. Although preliminary studies (Xu, 2024; Zhang et al., 2024; Xu et al., 2025e;c) show limited gains compared to optimized transformers, developing efficient alternative architectures for transformers could revolutionize large-scale IR.

Ultimately, strong foundation models have proven crucial for IR performance (Neelakantan et al., 2022; Ma et al., 2024b, *inter alia*). As IR applications expand, developing foundation models that balance computational efficiency with robust performance across tasks and modalities emerges as a key research priority.

Flexible and Scalable Relevance Estimators. As discussed in Section 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 & 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.

9.2 The Shifting Paradigm: From Search for Humans to Retrieval for AI

The integration of IR systems into other research domains presents new challenges. We discuss key implications for future IR modeling research as the primary "user" of retrieval shifts from humans to AI models.

The 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), tool usage (Schick et al., 2023; Patil et al., 2024b; Qin et al., 2024a; Patil et al., 2024a) and planning (Song et al., 2023). This shifting paradigm raises fundamental questions about task formulation, evaluation, and system optimization:

- Current IR research is grounded in human information-seeking behavior (Wilson, 2000; Marchionini, 2006; White & Roth, 2009, *inter alia*). When the end user becomes another ML model, we must reconsider how to define and assess *relevance*. For example, a document might be irrelevant to a human but contain a key factual nugget that an LLM needs to answer a question. This 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 (Lewis et al., 2020b; Petroni et al., 2021; Asai et al., 2024c). Future IR models should support end-to-end system optimization rather than focusing solely on ranking metrics (OpenAI, 2025; Huang et al., 2025).

Retrieval Augmented Generation. Retrieval Augmented Generation (RAG) refers to architectures that combine an external retrieval module with a generative model, enabling the system to access and condition on external knowledge before producing output. Early RAG formulations separate retrieval and generation as distinct components: a dense or sparse retriever extracts relevant passages for a given input, and a generator (usually an LLM) conditions on those retrieved results to produce text (Lewis et al., 2020b; Ram et al., 2023b; Mallen et al., 2023; Asai et al., 2024c; Chen et al., 2020b; Xu & Cohen, 2023). This architectural decoupling allows retrieval to be optimized independently from generation, which is useful for modularity and scalability. A typical RAG pipeline comprises:

- Corpus embedding and indexing (vector stores or sparse indexes) for fast retrieval;
- Retriever model (dense vectors or hybrid techniques) that returns top-k relevant items;
- Fusion mechanism that combines retrieved content with the query (e.g., concatenation or cross-attention) before passing it to the generator;
- Generator that produces the final answer based on fused context.

This simple architecture improves factual accuracy and reduces hallucination by grounding generation on external context. However, it also introduces key engineering and modeling considerations:

- **Retriever-Generator Coupling**: The architectural choice of where and how to combine retrieval results can affect quality and efficiency. Early approaches rely on simple concatenation of text passages into the generator's context window (Lewis et al., 2020b; Ram et al., 2023b), which do not alter the architecture of the generator. More advanced designs use cross-attention layers or rerankers to improve precision (Izacard & Grave, 2021; Park et al., 2023; An et al., 2025b).
- **Multi-Modal and Structured Retrieval**: Recent work explores RAG variants that extend beyond plain text. For example, GraphRAG and related graph-enhanced frameworks integrate structured knowledge sources, enabling retrieval of relational paths not easily captured by vector similarity alone (Edge et al., 2025; Li et al., 2025a). These architectures introduce additional components like graph encoders and reasoning layers, allowing the system to extract multi-hop or entity-centric context before generation.
- **Adaptive Retrieval**: Newer RAG pipelines incorporate reasoning capabilities into the generator, to dynamically decide whether to retrieval or to generate answer (Trivedi et al., 2023; Jiang et al., 2023b; Asai et al., 2024b).

These architectural choices have direct implications: strong retrieval can reduce generation errors but may increase pipeline complexity and latency; graph-aware models can support structured reasoning at the cost of larger indexes and additional components; adaptive retrieval strategies can improve effectiveness but require careful task planning and retrieval coordination.

Overall, RAG architectures represent a spectrum from simple retrieval-plus-generation pipelines to complex multi-component systems that integrate heterogeneous knowledge representations and adaptive retrieval planning to improve both relevance and reasoning capability. We refer readers to (Gao et al., 2023; Li et al., 2025e; Mei et al., 2025) for more comprehensive reviews of this topic.

The Rise of Autonomous Search Agents. Complex tasks often 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., 2025a; Ravfogel et al., 2024) and reasoning (Su et al., 2024; Shao et al., 2025; Liang et al., 2025). This has spurred the development of autonomous search agents. Existing attempts can be divided into two main directions. One line of work focuses on training more capable retrievers and rerankers. This involves creating new data pipelines to synthesize instruction-following and reasoning-focused training data (Oh et al., 2024; Weller et al., 2024; Shao et al., 2025, *inter alia*) and leveraging

strong backbone language models such as the proprietary GPT-4 family models (OpenAI et al., 2023) and open-weight Qwen (Yang et al., 2025a) to imbue the retriever with reasoning capabilities.

Another line of work treats the search/retrieval system as a static tool and focuses on improving a large reasoning model's (LRM) ability to use the retrieval tool to improve reasoning and handle in-context memory (Jin et al., 2025; Yu et al., 2025; Gao et al., 2025; Xu et al., 2025d, *inter alia*). In this setup, the LRM decides when, where, and how to conduct a search, and the results subsequently influence its further reasoning and decision-making (Nakano et al., 2021; He et al., 2025). This line of work, commonly referred to as **Deep Research Agents**, represents a paradigm shift from *retrieval as the end goal* to *retrieval as tools for LLMs*. Formally, we use the definition from Huang et al. (2025): deep research agents refer to AI agents powered by LLMs, integrating dynamic reasoning, adaptive planning, multi-iteration external data retrieval and tool use, and comprehensive analytical report generation for informational research tasks.

From the retrieval perspective, the retrieval usage of deep research agents can be broadly categorized into two types: (1) API-based search engines which interact with structured data sources, such as search engine APIs or scientific database APIs; (2) browser-based search engines, which simulate human-like interactions with web pages and enable real-time extraction of dynamic or unstructured content leveraging LLMs' long-context, code understanding and multimodality capabilities.

Both formulations have strengths and weaknesses. API-based retrieval, in line with traditional IR literature, is a fast, efficient, structured and scalable method to enable deep research agents' access to external knowledge (Schütze et al., 2008; Singh et al., 2025b). Browser-based retrieval (Nakano et al., 2021; OpenAI, 2025) offers the advantage of simulating real-time, interactive information-seeking behavior — conceptually similar to that of human users. However, it incurs additional latency and token consumption, and introduces further complexity arising from the web-browsing environment. Given the current landscape of retrieval integration within deep research agents, a central open question is how to design a hybrid retrieval architecture that combines the strengths of both paradigms to achieve a superior performance-efficiency tradeoff.

From the model training perspective, recent efforts have abstracted away the details of retrieval models, treating retrieval as static tools and instead focusing on improving the capabilities of LLM-based search agents. For example, many recent efforts aim to train LRMs to use search tools more effectively via reinforcement learning or specialized fine-tuning (Li et al., 2025b; Jin et al., 2025; Li et al., 2025c; Chen et al., 2025; Wu et al., 2025b;a, *inter alia*). This approach is central to agentic frameworks that orchestrate tool use for complex task completion (Wu et al., 2023a; Shinn et al., 2023; Asai et al., 2024b). Despite this exciting progress, key limitations remain. To enable retrievers' reasoning capability often requires strong backbone models (e.g., 7B scale), which can be 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 sometimes 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 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 remain to be studied. We refer readers to (Singh et al., 2025a; Liang et al., 2025; Xi et al., 2025; Lin et al., 2025) for more comprehensive reviews of this topic.

9.3 Retrieval Beyond Simple Relevance: Instruction-Following and Reasoning-Aware Retrieval

Instruction-Following Retrieval. Instruction-following retrieval extends the standard query-document formulation by conditioning the retriever on a natural-language instruction. Architecturally, this is typically implemented by jointly encoding the instruction and the query in a shared bi-encoder, for example through simple concatenation or lightweight attention layers, so that the query representation adapts to task intent (Weller et al., 2025a; Oh et al., 2024). Training objectives often use contrastive learning with instruction-conditioned positives and hard negatives, allowing the model to capture fine-grained task semantics without large cross-encoder computations (Weller et al., 2024). Instruction-aware retrievers benefit from curated or synthetic datasets pairing instructions, queries, and relevant passages, which improves robustness to paraphrasing and query phrasing variations. These design choices preserve scalability while improving performance on instruction-heavy tasks. Hybrid pipelines may further combine instruction-conditioned re-

trieval with downstream rerankers or fusion mechanisms, ensuring that retrieved evidence aligns with the specific needs of generation or reasoning tasks (Shao et al., 2024; Weller et al., 2025b).

Reasoning-Aware Retrieval. Reasoning-aware retrieval aims to retrieve evidence that supports multi-step inference and complex reasoning (Su et al., 2024). One line of work uses **graph-based architectures**, where the retriever constructs and encodes relational graphs or multi-hop chains of passages. These components — graph construction, encoding, path selection, and ranking — allow the retriever to return connected evidence that reflects reasoning dependencies (Liu et al., 2025a; Edge et al., 2025). While effective for compositional queries, these approaches increase system complexity, latency, and index size.

A different line of work, exemplified by ReasonIR (Shao et al., 2025) and RankR1 (Zhuang et al., 2025), improves reasoning capability through *data curation and task-aligned training* rather than altering the underlying "retrieve-then-rerank" pipeline. By generating multi-step reasoning datasets or instruction-guided examples, these methods teach standard retrievers and rerankers to prioritize evidence that is useful for downstream reasoning, achieving stronger performance without introducing graph-based components. From an evaluation perspective, reasoning-aware retrieval highlights the need for task-centered metrics that measure downstream reasoning success, emphasizing the tight connection between architectural or training choices and reasoning effectiveness.

9.4 Deployment, Robustness, and Trustworthiness

As modern IR systems become more powerful and integrated into high-stakes applications, ensuring their practical deployability and reliability is paramount.

The Efficiency-Effectiveness Tradeoff at Scale. Traditional retrieval systems face significant challenges when scaling to web-scale document corpus, and deploying such systems requires a blend of science and engineering expertise (Dean, 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; Nayak, 2019; OpenAI, 2024; Google, 2024, *inter alia*). These applications often require multiple rounds of retrieval and operate on dynamic corpuses, urging for efficient and effective retrieval. Mainstream inference optimization frameworks such as vLLM (Kwon et al., 2023) and SGLang (Zheng et al., 2024b) have provided support for embedding models. From the modeling perspective, an open question is to design and pre-train models explicitly for retrieval purposes (Warner et al., 2025; Nussbaum et al., 2025; Günther et al., 2024).

Ensuring Robustness in a Noisy and Adversarial World. We discuss a few challenges in IR models' deployment in noisy environments, 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) showed that neural retrievers are biased towards AI-generated documents. Xu et al. (2024a) showed 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 ad hoc retrieval and RAG LLM safety have discussed BERT-based models' brittleness to adversarial attacks (Wang et al., 2022b), as well as the threat of corpus poisoning where injected harmful documents lead to unsafe RAG system 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., 2025a), documents that contain biases and toxic materials can potentially jailbreak aligned LLMs. This observation highlights the importance for IR models to be robust to bias and toxic content.

- **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., 2025; 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. Several works have reported cross-encoder rerankers' performance drops on out-of-domain datasets (Mokrii et al., 2021; Thakur et al., 2021). In the context of retrieval-augmented generation, Cao et al. (2025) conducted a rigorous benchmarking, and found that 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 out-of-distribution (OOD) input (Gupta et al., 2024a).

We refer readers for more detailed discussions on IR models' robustness to dedicated surveys (Asai et al., 2024c; Liu et al., 2025c; Zhou et al., 2024). Addressing these robustness issues at the model architecture level is a critical and underexplored direction for future research.

10 Conclusions and Closing Thoughts

The journey of IR model architectures, as we have charted, is a story of escalating abstraction and semantic depth. Beginning with the foundational principles of term-based matching in Boolean, vector space, and probabilistic models, the field systematically evolved. Learning-to-Rank introduced the power of machine learning to combine diverse statistical features, but it was the advent of neural networks that marked the first major leap toward semantic understanding. These neural ranking models, with their ability to learn representations directly from text, began to bridge the lexical gap that had long constrained traditional methods. The arrival of pre-trained transformers like BERT then catalyzed a paradigm shift, providing a powerful, universal foundation for both highly-effective cross-encoder rerankers and efficient bi-encoder retrievers. Most recently, the ascent of LLMs has not only scaled up these existing architectures but has also introduced entirely new paradigms, such as generative retrieval and zero-shot listwise reranking, fundamentally reshaping what a retrieval system can do.

Throughout this evolution, a core architectural tension has persisted: the tradeoff between effectiveness and efficiency. This is the fundamental conflict between deep, fine-grained interaction models (like interaction-based neural rankers and cross-encoders) that offer high accuracy, and scalable representation-based models (like vector space models, dense retrievers) that enable fast, pre-computable search over massive collections. The enduring "retrieve-then-rerank" pipeline is a direct architectural answer to this tradeoff. Innovations like multi-vector models (e.g., ColBERT) and hybrid sparse-dense systems represent sophisticated attempts to find a better balance on this spectrum, pushing the Pareto frontier of what is possible.

Today, we stand at another inflection point. The primary "user" of IR is shifting from a human at a search bar to an AI model within a larger system. IR is no longer just a tool for finding documents; it is becoming a critical cognitive component for retrieval-augmented generation, autonomous agents, and complex reasoning systems. This shift, as outlined in Section 9, forces us to re-evaluate our core assumptions. Relevance is no longer solely about satisfying human information needs but about providing the precise factual or contextual information an AI needs to complete a downstream task. This demands new model architectures that are not only powerful and efficient but also instruction-aware, contextually flexible, and seamlessly integrable into end-to-end differentiable systems.

Looking ahead, the future of IR model architecture will be defined by its ability to meet these new demands. The grand challenges lie in building foundation models that are multimodal, multilingual, and computationally sustainable; in designing systems that are robust, trustworthy, and resistant to adversarial manipulation; and in developing autonomous search agents that can reason, plan, and interact with the world's information on our behalf. As IR becomes ever more deeply embedded in the fabric of artificial intelligence, its continued evolution will be crucial, not just for the future of search, but for the future of intelligent systems themselves.

Acknowledgments

This work was partially supported by grants from NSF (DMS 2134223 and IIS 2205418).

References

Qingyao Ai, Keping Bi, Jiafeng Guo, and W. Bruce Croft. Learning a deep listwise context model for ranking refinement. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 135–144, 2018a.

Qingyao Ai, Jiaxin Mao, Yiqun Liu, and W. Bruce Croft. Unbiased learning to rank: Theory and practice. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 2305–2306, New York, NY, USA, 2018b. Association for Computing Machinery. doi: 10.1145/3269206.3274274.

Qingyao Ai, Xuanhui Wang, Sebastian Bruch, Nadav Golbandi, Michael Bendersky, and Marc Najork. Learning groupwise multivariate scoring functions using deep neural networks. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 85–92, 2019.

Sophia Althammer, Sebastian Hofstätter, and Allan Hanbury. Cross-domain retrieval in the legal and patent domains: a reproducability study. In *European Conference on Information Retrieval*, 2020.

Bang An, Shiyue Zhang, and Mark Dredze. RAG LLMs are not safer: A safety analysis of retrieval-augmented generation for large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 5444–5474, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. URL https://aclanthology.org/2025.naacl-long.281/.

Yuwei An, Yihua Cheng, Seo Jin Park, and Junchen Jiang. HyperRAG: Enhancing quality-efficiency tradeoffs in retrieval-augmented generation with reranker kv-cache reuse. *arXiv preprint arXiv:2504.02921*, 2025b.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. Composable sparse fine-tuning for cross-lingual transfer. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1778–1796, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.125. URL https://aclanthology.org/2022.acl-long.125/.

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

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SyK00v5xx.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421/.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. Task-aware retrieval with instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3650–3675, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.225. URL https://aclanthology.org/2023.findings-acl.225/.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. OpenScholar: Synthesizing scientific literature with retrieval-augmented LMs. *arXiv preprint arXiv:2411.14199*, 2024a.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The 12th International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=hSyW5go0v8.

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

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press; Addison-Wesley, 1st edition, 1999.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

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

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

Soyuj Basnet, Jerry Gou, Antonio Mallia, and Torsten Suel. Deeperimpact: Optimizing sparse learned index structures. *arXiv preprint arXiv:2405.17093*, 2024.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In *1st Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=IW1PR7vEBf.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Adam Berger and John Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 222–229, 1999.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. Autoregressive search engines: Generating substrings as document identifiers. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Z4kZxAjg8Y.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 405–414, 2018.

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

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. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pp. 2206–2240. PMLR, 2022.

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

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30:107–117, 1998. URL http://www-db.stanford.edu/~backrub/google.html.

Andrei Z. Broder, David Carmel, Michael Herscovici, Aya Soffer, and Jason Zien. Efficient query evaluation using a two-level retrieval process. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, pp. 426–434, 2003.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sebastian Bruch, Masrour Zoghi, Michael Bendersky, and Marc Najork. 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*, pp. 1241–1244, 2019.

Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. 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*, pp. 152–162, 2024.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 535–541, 2006.

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

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

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

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

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

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2607–2622, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.131. URL https://aclanthology.org/2025.acl-long.131/.

Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin. Training and testing low-degree polynomial data mappings via linear svm. *Journal of Machine Learning Research*, 11 (48):1471–1490, 2010. URL http://jmlr.org/papers/v11/chang10a.html.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, M. de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. Continual learning for generative retrieval over dynamic corpora. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023a.

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

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10635–10644, 2020a.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Ming-Ting Sun, Xinxin Zhu, and J. Liu. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. *ArXiv preprint arXiv:2305.18500*, 2023b.

Wei-Fan Chen, Shahbaz Syed, Benno Stein, Matthias Hagen, and Martin Potthast. Abstractive snippet generation. In *Proceedings of the Web Conference 2020*, WWW '20, pp. 1309–1319, New York, NY, USA, 2020b. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380206. URL https://doi.org/10.1145/3366423.3380206.

Xuanang Chen, Ben He, Kai Hui, Le Sun, and Yingfei Sun. Simplified TinyBERT: Knowledge distillation for document retrieval. In *European Conference on Information Retrieval*, pp. 241–248. Springer, 2021.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3576–3588, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL https://aclanthology.org/2021.naacl-main.280/.

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

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? An Analysis of BERT's Attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828/.

Charles L. A. Clarke and Laura Dietz. LLM-based relevance assessment still can't replace human relevance assessment. *arXiv preprint arXiv:2412.17156*, 2024.

Daniel Cohen and W. Bruce Croft. End to end long short term memory networks for non-factoid question answering. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, pp. 143–146, New York, NY, USA, 2016. Association for Computing Machinery. doi: 10.1145/2970398.2970438.

Daniel Cohen, Bhaskar Mitra, Oleg Lesota, Navid Rekabsaz, and Carsten Eickhoff. Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(7), 2011.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020a. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536. URL https://aclanthology.org/2020.acl-main.536/.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 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, pp. 758–759, New York, NY, USA, 2009. Association for Computing Machinery. doi: 10.1145/1571941.1572114.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Kunal Dahiya, Ananye Agarwal, Deepak Saini, Jian Jiao, Amit Singh, Sumeet Agarwal, Purushottam Kar, Manik Varma, et al. Siamesexml: Siamese networks meet extreme classifiers with 100m labels. In *International Conference on Machine Learning*, pp. 2330–2340. PMLR, 2021.

Kunal Dahiya, Sachin Yadav, Sushant Sondhi, Deepak Saini, Sonu Mehta, Jian Jiao, Sumeet Agarwal, Purushottam Kar, and Manik Varma. Deep encoders with auxiliary parameters for extreme classification. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 358–367, 2023.

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

Zhuyun Dai and Jamie Callan. 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*, 2019a.

Zhuyun Dai and Jamie Callan. Context-aware sentence/passage term importance estimation for 1st stage retrieval. *arXiv preprint arXiv:1910.10687*, 2019b.

Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, WSDM '18, pp. 126–134, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3159652.3159659.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. In *The 11th International Conference on Learning Representations*, 2023.

Soham Dan and Dan Roth. On the effects of transformer size on in- and out-of-domain calibration. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2096–2101, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.180. URL https://aclanthology.org/2021.findings-emnlp.180/.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *Proceedings of the 41st International Conference on Machine Learning*, pp. 10041 – 10071, 2024.

Arpan Dasgupta, Siddhant Katyan, Shrutimoy Das, and Pawan Kumar. Review of extreme multilabel classification. *arXiv preprint arXiv:2302.05971*, 2023.

Denis Andrei de Araujo, Carolina Müller, Rove Chishman, and Sandro José Rigo. Information extraction for legal knowledge representation–a review of approaches and trends. *Revista Brasileira de Computação Aplicada*, 6(2):2–19, 2014.

Nicola De Cao, G. Izacard, S. Riedel, and F. Petroni. Autoregressive entity retrieval. In *ICLR 2021-9th International Conference on Learning Representations*, volume 2021. ICLR, 2021.

Jeffrey Dean. Challenges in building large-scale information retrieval systems: invited talk. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, 2009. doi: 10.1145/1498759.1498761.

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

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

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Shehzaad Dhuliawala, Leonard Adolphs, Rajarshi Das, and Mrinmaya Sachan. Calibration of machine reading systems at scale. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1682–1693, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.133. URL https://aclanthology.org/2022.findings-acl.133/.

Qian Dong, Yiding Liu, Suqi Cheng, Shuaiqiang Wang, Zhicong Cheng, Shuzi Niu, and Dawei Yin. Incorporating explicit knowledge in pre-trained language models for passage re-ranking. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022a.

Qian Dong, Shuzi Niu, Tao Yuan, and Yucheng Li. Disentangled graph recurrent network for document ranking. *Data Science and Engineering*, 7:30 – 43, 2022b.

Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, volume 15, pp. 21, 1997.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2025.

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. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Adel Elmahdy, Sheng-Chieh Lin, and Amin Ahmad. Synergistic approach for simultaneous optimization of monolingual, cross-lingual, and multilingual information retrieval. *arXiv preprint arXiv:2408.10536*, 2024.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. doi: 10.1016/0364-0213(90)90002-E.

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

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

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. CLIP2Video: Mastering video-text retrieval via image CLIP. *ArXiv preprint arXiv:2106.11097*, 2021.

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

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2025.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62/.

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

Christian Fluhr, Robert E. Frederking, Doug Oard, Akitoshi Okumura, Kai Ishikawa, and Kenji Satoh. Multilingual (or cross-lingual) information retrieval. *Proceedings of the Multilingual Information Management: Current Levels and Future Abilities*, pp. 10–13, 1999.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*, 2021a.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE: Sparse lexical and expansion model for 1st stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2288–2292, New York, NY, USA, 2021b. Association for Computing Machinery. doi: 10.1145/3404835.3463098.

Thibault Formal, Stéphane Clinchant, Hervé Déjean, and Carlos Lassance. SPLATE: Sparse late interaction retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2635–2640, 2024. doi: 10.1145/3626772.3657968.

Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pp. 23–37. Springer, 1995.

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

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

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Dan Fu, Simran Arora, Jessica Grogan, Isys Johnson, Evan Sabri Eyuboglu, Armin Thomas, Benjamin Spector, Michael Poli, Atri Rudra, and Christopher Ré. Monarch Mixer: A simple sub-quadratic GEMM-based architecture. *Advances in Neural Information Processing Systems*, 36:77546–77603, 2023.

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

Valentin Gabeur, Chen Sun, Alahari Karteek, and Cordelia Schmid. Multi-modal transformer for video retrieval. *arXiv preprint arXiv: 2007.10639*, 2020.

Luke Gallagher, Ruey-Cheng Chen, Roi Blanco, and J. Shane Culpepper. Joint optimization of cascade ranking models. *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, 2019.

Jianfeng Gao, Patrick Pantel, Michael Gamon, Xiaodong He, and Li Deng. Modeling interestingness with deep neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2–13, 2014.

Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl, 2025. URL https://arxiv.org/abs/2508.07976.

Luyu Gao and Jamie Callan. Condenser: a pre-training architecture for dense retrieval. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 981–993, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.75. URL https://aclanthology.org/2021.emnlp-main.75/.

Luyu Gao and Jamie Callan. Long document re-ranking with modular re-ranker. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

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

Luyu Gao, Zhuyun Dai, and Jamie Callan. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3030–3042, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.241. URL https://aclanthology.org/2021.naacl-main.241/.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552/.

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

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

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind one embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014.

Google. Grounding with google search. https://ai.google.dev/gemini-api/docs/grounding?lang=python, 2024.

Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-Pool: Cross-modal language-video attention for text-video retrieval. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4996–5005, 2022.

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John P. McCrae. Cross-lingual sentence embedding using multi-task learning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9099–9113, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.716. URL https://aclanthology.org/2021.emnlp-main.716/.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

Roksana Goworek, Olivia Macmillan-Scott, and Eda B. Özyiğit. Bridging language gaps: Advances in cross-lingual information retrieval with multilingual LLMs. *arXiv preprint arXiv:2510.00908*, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *1st Conference on Language Modeling*, 2024.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, Maximilian Werk, Nan Wang, and Han Xiao. Jina Embeddings 2: 8192-token general-purpose text embeddings for long documents. arXiv preprint arXiv:2310.19923, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 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, pp. 55–64, New York, NY, USA, 2016a. Association for Computing Machinery. doi: 10.1145/2983323.2983769.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 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, pp. 701–710, New York, NY, USA, 2016b. Association for Computing Machinery. doi: 10.1145/2983323.2983768.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067, 2020.

Ashim Gupta, Rishanth Rajendhran, Nathan Stringham, Vivek Srikumar, and Ana Marasovic. Whispers of doubt amidst echoes of triumph in NLP robustness. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5533–5590, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.310. URL https://aclanthology.org/2024.naacl-long.310/.

Nilesh Gupta, Fnu Devvrit, Ankit Singh Rawat, Srinadh Bhojanapalli, Prateek Jain, and Inderjit S. Dhillon. Dual-encoders for extreme multi-label classification. In *The 12th International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=dNe1T0Ahby.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The journal of machine learning research*, 13(1):307–361, 2012.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pp. 3929–3938. PMLR, 2020.

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

Hua He, Kevin Gimpel, and Jimmy Lin. Multi-perspective sentence similarity modeling with convolutional neural networks. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1576–1586, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1181. URL https://aclanthology.org/D15-1181/.

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

Zhankui He, Zhouhang Xie, Harald Steck, Dawen Liang, Rahul Jha, Nathan Kallus, and Julian McAuley. Reindex-then-adapt: Improving large language models for conversational recommendation. *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, 2024.

Djoerd Hiemstra and Wessel Kraaij. Twenty-one at TREC-7: Ad-hoc and cross-language track. In *7th Text REtrieval Conference, TREC-7 1998*, pp. 227–238. National Institute of Standards and Technology, 1999.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015.

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

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. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

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

Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and Allan Hanbury. 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*, pp. 2021–2024, 2020b.

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

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 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*, pp. 113–122, 2021a.

Sebastian Hofstätter, Bhaskar Mitra, Hamed Zamani, Nick Craswell, and Allan Hanbury. Intra-document cascading: Learning to select passages for neural document ranking. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021b.

Sebastian Hofstätter, Omar Khattab, Sophia Althammer, Mete Sertkan, and Allan Hanbury. 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*, pp. 737–747, 2022.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in Statistics: Methodology and Distribution*, pp. 162–190. Springer, 1992.

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

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 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, pp. 2042–2050, Cambridge, MA, USA, 2014. MIT Press.

Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Qiguang Chen, et al. OWL: Optimized workforce learning for general multi-agent assistance in real-world task automation. *arXiv preprint arXiv:2505.23885*, 2025.

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

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

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 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, pp. 2333–2338, New York, NY, USA, 2013. Association for Computing Machinery. doi: 10.1145/2505515.2505665.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of LLM agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024a.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025.

Zhiqi Huang, Puxuan Yu, Shauli Ravfogel, and James Allan. Language concept erasure for language-invariant dense retrieval. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13261–13273, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.736. URL https://aclanthology.org/2024.emnlp-main.736/.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. PACRR: A position-aware neural IR model for relevance matching. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1049–1058, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1110. URL https://aclanthology.org/D17-1110/.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. Co-PACRR: A context-aware neural ir model for ad-hoc retrieval. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, WSDM '18, pp. 279–287, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3159652.3159689.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkxgnnNFvH.

Shayekh Bin Islam, Md Asib Rahman, K. S. M. Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. Open-RAG: Enhanced retrieval augmented reasoning with open-source large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14231–14244, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.831. URL https://aclanthology.org/2024.findings-emnlp.831/.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 935–944, 2016.

Himanshu Jain, Venkatesh Balasubramanian, Bhanu Chunduri, and Manik Varma. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, pp. 528–536, 2019. doi: 10.1145/3289600.3290979.

Soroush Javdan, Pragash Krishnamoorthy, and Olga Baysal. CREST: Improving interpretability and effectiveness of troubleshooting at ericsson through criterion-specific trouble report retrieval. *arXiv preprint arXiv:2511.17417*, 2025.

Frederick Jelinek. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of Workshop on Pattern Recognition in Practice*, 1980.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40 (Supplement_1):i119–i129, 2024. doi: 10.1093/bioinformatics/btae238.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023a.

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

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Thorsten Joachims. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 217–226, New York, NY, USA, 2006. Association for Computing Machinery. doi: 10.1145/1150402.1150429.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, pp. 781–789, 2017.

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

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

Yeong-Joon Ju, Ho-Joong Kim, and Seong-Whan Lee. MIRe: Enhancing multimodal queries representation via fusion-free modality interaction for multimodal retrieval. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5350–5363, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.279. URL https://aclanthology.org/2025.findings-acl.279/.

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

SeongKu Kang, Shivam Agarwal, Bowen Jin, Dongha Lee, Hwanjo Yu, and Jiawei Han. Improving retrieval in theme-specific applications using a corpus topical taxonomy. *Proceedings of the ACM Web Conference 2024*, 2024.

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

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014.

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

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165. PMLR, 2020.

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

Michael Kearns and Leslie Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, 41(1):67–95, 1994.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

Omar Khattab and Matei Zaharia. 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*, pp. 39–48, 2020.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_nGgzQjzaRy.

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

Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjaneyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. EmbedDistill: A geometric knowledge distillation for information retrieval. *arXiv preprint arXiv:2301.12005*, 2023.

Sungyeon Kim, Xinliang Zhu, Xiaofan Lin, Muhammet Bastan, Douglas Gray, and Suha Kwak. GENIUS: A generative framework for universal multimodal search. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19659–19669, 2025.

Yoon Kim. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181/.

Varsha Kishore, Chao Wan, Justin Lovelace, Yoav Artzi, and Kilian Q. Weinberger. Incdsi: Incrementally updatable document retrieval. In *International Conference on Machine Learning*, pp. 17122–17134. PMLR, 2023.

Fanheng Kong, Jingyuan Zhang, Yahui Liu, Hongzhi Zhang, Shi Feng, Xiaocui Yang, Daling Wang, Yu Tian, Qi Wang, Fuzheng Zhang, and Guorui Zhou. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *arXiv preprint arXiv:2505.19650*, 2025.

Weize Kong, Jeffrey M. Dudek, Cheng Li, Mingyang Zhang, and Michael Bendersky. 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*, pp. 2399–2403, 2023.

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

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

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning, PMLR*, 32(2):1188–1196, 22–24 Jun 2014. URL https://proceedings.mlr.press/v32/le14.html.

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

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *The 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lgsyLSsDRe.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*, 2019.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015. doi: 10.1162/tacl_a_00134. URL https://aclanthology.org/Q15-1016/.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

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

Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. PARADE: Passage representation aggregation fordocument reranking. *ACM Transactions on Information Systems*, 42(2), September 2023a. doi: 10.1145/3600088.

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

Dongfang Li, Baotian Hu, and Qingcai Chen. Calibration meets explanation: A simple and effective approach for model confidence estimates. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2775–2784, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.178. URL https://aclanthology.org/2022.emnlp-main.178/.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *AAAI Conference on Artificial Intelligence*, 2019.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022b.

Minghan Li, Sheng-Chieh Lin, Xueguang Ma, and Jimmy Lin. 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, pp. 1954–1959, New York, NY, USA, 2023c. Association for Computing Machinery. doi: 10.1145/3539618.3591977.

Mufei Li, Siqi Miao, and Pan Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *The 13th International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=JvkuZZ04O7.

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

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

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

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3), May 2025d. doi: 10.1145/3722552.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, et al. Towards agentic RAG with deep reasoning: A survey of RAG-reasoning systems in LLMs. *arXiv preprint arXiv:2507.09477*, 2025e.

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

Jintao Liang, Gang Su, Huifeng Lin, You Wu, Rui Zhao, and Ziyue Li. Reasoning RAG via system 1 or system 2: A survey on reasoning agentic retrieval-augmented generation for industry challenges. *arXiv preprint arXiv:2506.10408*, 2025.

Jimmy Lin. A proposed conceptual framework for a representational approach to information retrieval. *ACM SIGIR Forum*, 55:1–29, 2021.

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

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

Juexin Lin, Sachin Yadav, Feng Liu, Nicholas Rossi, Praveen Reddy Suram, Satya Chembolu, Prijith Chandran, Hrushikesh Mohapatra, Tony Lee, Alessandro Magnani, and Ciya Liao. Enhancing relevance of embedding-based retrieval at Walmart. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024.

Minhua Lin, Zongyu Wu, Zhichao Xu, Hui Liu, Xianfeng Tang, Qi He, Charu Aggarwal, Hui Liu, Xiang Zhang, and Suhang Wang. A comprehensive survey on reinforcement learning-based agentic search: Foundations, roles, optimizations, evaluations, and applications. *arXiv preprint arXiv:2510.16724*, 2025.

Robert Litschko, Ivan Vulić, and Goran Glavaš. Parameter-efficient neural reranking for cross-lingual and multilingual retrieval. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1071–1082, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.90/.

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

Hao Liu, Zhengren Wang, Xi Chen, Zhiyu Li, Feiyu Xiong, Qinhan Yu, and Wentao Zhang. HopRAG: Multi-hop reasoning for logic-aware retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1897–1913, Vienna, Austria, July 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.97. URL https://aclanthology.org/2025.findings-acl.97/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 115–124, 2017.

Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. Hm-RAG: Hierarchical multi-agent multimodal retrieval augmented generation. *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025b.

Qi Liu, Bo Wang, Nan Wang, and Jiaxin Mao. Leveraging passage embeddings for efficient listwise reranking with large language models. In *Proceedings of the ACM on Web Conference 2025*, 2024b. doi: 10.1145/3696410.3714554.

Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. HiT: Hierarchical transformer with momentum contrast for video-text retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11895–11905, 2021.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, March 2009. doi: 10.1561/1500000016.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, and Maarten de Rijke. Robust information retrieval. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, pp. 1008–1011, 2025c.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2395–2405, 2018.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. Zero-shot entity linking by reading entity descriptions. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3449–3460, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1335. URL https://aclanthology.org/P19-1335/.

David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision*, volume 2, pp. 1150–1157, 1999.

Zhengdong Lu and Hang Li. A deep architecture for matching short texts. *Advances in neural information processing systems*, 26, 2013.

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

Guangyuan Ma, Xing Wu, Zijia Lin, and Songlin Hu. Drop your decoder: Pre-training with bag-of-word prediction for dense passage retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1818–1827, 2024a.

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

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

Sean MacAvaney, Luca Soldaini, and Nazli Goharian. Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning. *Advances in Information Retrieval*, 12036:246 – 254, 2019a.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1101–1104, 2019b.

Sean MacAvaney, Franco Maria Nardini, R. Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. Efficient document re-ranking for transformers by precomputing term representations. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. ABNIRML: Analyzing the behavior of neural IR models. *Transactions of the Association for Computational Linguistics*, 10:224–239, 2022. doi: 10.1162/tacl_a_00457. URL https://aclanthology.org/2022.tacl-1.13/.

David J. C. MacKay and Linda C. Bauman Peto. A hierarchical Dirichlet language model. *Natural language engineering*, 1(3):289–308, 1995.

Yu A. Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

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

Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4): 41–46, April 2006. doi: 10.1145/1121949.1121979.

Sanket Vaibhav Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Q Tran, Jinfeng Rao, Marc Najork, Emma Strubell, and Donald Metzler. DSI++: Updating transformer memory with new documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8198–8213, 2023.

Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. A survey of Multimodal Retrieval-Augmented Generation. arXiv preprint arXiv:2504.08748, 2025.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

David R. H. Miller, Tim Leek, and Richard M. Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 214–221, 1999.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Ann Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *arXiv preprint arXiv:2106.07998*, 2021.

Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana. A dual embedding space model for document ranking. *arXiv preprint arXiv:1602.01137*, 2016.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 1291–1299, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. doi: 10.1145/3038912.3052579.

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

Bhaskar Mitra, Sebastian Hofstätter, Hamed Zamani, and Nick Craswell. 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*, pp. 1697–1702, 2021.

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

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

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

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

Iurii Mokrii, Leonid Boytsov, and Pavel Braslavski. A systematic evaluation of transfer learning and pseudo-labeling with BERT-based ranking models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 2081–2085, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3404835.3463093.

Niklas Muennighoff. SGPT: GPT sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*, 2022.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *The 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BC4lIvfSzv.

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

Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. Improving document ranking with dual word embeddings. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pp. 83–84, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. doi: 10.1145/2872518.2889361.

Pandu Nayak. Understanding searches better than ever before. https://blog.google/products/search/search-language-understanding-bert/, 2019.

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

Thong Nguyen, Sean MacAvaney, and Andrew Yates. A unified framework for learned sparse retrieval. In *European Conference on Information Retrieval*, pp. 101–116. Springer, 2023.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1864–1874, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.146. URL https://aclanthology.org/2022.findings-acl.146/.

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. Large dual encoders are generalizable retrievers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9844–9855, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.669. URL https://aclanthology.org/2022.emnlp-main.669/.

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

Priyank Nigam, Yiwei Song, Vijai Mohan, Vihan Lakshman, Weitian Ding, Ankit Shingavi, Choon Hui Teo, Hao Gu, and Bing Yin. Semantic product search. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

Rodrigo Nogueira, Jimmy Lin, and A. I. Epistemic. From doc2query to docTTTTTquery. *Online preprint*, 6(2), 2019a.

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

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

Cicero Nogueira dos Santos, Xiaofei Ma, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. Beyond [CLS] through ranking by generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1722–1727, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.134. URL https://aclanthology.org/2020.emnlp-main.134/.

Zach Nussbaum and Brandon Duderstadt. Training sparse mixture of experts text embedding models. arXiv preprint arXiv:2502.07972, 2025.

Zach Nussbaum, John Xavier Morris, Andriy Mulyar, and Brandon Duderstadt. Nomic Embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=IPmzyQSiQE.

Douglas Oard, Carol Peters, Miguel Ruiz, Robert Frederking, Judith Klavans, and Paraic Sheridan. Multilingual Information Discovery and AccesS (MIDAS). *D-Lib Magazine*, 5(10), 1999. doi: 10.1045/october99-oard.

Douglas W. Oard and Bonnie J. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19 CS-TR-3615, University of Maryland Institute for Advanced Computer Studies, 1996.

Douglas W. Oard and Bonnie J. Dorr. Evaluating cross-language text filtering effectiveness. In *Cross-Language Information Retrieval*, pp. 151–161. Springer, 1998.

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

Bruno Oliveira and Carla Teixeira Lopes. From 10 blue links pages to feature-full search engine results pages - analysis of the temporal evolution of SERP features. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pp. 338–345. ACM, March 2023. doi: 10.1145/3576840.3578307.

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. Neural information retrieval: At the end of the early years. *Information Retrieval Journal*, 21:111–182, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Introducing ChatGPT. https://openai.com/index/chatgpt/, 2022.

OpenAI. Introducing ChatGPT search. https://openai.com/index/introducing-chatgpt-search/, 2024.

OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, et al. GPT-4o system card. arXiv preprint arXiv:2410.21276, 2024.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. *IEEE Transactions on Audio, Speech and Language Processing*, 24(4):694–707, 2016. doi: 10.1109/TASLP.2016.2520371.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text matching as image recognition. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, AAAI'16, pp. 2793–2799. AAAI Press, 2016.

Liang Pang, Jun Xu, Qingyao Ai, Yanyan Lan, Xueqi Cheng, and Ji-Rong Wen. 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*, pp. 499–508, 2020.

Eunhwan Park, Sung-Min Lee, Dearyong Seo, Seonhoon Kim, Inho Kang, and Seung-Hoon Na. RINK: Reader-inherited evidence reranker for table-and-text open domain question answering authors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13446–13456, 2023.

Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Advances in Neural Information Processing Systems*, 2024a.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large Language Model connected with massive APIs. In *The 38th Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=tBRNC6YemY.

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. RWKV: Reinventing RNNs for the transformer era. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.936. URL https://aclanthology.org/2023.findings-emnlp.936/.

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

Wenjun Peng, Guiyang Li, Yue Jiang, Zilong Wang, Dan Ou, Xiaoyi Zeng, Derong Xu, Tong Xu, and Enhong Chen. Large language model based long-tail query rewriting in Taobao search. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 20–28, New York, NY, USA, 2024b. Association for Computing Machinery. doi: 10.1145/3589335.3648298.

Gustavo Penha and Claudia Hauff. On the calibration and uncertainty of neural learning to rank models for conversational search. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 160–170, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.12. URL https://aclanthology.org/2021.eacl-main.12/.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, 2021.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL https://aclanthology.org/2020.emnlp-main.617/.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 487–503, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.39. URL https://aclanthology.org/2021.eacl-main.39/.

Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–281, 1998.

Jeff Pool, Abhishek Sawarkar, and Jay Rodge. Accelerating inference with sparsity using the NVIDIA Ampere architecture and NVIDIA TensorRT. NVIDIA Technical Blog, July 2021. URL https://developer.nvidia.com/blog/accelerating-inference-with-sparsity-using-ampere-and-tensorrt/.

Jacob Portes, Alexander Trott, Sam Havens, Daniel King, Abhinav Venigalla, Moin Nadeem, Nikhil Sardana, Daya Khudia, and Jonathan Frankle. MosaicBERT: A bidirectional encoder optimized for fast pretraining. *Advances in Neural Information Processing Systems*, 36:3106–3130, 2023.

Jes'us Andr'es Portillo-Quintero, José carlos Ortíz-Bayliss, and Hugo Terashima-Mar'in. A straightforward framework for video retrieval using CLIP. In *Mexican Conference on Pattern Recognition*, 2021.

Yashoteja Prabhu and Manik Varma. FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 263–272, 2014.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference*, pp. 993–1002, 2018.

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

Ronak Pradeep, Kai Hui, Jai Gupta, Adam Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Tran. How does generative retrieval scale to millions of passages? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1305–1321, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.83. URL https://aclanthology.org/2023.emnlp-main.83/.

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

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

Jiarui Qin, Jiachen Zhu, Bo Chen, Zhirong Liu, Weiwen Liu, Ruiming Tang, Rui Zhang, Yong Yu, and Weinan Zhang. RankFlow: Joint optimization of multi-stage cascade ranking systems as flows. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

Tao Qin and Tie-Yan Liu. Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013.

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

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The 12th International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=dHng2O0Jjr.

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

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. Large language models are effective text rankers with pairwise ranking prompting. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1504–1518, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.97. URL https://aclanthology.org/2024.findings-naacl.97/.

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

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

Qwen Team. QwQ-32B: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/, March 2025.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language-models.pdf, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

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

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trung Hieu Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. *arXiv preprint arXiv:2305.05065*, 2023.

Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. What are you token about? Dense retrieval as distributions over the vocabulary. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2481–2498, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.140. URL https://aclanthology.org/2023.acl-long.140/.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023b.

Viresh Ranjan, Nikhil Rasiwasia, and C. V. Jawahar. Multi-label cross-modal retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4094–4102, 2015.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 251–260, 2010.

Shauli Ravfogel, Valentina Pyatkin, Amir David Nissan Cohen, Avshalom Manevich, and Yoav Goldberg. Description-based text similarity. In *1st Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=W8Rv1jVycX.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on*

Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2825–2835, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.224. URL https://aclanthology.org/2021.emnlp-main.224/.

Zhaochun Ren, Xiangnan He, Dawei Yin, and M. de Rijke. Information discovery in e-commerce. *Foundations and Trends in Information Retrieval*, 18:417–690, 2024.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, April 2009. doi: 10.1561/1500000019.

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

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference (TREC-3), NIST Special Publication*, pp. 109–126. National Institute of Standards & Technology, 1995.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54/.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5919–5930, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.477. URL https://aclanthology.org/2020.emnlp-main.477/.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. Improving passage retrieval with zero-shot question generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3781–3797, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.249. URL https://aclanthology.org/2022.emnlp-main.249/.

Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.

Gerard Salton. Interactive information retrieval. Technical report, Cornell University, 1969.

Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194, 1970.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

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

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.272. URL https://aclanthology.org/2022.naacl-main.272/.

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

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

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

Aliaksei Severyn and Alessandro Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 373–382, 2015.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. Scaling retrieval-based language models with a trillion-token datastore. In *The 38th Annual Conference on Neural Information Processing Systems*, 2024.

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, and Luke Zettlemoyer. ReasonIR: Training retrievers for reasoning tasks. In *2nd Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=kkBCNLMbGj.

Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2160–2167. IEEE, 2012.

Xinjie Shen, Zhichao Geng, and Yang Yang. Exploring l0 sparsification for inference-free sparse retrievers. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 2572–2576, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3726302.3730192.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pp. 373–374, New York, NY, USA, 2014a. Association for Computing Machinery. doi: 10.1145/2567948.2577348.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pp. 101–110, New York, NY, USA, 2014b. Association for Computing Machinery. doi: 10.1145/2661829.2661935.

Peng Shi, He Bai, and Jimmy Lin. Cross-lingual training of neural models for document ranking. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2768–2773, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.249. URL https://aclanthology.org/2020.findings-emnlp.249/.

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

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv preprint arXiv:2501.09136*, 2025a.

Amanpreet Singh, Joseph Chee Chang, Dany Haddad, Aakanksha Naik, Jena D. Hwang, Rodney Kinney, Daniel S Weld, Doug Downey, and Sergey Feldman. Ai2 scholar QA: Organized literature synthesis with attribution. In Pushkar Mishra, Smaranda Muresan, and Tao Yu (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 513–523, Vienna, Austria, July 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-demo.49. URL https://aclanthology.org/2025.acl-demo.49/.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2219–2228, 2018.

Richard Socher, Andrej Karpathy, Quoc Le, Christopher D. Manning, and Andrew Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the association for computational linguistics*, 2:207–218, 2014.

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

Fei Song and W. Bruce Croft. A general language model for information retrieval. In *Proceedings of the 8th International Conference on Information and Knowledge Management*, pp. 316–321, 1999.

Meina Song, Qing Liu, and E. Haihong. Deep hierarchical attention networks for text matching in information retrieval. *2018 International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pp. 476–481, 2018.

Daniel Sonntag and Hans-Jürgen Profitlich. An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation. *Artificial intelligence in Medicine*, 93: 13–28, 2018.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 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, pp. 12–22, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3673791.3698415.

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

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

Yongye Su, Zeya Zhang, Jane Kou, Cheng Ju, Shubhojeet Sarkar, Yamin Wang, Ji Liu, and Shengbo Guo. Modernizing facebook scoped search: Keyword and embedding hybrid retrieval with LLM evaluation. arXiv preprint arXiv:2509.13603, 2025.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 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*, pp. 14918–14937, 2023.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. ERNIE: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In Chengqing Zong and Michael Strube (eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1150. URL https://aclanthology.org/P15-1150/.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, and Maarten de Rijke. Recent advances in generative information retrieval. *Companion Proceedings of the ACM Web Conference 2024*, 2023.

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. Transformer memory as a differentiable search index. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Vu-B0clPfq.

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

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

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

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

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Howard Turtle and James Flood. Query evaluation: strategies and optimizations. *Information Processing & Management*, 31(6):831–850, 1995.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. UDapter: Language adaptation for truly Universal Dependency parsing. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2302–2315, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.180. URL https://aclanthology.org/2020.emnlp-main.180/.

C. Van Rijsbergen. Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle Upon Tyne Seminar on Data Base Systems*, volume 79, pp. 1–14, 1979.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, et al. EmbeddingGemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. A primer on kernel methods. In *Kernel Methods in Computational Biology*, pp. 35–70. MIT press, 2004.

David Wan, Han Wang, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. CLaMR: Contextualized late-interaction for multimodal content retrieval. *arXiv preprint arXiv:2506.06144*, 2025.

Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. A deep architecture for semantic matching with multiple positional sentence representations. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, AAAI'16, pp. 2835–2841. AAAI Press, 2016.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 154–162, 2017.

Haixun Wang and Taesik Na. Rethinking e-commerce search. *ACM SIGIR Forum*, 57:1–19, 2023.

Jiongxiao Wang, Qiaojing Yan, Yawei Wang, Yijun Tian, Soumya Smruti Mishra, Zhichao Xu, Megha Gandhi, Panpan Xu, and Lin Lee Cheong. Reinforcement learning for self-improving agent with skill library, 2025a. URL https://arxiv.org/abs/2512.17102.

Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

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

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

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

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions. *arXiv preprint arXiv:2308.14263*, 2024b.

Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pp. 610–618, 2018.

Xuewei Wang, Qiang Jin, Shengyu Huang, Min Zhang, Xi Liu, Zhengli Zhao, Yukun Chen, Zhengyu Zhang, Jiyan Yang, Ellie Wen, Sagar Chordia, Wenlin Chen, and Qin Huang. Towards the better ranking consistency: A multi-task learning framework for early stage ads ranking. *arXiv preprint arXiv:2307.11096*, 2023b.

Yumeng Wang, Lijun Lyu, and Avishek Anand. BERT rankers are brittle: A study using adversarial document perturbations. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '22, pp. 115–120, New York, NY, USA, 2022b. Association for Computing Machinery. doi: 10.1145/3539813.3545122.

Zora Zhiruo Wang, Akari Asai, Xinyan Velocity Yu, Frank F. Xu, Yiqing Xie, Graham Neubig, and Daniel Fried. CodeRAG-bench: Can retrieval augment code generation? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3199–3214, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.176. URL https://aclanthology.org/2025.findings-naacl.176/.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.127. URL https://aclanthology.org/2025.acl-long.127/.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. In *Computer Vision – ECCV 2024*, volume 15145 of *Lecture Notes in Computer Science*, pp. 387–404. Springer, Cham, 2025. doi: 10.1007/978-3-031-73021-4_23.

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

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

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. FollowIR: Evaluating and teaching information retrieval models to follow instructions. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 11926–11942, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.597. URL https://aclanthology.org/2025.naacl-long.597/.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. Rank1: Test-time compute for reranking in information retrieval. In *2nd Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=Pg0PAvbhGv.

Ryen W. White and Resa A. Roth. *Exploratory search: Beyond the query-response paradigm*. Morgan & Claypool Publishers, 2009.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2016.

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

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

Marco Wrzalik and Dirk Krechel. CoRT: Complementary rankings from transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4194–4204, 2021.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. WebDancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*, 2025a.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. WebWalker: Benchmarking LLMs in web traversal. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10290–10305, Vienna, Austria, July 2025b. Association for Computational Linguistics. URL https://aclanthology.org/2025.acl-long.508/.

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

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

Xing Wu, Guangyuan Ma, Meng Lin, Zijia Lin, Zhongyuan Wang, and Songlin Hu. Contextual masked auto-encoder for dense passage retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4738–4746, 2023b.

Yunjia Xi, Jianghao Lin, Yongzhao Xiao, Zheli Zhou, Rong Shan, Te Gao, Jiachen Zhu, Weiwen Liu, Yong Yu, and Weinan Zhang. A survey of LLM-based deep search agents: Paradigm, optimization, evaluation, and challenges. *arXiv preprint arXiv:2508.05668*, 2025.

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

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

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

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 538–548, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.35. URL https://aclanthology.org/2022.emnlp-main.35/.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024a.

Quanting Xie, So Yeon Min, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Russ Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. Embodied-RAG: General non-parametric embodied memory for retrieval and generation. In *Language Gamification - NeurIPS 2024 Workshop*, 2024b. URL https://openreview.net/forum?id=U8p8zpK3jL.

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 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, pp. 55–64, New York, NY, USA, 2017. Association for Computing Machinery. doi: 10.1145/3077136.3080809.

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

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

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel De Souza Pereira Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-Embed-Nemotron: A unified multimodal retrieval model for text, image, audio, and video. *arXiv preprint arXiv:2510.03458*, 2025a.

Shicheng Xu, Danyang Hou, Liang Pang, Jingcheng Deng, Jun Xu, Huawei Shen, and Xueqi Cheng. 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, pp. 208–217, New York, NY, USA, 2024a. Association for Computing Machinery. doi: 10.1145/3626772.3657750.

Zhichao Xu. RankMamba: Benchmarking Mamba's document ranking performance in the era of transformers. *arXiv preprint arXiv:2403.18276*, 2024.

Zhichao Xu and Daniel Cohen. A lightweight constrained generation alternative for query-focused summarization. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 1745–1749, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591936. URL https://doi.org/10.1145/3539618.3591936.

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

Zhichao Xu, Ashim Gupta, Tao Li, Oliver Bentham, and Vivek Srikumar. Beyond perplexity: Multi-dimensional safety evaluation of LLM compression. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15359–15396, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.901. URL https://aclanthology.org/2024.findings-emnlp.901/.

Zhichao Xu, Hemank Lamba, Qingyao Ai, Joel Tetreault, and Alex Jaimes. CFE2: Counterfactual editing for search result explanation. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 145–155, 2024c.

Zhichao Xu, Aosong Feng, Yijun Tian, Haibo Ding, and Lin Lee Cheong. CSPLADE: Learned sparse retrieval with causal language models. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 99–114, Mumbai, India, December 2025b. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. URL https://aclanthology.org/2025.ijcnlp-long.7/.

Zhichao Xu, Zhiqi Huang, Shengyao Zhuang, and Vivek Srikumar. Distillation versus contrastive learning: How to train your rerankers. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 564–578, Mumbai, India, December 2025c. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-ijcnlp.33/.

Zhichao Xu, Minheng Wang, Yawei Wang, Wenqian Ye, Yuntao Du, Yunpu Ma, and Yijun Tian. Recon: Reasoning with condensation for efficient retrieval-augmented generation, 2025d. URL https://arxiv.org/abs/2510.10448.

Zhichao Xu, Jinghua Yan, Ashim Gupta, and Vivek Srikumar. State space models are strong text rerankers. In Vaibhav Adlakha, Alexandra Chronopoulou, Xiang Lorraine Li, Bodhisattwa Prasad Majumder, Freda Shi, and Giorgos Vernikos (eds.), *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pp. 152–169, Albuquerque, NM, May 2025e. Association for Computational Linguistics. URL https://aclanthology.org/2025.repl4nlp-1.12/.

Zhichao Xu, Shengyao Zhuang, Xueguang Ma, Bingsen Chen, Yijun Tian, Fengran Mo, Jie Cao, and Vivek Srikumar. Rethinking on-policy optimization for query augmentation. *arXiv preprint arXiv:2510.17139*, 2025f.

Zhichao Xu, Shengyao Zhuang, Crystina Zhang, Xueguang Ma, Yijun Tian, Maitrey Mehta, Jimmy Lin, and Vivek Srikumar. LACONIC: Dense-level effectiveness for scalable sparse retrieval via a two-phase training curriculum. *arXiv preprint arXiv:2601.01684*, 2026.

Le Yan, Zhen Qin, Honglei Zhuang, Rolf Jagerman, Xuanhui Wang, Michael Bendersky, and Harrie Oosterhuis. Consolidating ranking and relevance predictions of large language models through post-processing. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 410–423, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.25. URL https://aclanthology.org/2024.emnlp-main.25/.

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

Eugene Yang, Dawn J. Lawrie, and James Mayfield. Distillation for multilingual information retrieval. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

Eugene Yang, Andrew Yates, Kathryn Ricci, Orion Weller, Vivek Chari, Benjamin Van Durme, and Dawn Lawrie. Rank-K: Test-time reasoning for listwise reranking. *arXiv preprint arXiv:2505.14432*, 2025b.

Liu Yang, Qingyao Ai, Jiafeng Guo, and W. Bruce Croft. aNMM: Ranking short answer texts with attention-based neural matching model. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pp. 287–296, New York, NY, USA, 2016a. Association for Computing Machinery. doi: 10.1145/2983323.2983818.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta networks: Improving Mamba2 with Delta rule. In *The 13th International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=r8H7xhYPwz.

Tao Yang, Zhichao Xu, and Qingyao Ai. 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*, pp. 234–244, 2023a.

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

Tao Yang, Zhichao Xu, Zhenduo Wang, Anh Tran, and Qingyao Ai. Marginal-certainty-aware fair ranking algorithm. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, pp. 24–32, 2023c.

Xiao Yang, Peifeng Yin, Abe Engle, Jinfeng Zhuang, and Ling Leng. MTMD: A multi-task multi-domain framework for unified ad lightweight ranking at pinterest. *arXiv preprint arXiv:2510.09857*, 2025d.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, San Diego, California, June 2016b. Association for Computational Linguistics. doi: 10.18653/v1/N16-1174. URL https://aclanthology.org/N16-1174/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

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

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

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for computational linguistics*, 4:259–272, 2016.

Soyoung Yoon, Eunbi Choi, Jiyeon Kim, Hyeongu Yun, Yireun Kim, and Seung-won Hwang. ListT5: Listwise reranking with fusion-in-decoder improves zero-shot retrieval. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2287–2308, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.125. URL https://aclanthology.org/2024.acl-long.125/.

Soyoung Yoon, Gyuwan Kim, Gyu-Hwung Cho, and Seung won Hwang. AcuRank: Uncertainty-aware adaptive computation for listwise reranking. *arXiv preprint arXiv:2505.18512*, 2025.

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

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in neural information processing systems*, 32, 2019.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent, 2025. URL https://arxiv.org/abs/2507.02259.

Hsiang-Fu Yu, Kai Zhong, Jiong Zhang, Wei-Cheng Chang, and Inderjit S. Dhillon. PECOS: Prediction for enormous and correlated output spaces. *Journal of Machine Learning Research*, 23(98):1–32, 2022a.

Puxuan Yu and James Allan. 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*, pp. 1637–1640, 2020.

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

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

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

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. Chain-of-Note: Enhancing robustness in retrieval-augmented language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14672–14685, Miami, Florida, USA, November 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.813. URL https://aclanthology.org/2024.emnlp-main.813/.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. COCO-DR: Combating the distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1462–1479, Abu Dhabi, United Arab Emirates, December 2022c. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.95. URL https://aclanthology.org/2022.emnlp-main.95/.

Hamed Zamani and W. Bruce Croft. Joint modeling and optimization of search and recommendation. In *Biennial Conference on Design of Experimental Search & Information Retrieval Systems*, 2018.

Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 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*, pp. 497–506, 2018.

Rabih Zbib, Lingjun Zhao, Damianos Karakos, William Hartmann, Jay DeYoung, Zhongqiang Huang, Zhuolin Jiang, Noah Rivkin, Le Zhang, Richard Schwartz, et al. Neural-network lexical translation for cross-lingual ir from text and speech. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 645–654, 2019.

Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. Scalable and effective generative information retrieval. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 1441–1452, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3589334.3645477.

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

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

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, Yin-Hua Lu, and Yu Shi. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. RepBERT: Contextualized text embeddings for 1st-stage retrieval. *arXiv preprint arXiv:2006.15498*, 2020.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1503–1512, 2021.

Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. RepoCoder: Repository-level code completion through iterative retrieval and generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2471–2484, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.151. URL https://aclanthology.org/2023.emnlp-main.151/.

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

Huaiwen Zhang, Yang Yang, Fan Qi, Shengsheng Qian, and Changsheng Xu. Debiased video-text retrieval via soft positive sample calibration. *IEEE Transactions on Circuits and Systems for Video Technology*, 33:5257–5270, 2023b.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. Knowing more about questions can help: Improving calibration in question answering. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1958–1970, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.172. URL https://aclanthology.org/2021.findings-acl.172/.

Weinan Zhang, Junwei Liao, Ning Li, Kounianhua Du, and Jianghao Lin. Agentic information retrieval. *arXiv preprint arXiv:2410.09713*, 2025a.

Xinyu Zhang, Andrew Yates, and Jimmy Lin. Comparing score aggregation approaches for document retrieval with pretrained transformers. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (eds.), *Advances in Information Retrieval*, pp. 150–163, Cham, 2021b. Springer International Publishing.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. Toward best practices for training multilingual dense retrieval models. *ACM Transactions on Information Systems*, 42:1–33, 2022.

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

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

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 1435–1448, 2023.

Kai Zheng, Haijun Zhao, Rui Huang, Beichuan Zhang, Na Mou, Yanan Niu, Yang Song, Hongning Wang, and Kun Gai. Full stage learning to rank: A unified framework for multi-stage systems. *Proceedings of the ACM Web Conference 2024*, 2024a.

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. SGLang: Efficient execution of structured language model programs. *Advances in Neural Information Processing Systems*, 37:62557–62583, 2024b.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13764–13775, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.849. URL https://aclanthology.org/2023.emnlp-main.849/.

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

Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K. Reddy. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*, pp. 2472–2482, 2019.

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

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 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, pp. 2308–2313, New York, NY, USA, 2023a. Association for Computing Machinery. doi: 10.1145/3539618.3592047.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 358–370, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.31. URL https://aclanthology.org/2024.naacl-short.31/.

Shengyao Zhuang, Hang Li, and Guido Zuccon. 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*, pp. 463–470. Springer, 2021.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. Open-source large language models are strong zero-shot query likelihood models for document ranking. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8807–8817, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.590. URL https://aclanthology.org/2023.findings-emnlp.590/.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 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, pp. 38–47, New York, NY, USA, 2024b. Association for Computing Machinery. doi: 10.1145/3626772.3657813.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. Rank-R1: Enhancing reasoning in LLM-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*, 2025.

Justin Zobel, Alistair Moffat, and Kotagiri Ramamohanarao. Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems (TODS)*, 23(4):453–490, 1998.