Advances in Artificial Intelligence and Machine Learning; Review (Ahead of Print) Received 07-06-2026; Accepted 05-09-2026; Published 12-09-2026

Machine Learning for Ranking in Search Systems: A Scoping Review

Alexander Gamarnik gamarnikaa@student.bmstu.ru
Department of Robotics and Complex Automation
Bauman Moscow State Technical University
BMSTU
5/1, 2nd Baumanskaya St., Moscow, 105005, Russia

Corresponding Author: Alexander Gamarnik

Copyright © 2026 Alexander Gamarnik. 2026 Alexander Gamarnik. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The exponential growth of digital information has made machine learning (ML)-based ranking a crucial mechanism for modern search systems, enabling them to manage complex queries and deliver highly relevant results across various domains. This scoping review aims to systematically characterize the current state of machine learning for ranking in search systems. Following PRISMA-ScR guidelines, a comprehensive literature search was conducted across the IEEE Xplore and Lens.org databases for peer-reviewed studies published between 2023 and 2026, selecting 78 eligible papers for data extraction and thematic synthesis. Key findings reveal a clear paradigm shift, with transformer-based architectures and Large Language Models (LLMs) dominating the recent literature, particularly at the re-ranking stage. In contrast, neural ranking and classical learning-to-rank (LTR) approaches remain less frequent. Evaluation mostly relies on standardized public benchmarks like MS MARCO and BEIR, utilizing offline ranking metrics (e.g., nDCG, MRR), though efficiency and generation quality metrics are increasingly reported for LLMs. General and web search remain the primary application domains, along with emerging generative search pipelines. Critical challenges persist regarding high inference costs, latency, domain generalization, data bias, and the ”black-box” nature of these models. This review provides significant value for both science and practice by mapping the rapid structural evolution of the field, establishing a clear taxonomy of current methodologies, and highlighting critical research gaps. For practitioners and researchers alike, it offers a foundational roadmap for addressing scalability and interpretability constraints, guiding the future development of efficient, transparent ranking models for production-ready search systems.

Keywords: Machine learning, Learning to rank, Search system, Information retrieval, Ranking problem.

1. INTRODUCTION

The exponential growth of digital information has precipitated a paradigm shift in search technology, introducing AI-driven approaches to manage complex queries and large data corpora [1, 2]. At the core of this challenge is the search system, which is defined as a system that, in response to an

Citation: Alexander Gamarnik. Machine Learning for Ranking in Search Systems: A Scoping Review.
Advances in Artificial Intelligence and Machine Learning. 2026. (Ahead of Print) https://dx.doi.org/10.54364/AAIML.2026.65344

explicit query, retrieves and ranks items from a collection and returns them ordered by relevance. Search systems manifest across a variety of domains, including web search, e-commerce, enterprise search, academic literature retrieval, and, more recently, conversational and Retrieval-Augmented Generation (RAG) pipelines. These systems are distinct from recommender systems. While search is explicitly query-driven, recommender systems rely primarily on user profiles and contextual behaviors without an explicit query, so such systems are excluded from the scope of this review.

The central mechanism of any search system is its ranking function. Historically, information retrieval relied on classical, heuristic ranking functions, such as BM25, TF-IDF, and Vector Space Models, which are based on exact lexical matching [3]. However, to address the semantic gap and to capture complex relationships between queries and documents [4, 5], the field has transitioned toward ML-based ranking, which is defined as the task of ordering a set of candidate items by their estimated relevance to a query, producing a ranked result list where the ordering function is learned from data rather than fixed by heuristics.

This transition has occurred in several waves. It began with traditional LTR algorithms using gradient-boosted decision trees (GBDT) (e.g., LambdaMART [6]), neural ranking models [7] and progressed to rankers employing transformer-based bi-encoders and cross-encoders [8]. Most recently, LLMs have revolutionized the field, demonstrating remarkable capabilities for relevance ranking [9]. Today, ML-based ranking covers the entire retrieval pipeline, including first-stage retrieval, second-stage re-ranking, and end-to-end ranking architectures.

Despite the rapid integration of various ML-based ranking models into search pipelines, there remains a significant gap in the comprehensive systematization of this literature. While recent studies include excellent focused reviews, such as the comprehensive survey on model architectures in information retrieval [10], which tracks the structural evolution from term-based models to modern LLMs, the broader landscape of ML-based ranking, trends of recent years (2023-2026), evaluation methodologies and application domains remain undefined.

To address this gap, this scoping review systematically maps the literature on ML-based ranking within search systems. Following the foundational framework [10], on model architectures for initial analysis, studies are categorised by model architectures. Metrics and datasets used to evaluate these models, as well as their application domains, are also examined. Finally, limitations, challenges, and research gaps are identified. To complete this scoping review multiple research questions (RQ) are prepared:

- **RQ1**: What ML approaches are applied for ranking in search systems?
- **RQ2**: What data is used to train and evaluate ML models for ranking in search systems?
- **RQ3**: How are ML approaches for ranking in search systems evaluated, in terms of metrics and evaluation methodology?
- **RQ4**: In what types of search systems and application domains are ML approaches for ranking studied?
- **RQ5**: What challenges, limitations, and research gaps are reported in the literature on ML methods for ranking in search systems?

2. METHODOLOGY

2.1 Protocol

The protocol for this scoping review was developed in accordance with the PRISMA extension for Scoping Reviews (PRISMA-ScR) guidelines.

2.2 Eligibility Criteria

A set of inclusion and exclusion criteria was developed to guide the selection of studies for this scoping review. The full criteria are presented in TABLE 1. To be eligible for thematic analysis, each study had to meet all of the inclusion criteria and none of the exclusion criteria.

Table 1: Definition of inclusion and exclusion criteria.

Criteria	**Decision**
Machine learning-based ranking applied in search systems is the central focus.	Inclusion
Journal article or conference paper	Inclusion
Has a DOI	Inclusion
Peer-reviewed	Inclusion
Publication year: 2023 or newer	Inclusion
Language: English or Russian	Inclusion
Published before 2023	Exclusion
Review paper	Exclusion
Non-peer-reviewed (preprints, theses, reports)	Exclusion
Off-topic (ML-based ranking is not the main theme)	Exclusion
Off-scope (ML-based ranking is not applied to a search system)	Exclusion
Full text not available (only able to use those with free access)	Exclusion
Keynotes, editorials, short papers (< 4 pages)	Exclusion
Duplicate record	Exclusion

For the purposes of this review, ranking includes first-stage retrieval, re-ranking stage, and end-to-end ranking pipelines. The review included studies using classical LTR, neural ranking, transformer-based ranking, and LLM-based ranking approaches. Operational definitions and coding rules for model families and ranking stages are provided in the ”Classification framework and coding rules” subsection.

The following are considered out of scope unless used solely as a baseline: classical heuristic ranking functions (BM25, TF-IDF, VSM cosine, fixed query-likelihood models), boolean/exact-match filtering, link-analysis heuristics (PageRank, HITS), pure classification or regression with no ordering output.

A search system is defined in the Introduction. In-scope system types include web search, e-commerce and product search, enterprise and site search, academic and literature search, code search, vertical search, conversational and RAG-based retrieval where the ranking component is the focus, and query-driven question answering and passage retrieval. The following are explicitly out of scope: recommender systems and other non-query-driven ranking. Medical and legal search,

which refer to a distinct subfield with specialized data and evaluation, are considered off-scope for this review. Cross-modal and multimodal retrieval where ranked items are non-textual is considered off-scope as it relies on different model architectures.

2.3 Information Sources

Two databases were searched: IEEE Xplore and Lens.org, which were selected on the basis of free accessibility and their coverage of the relevant literature: IEEE Xplore provides in-depth coverage of computer science and engineering venues, while Lens.org is a cross-publisher scholarly index that aggregates records from multiple publishers. The search was conducted on 14 April 2026.

ACM Digital Library, Scopus, Web of Science, SpringerLink, ScienceDirect, and ACL Anthology were not searched separately to avoid unnecessary duplication and to maintain an openly reproducible search strategy. Lens.org was selected as a cross-publisher scholarly aggregator that provides broad coverage of records associated with multiple publishers, including ACM, Springer, and Elsevier publications, although complete coverage of these sources was not assumed. Its free accessibility enables researchers to inspect, repeat, and validate the provided search strategy without requiring commercial database subscriptions.

2.4 Search

The search query was built from three concept blocks combined with the Boolean operator AND. Within each block, synonyms and term variants were combined with OR to maximize sensitivity. The search was applied to the title and abstract fields or to all metadata fields depending on the database. The complete search query is given in TABLE 2.

Table 2: Search query.

Search query
(”information retrieval” OR ”search engine*” OR ”web search” OR ”retrieval” OR ”search system*” OR ”RAG”) AND (”learning to rank” OR ”learning-to-rank” OR ”re-rank*” OR ”rerank*” OR ”rank*” OR ”first-stage*”) AND (”machine learning” OR ”deep learning” OR ”neural network*” OR ”transformer*” OR ”BERT” OR ”large language model*” OR ”LLM” OR ”cross-encoder” OR ”gradient boost*”)

The query was adapted to the specific syntax of each database: IEEE Xplore command search over document metadata; Lens.org scholarly search restricted to title, abstract, and keywords. The following limits were applied: publication year 2023 onward, document type, and language (English or Russian).

2.5 Selection of Sources of Evidence

Records were identified by running the search query in IEEE Xplore and Lens.org. In both databases, the search was conducted in title/abstract fields or available metadata, depending on the database

functionality. Database filters were applied to restrict results to studies published in 2023 or later and to document types corresponding to journal articles and conference papers from Lens.org.

Search results from the two databases were exported and merged into a single record set. Duplicate records were removed by comparing paper titles. After deduplication, records were further checked for eligibility at the record level: review papers identified in IEEE Xplore were removed, and papers that are written neither in English nor in Russian were excluded based on the title when the language was evident.

The remaining records underwent title and abstract screening against the inclusion and exclusion criteria. Studies judged potentially eligible were then sought for full-text retrieval. Records for which the full text was not freely available were excluded because the review was limited to papers accessible through free full-text sources. The retrieved full texts were then assessed against the eligibility criteria during full-text screening. The selection process and exclusion counts are reported in the PRISMA flow diagram.

2.6 Data Charting Process

Data were charted from each study by a single reviewer using a structured form developed in Google Sheets, designed around the research questions, and refined iteratively during extraction. To reduce the risk of reviewer-related bias, uncertainties or ambiguous cases identified during charting process were discussed and resolved through consensus with a colleague. However, the initial screening, data extraction, and coding were primarily conducted by a single reviewer which is a potential limitation.

2.7 Critical Appraisal of Sources of Evidence

No formal critical appraisal of the included studies was conducted due to the mapping-oriented purpose of this scoping review, which was to systematize the recent literature on machine learning-based ranking in search systems rather than to determine the comparative effectiveness of individual studies. The absence of formal critical appraisal is acknowledged as a limitation of this review.

2.8 Classification Framework and Coding Rules

A structured classification framework was used to ensure consistent data charting and synthesis across the included studies. The framework covered four dimensions: model family, ranking stage, evaluation metrics, and application domain.

Model family. The classification of model families was adapted from the architecture-oriented framework proposed in [10]. Four model-family categories were used: classical LTR, neural ranking, transformer-based ranking, and LLM-based ranking. In classical LTR core idea is to train a model that can optimally combine a wide array of features to predict the relevance of documents to a query (e.g., GBDT, LambdaMART, pointwise/pairwise/listwise formulations). Neural ranking includes non-transformer neural architectures that learn representations from data, in-

cluding recurrent neural networks, long short-term memory networks, graph neural networks, and earlier representation-based or interaction-based ranking models. Transformer-based rankers include encoder-style pre-trained architectures used for retrieval or ranking, such as BERT-based bi-encoders, cross-encoders, and late-interaction models such as ColBERT. LLM-based rankers were treated as a separate category, although they are also transformer-based in a technical sense. This category includes generative large language models used for ranking through zero-shot or few-shot prompting, instruction prompting, or instruction/fine-tuning. This distinction was made because LLM-based ranking commonly relies on generative inference and prompt-based ranking strategies, whereas encoder-style transformer rankers generally compute relevance scores through representation matching or cross-encoding.

Ranking stage. Studies were classified into three ranking stages. First-stage ranking refers to candidate generation or retrieval from a large corpus, where computational efficiency is required because the model searches across many possible documents. Re-ranking refers to the reordering of a smaller candidate set that has already been retrieved by an initial retrieval method. End-to-end ranking refers to systems in which retrieval and ranking are jointly optimized or cannot be separated into distinct first-stage and re-ranking components.

Evaluation metrics. Evaluation measures were coded into four groups: ranking metrics, non-ranking set-based metrics, efficiency metrics, and generation-quality metrics. Ranking metrics assess the quality of the ranked list and include measures such as normalized Discounted Cumulative Gain (nDCG), Mean Reciprocal Rank (MRR), and Mean Average Precision (MAP). Non-ranking metrics include measures such as Precision@k and Recall@k. Efficiency metrics cover latency, inference time, computational cost, throughput, and related resource requirements. Generation-quality metrics include measures such as BLEU and ROUGE and were recorded when ranking was evaluated as part of a generative search, question-answering, or RAG system.

Application domain. Studies were classified into five application-domain categories. General and web search includes general-purpose information retrieval, web search, document ranking, and multilingual retrieval. Task-specific search includes search systems designed for a particular domain or task, such as telecommunications, code search, education, or manufacturing, where the study does not fall into one of the more specific categories below. Generative search includes question answering, retrieval-augmented generation, and LLM-based retrieval systems in which ranking supports the generation process. Scientific search includes academic literature retrieval, scientific document search, and evidence retrieval. E-commerce search includes product search, catalogue retrieval, and other commerce-oriented search tasks.

2.9 Data Items

For each study, the following data were charted: bibliographic data (authors, year, paper type), model architecture, ranking stage, dataset(s) used, evaluation metrics, evaluation setting, reported challenges, limitations, future directions, application domain.

2.10 Synthesis of Results

Charted data were synthesized using a descriptive-analytical approach. Categorical items were summarized using predefined categories from a previous survey [10] as well as a self-proposed taxonomy with frequency counts, and cross-tabulated to examine patterns and temporal trends across 2023–2026. Open-ended items were analyzed via inductive thematic analysis. Results are presented as summary tables, charts, and a narrative synthesis organized by research question.

3. RESULTS

3.1 Search Results

The database searches were conducted on 14 April 2026. The search returned 584 records from Lens.org and 94 records from IEEE Xplore, yielding 678 records in total. After merging the results from both databases, 36 duplicate records were removed on the basis of matching titles. Prior to screening, records were filtered at the metadata level: 8 review papers were removed, followed by 1 record that was not written in English or Russian, leaving 633 records eligible for screening. The complete dataset is available on GitHub [https://github.com/alex-werben/ml-ranking-search].

During title and abstract screening, records were assessed against the eligibility criteria for scope, topic, and central focus. This stage excluded 479 records, leaving 154 records considered potentially eligible and advanced to full-text assessment. When retrieving full texts, 6 records were found not to be freely available despite being indexed as open access and were therefore excluded, leaving 148 records for detailed screening.

Full-text screening assessed each record against all eligibility criteria. A further 70 records were excluded at this stage, resulting in a final corpus of 78 studies included in the review. The full selection process and exclusion counts at each stage are summarized in the PRISMA flow diagram (FIGURE 1).

3.2 Overview of Included Studies

The 78 included studies were published between 2023 and 2026. 42 were conference papers and 36 were journal articles. 68 studies were retrieved from Lens.org and 10 from IEEE Xplore. There were 15 studies from 2023, 17 from 2024, 35 from 2025, 11 from 2026. The lower number of studies for 2026, as shown in FIGURE 2, does not indicate a decline in publication activity because the search was conducted in April 2026. The predominance of Lens.org records, as shown in FIGURE 3, is a consequence of the database design and should not be interpreted as evidence that this source represents the entire field. Full characteristics of all included studies are provided in Appendix A.

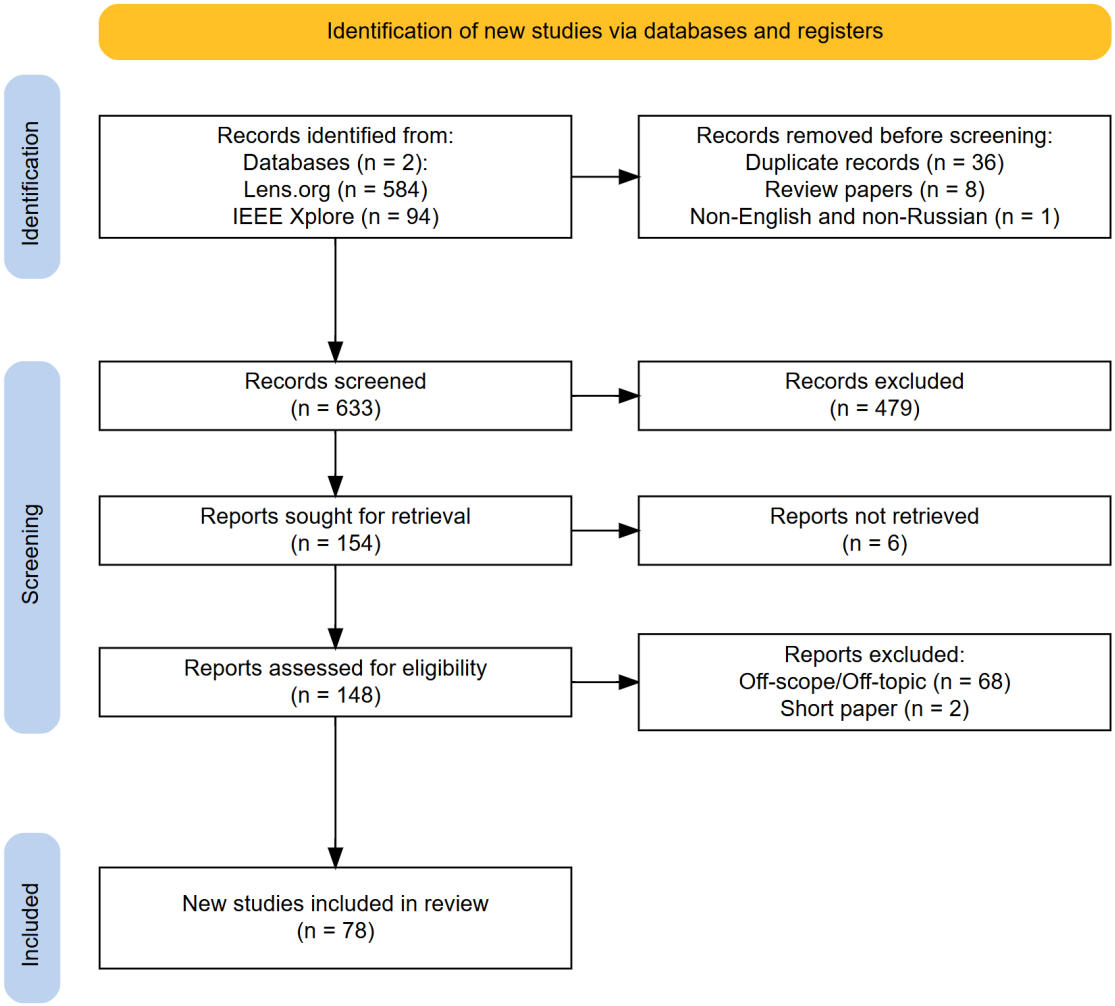

Figure 1: PRISMA flow diagram.

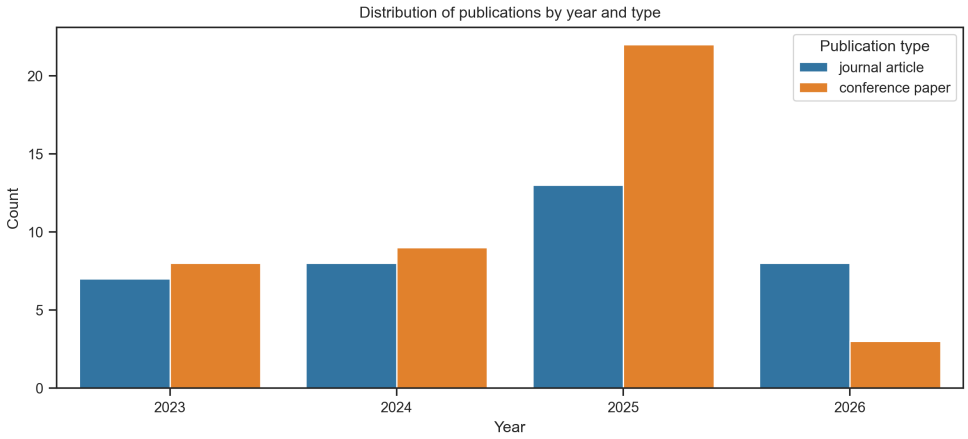

Figure 2: Distribution of publications by year and type.

3.3 Findings

3.3.1 RQ1. What ML approaches are applied for ranking in search systems?

For RQ1 each study was assigned to a specific category of model family used in each paper. Model families are defined in Methodology. Frequency counts were computed for the model architectures and ranking stages reported across the reviewed studies. A summary of these counts is provided in TABLE 3, TABLE 4, respectively. The most prevalent category is transformer-based models (n = 38), including architectures like BERT [11], [8]. Transformer-based models are utilized with two primary paradigms [10]: cross-encoders and bi-encoders [11, 12]. Additionally, late-interaction architectures like ColBERT take place in some papers [12, 13]. LLMs such as GPT-4, LLaMA, etc., are often applied as powerful, zero-shot re-ranking models [14, 15], [9], representing a big share of studies (n = 26). They are frequently applied using pointwise, pairwise, listwise, setwise prompting paradigms for re-ranking stage [16, 17]. LLMs are primarily used as rerankers. [16] states that for LLMs 'one of their main uses in ranked retrieval has been as rerankers'. Neural ranking approaches make up a relatively small segment (n = 9), capturing representation-based and interaction-based relationships. Model architectures like Graph Neural Networks (GNNs) were presented [5], [18, 19], also there were architectures based on Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) [20, 21]. Classical LTR models were the least frequently represented (n = 5), utilizing tree-based methods [6], logistic regression [22], and application of a meta-classifier for improving RAG pipeline [23]. FIGURE 4, presents the temporal distribution of publications by model architecture. The data reveal a growing trend since 2025, with an increasing number of studies adopting LLM-based and transformer-based approaches. However, as the literature search was conducted in April 2026, the counts for this year are expected to represent only about a quarter of the total publications that will ultimately appear by the end of the year. In contrast, studies focusing on neural rankers and classical LTR models consistently account for a smaller share of publications each year. FIGURE 5, further breaks down the distribution of publications by both ranking stage and model architecture. This highlights a clear pattern that the majority of transformer-based and LLM-based architectures are applied at the re-ranking stage, especially for LLM-based models.

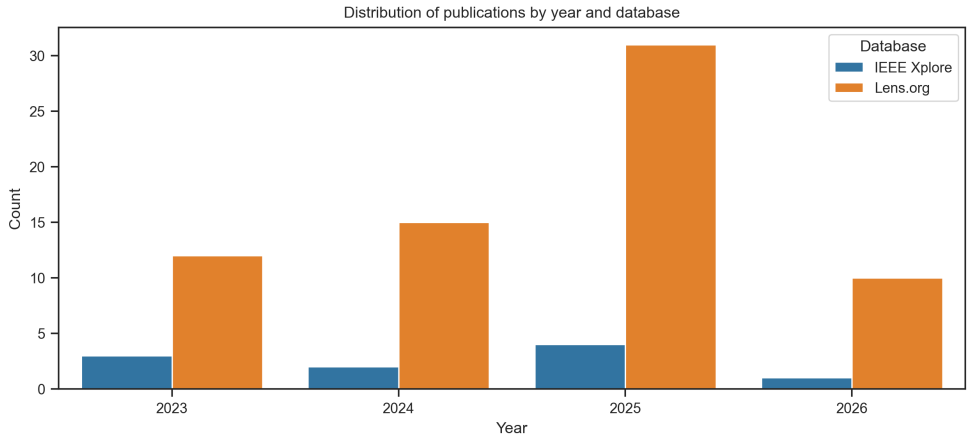

Figure 3: Distribution of publications by year and database.

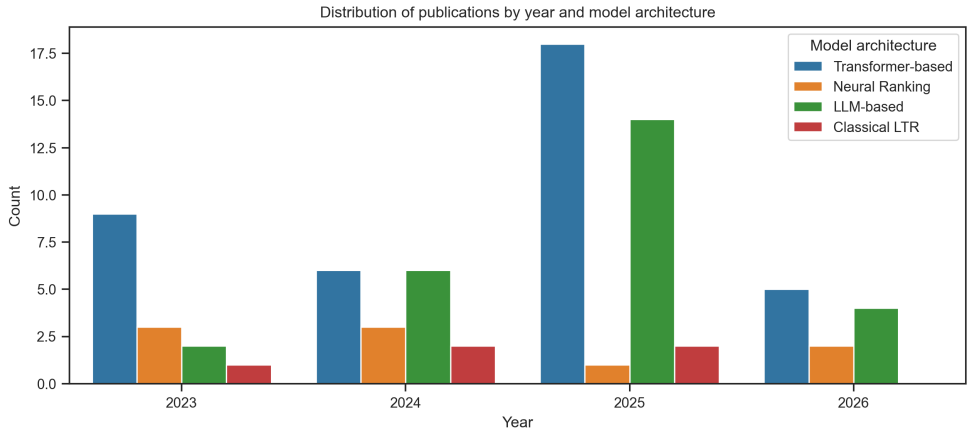

Figure 4: Distribution of publications by model architecture and publication year.

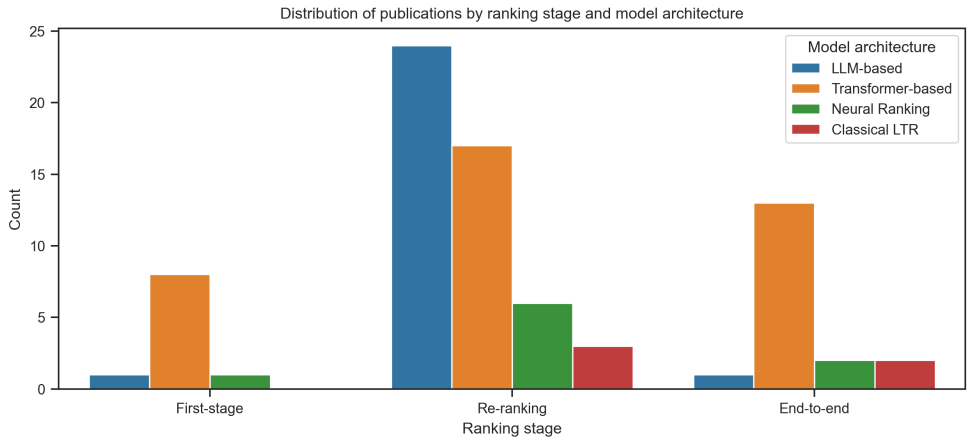

Figure 5: Distribution of publications by ranking stage and model architecture.

Table 3: Model architectures distribution in the reviewed studies.

Model family	**Number of studies**	**Percentage (%)**
Classical LTR	5	6.41
Neural ranking	9	11.54
Transformer-based	38	48.72
LLM-based	26	33.33

Table 4: Ranking stage distribution in the reviewed studies.

Ranking stage	**Number of studies**	**Percentage (%)**
First-stage	10	12.82
Re-ranking	50	64.1
End-to-end	18	23.08

3.3.2 RQ2. What data is used to train and evaluate ML models for ranking in search systems?

Analysis of the datasets used for evaluation showed that studies mostly rely on standardized public benchmarks. MS MARCO appeared most often (n = 34). [15], utilized MS MARCO 'to randomly sample 10K queries for a permutation distillation technique'. BEIR is one of the most used benchmarking suites, its full set or subsets appeared in 22 studies, TREC Deep Learning (DL) is also among the most used datasets, appearing in 18 studies. [17] used both BEIR and TREC DL 'to compare performance of various pointwise reranking models'. Datasets specifically designed for Question Answering (QA) - Natural Questions (NQ) (n = 6), Hotpot QA (n = 7), although originally part of BEIR suite, were also reported independently in some studies [24]. TREC Robust04, a part of BEIR, was used independently from it, in 6 studies. Finally, there are 9 studies using proprietary datasets for evaluation [25]. BRIGHT, a new, specialized AI benchmark built to test reasoning-intensive retrieval, was used in 4 studies. There were 23 studies, which used datasets for evaluation that appeared from 1 to 3 times across all papers. TABLE 5, displays the frequency counts of the most used datasets, highlighting a concentration of evaluations on MS MARCO, BEIR, and TREC DL that improves comparability across studies.

Table 5: Datasets used in the reviewed studies.

Dataset	**Number of studies**	**Percentage (%)**
MS_MARCO	34	43.59
BEIR	22	28.21
TREC_DL	18	23.08
Proprietary	9	11.54
Hotpot QA	7	8.97
NQ	6	7.69
TREC Robust04	6	7.69
BRIGHT	4	5.13
Other (appeared 1-3 times)	23	29.49

3.3.3 RQ3. How are ML approaches for ranking in search systems evaluated, in terms of metrics and evaluation methodology?

Analysis of evaluation methodology showed that proposed models were evaluated with offline benchmarks. No online tests were reported. One study reported conducting a user study in addition to offline benchmarking [26]. Evaluation metrics were coded according to the framework described in the Methodology. Frequency counts are shown at TABLE 6. Ranking metrics were used in 89.74% of studies, metrics such as nDCG, MRR, MAP are the most used among ranking metrics [17], [12]. Non-ranking metrics appeared in 64.1% of studies, confirming that Precision, Recall metrics remain fundamental baselines for many ranking models [27]. [2] measures retrieval coverage using Precision@K and Recall@K. Efficiency metrics were used in 43 studies, mostly in papers studying transformer-based and LLM-based models, as shown in TABLE 7. Efficiency evaluation was particularly common in LLM-based studies (19 of 26 studies) and less common in transformer-based studies (17 of 38), showing that computational cost is a more explicit concern when ranking relies on generative inference rather than encoder-based scoring alone. [28], measures trade-off between latency and nDCG [29], calculates latency in terms of seconds per query. Generation quality metrics were used in 6 studies. [30], used BLEU and ROUGE-L to evaluate model quality.

Table 6: Top evaluation metric categories used in the reviewed studies.

Evaluation metrics	**Number of studies**	**Percentage (%)**
Ranking	70	89.74
Non-ranking	50	64.1
Efficiency	43	55.13
Generation quality	6	7.69

Table 7: Evaluation metric categories by model architecture in the reviewed studies.

	Ranking	Non-ranking	Efficiency	Generation quality
Classical LTR	4	3	1	0
Neural ranking	8	7	6	1
Transformer-based	33	26	17	2
LLM-based	25	14	19	3

3.3.4 RQ4. In what types of search systems and application domains are ML approaches for ranking studied?

Application domains were assigned using the coding rules defined in the Methodology. Frequency counts are presented in TABLE 8. The predominance of general and web search indicates that the evidence base is centred on broad retrieval tasks. In addition, a cross-tabulation of application domain by model architecture was computed, with the results shown in TABLE 9.

General and web search applications are among the most popular, and occur in half of studies. A big number of studies in the web search domain propose transformer-based and LLM-based models. [11], proposed BERT encoder for document ranking, [16], explored efficiency of zero-

shot LLM architecture for document retrieval and ranking. Generative search is specific mostly for transformer-based and LLM-based models. [31], studied RAG system, [32], proposed approach for QA. [23], proposes use of classical LTR meta-classifier to increase the quality of the RAG pipeline. Scientific search appeared in several studies, focusing on highly structured or academic corpora. [33], ranks mathematical expressions. In E-commerce search studies focused on product catalogues, attribute matching. [34], investigated product search with sparse retrieval [35], examined vulnerabilities in product recommendations after conversational search. Task-specific search relates to highly-specific domains, like information retrieval for telecommunication [1], code search [36, 37], manufacturing [38].

Table 8: Application domains in the reviewed studies.

Application domain	**Number of studies**	**Percentage (%)**
General and Web search	39	50
Task-specific search	16	20.51
Generative search	15	19.23
Scientific search	4	5.13
E-commerce search	4	5.13

Table 9: Application domains by model architecture in the reviewed studies.

	General and Web search	Task-specific search	Generative search	Scientific search	E-commerce search
Classical LTR	4	0	1	0	0
Neural ranking	6	1	1	1	0
Transformer-based	16	8	9	2	3
LLM-based	13	7	4	1	1

3.3.5 RQ5. What challenges, limitations, and research gaps are reported in the literature on ML methods for ranking in search systems?

Inference Cost, Scalability, Latency and Computational Overhead. A frequently reported challenge across the literature was the high computational expense and inference latency associated with cross-encoders [39], [4], bi-encoders [8], [13], and LLM-based re-rankers [40, 41], [29], [15]. These high computational costs hinder deployment of these models in real-time search applications. Moreover, in several studies trade-off between candidate pool size and system efficiency was reported [24], [42]. Finally, dependency on proprietary LLM APIs was reported as a barrier to practical deployment and scalability [17], [30], [43].

Data Limitations and Domain Generalization. Researchers repeatedly cite a heavy reliance on extensive human-annotated datasets as a critical limitation, resulting in models that struggle to generalize to unseen, low-resource, or highly specialized domains [1], [17], [25], [37]. [1], states that 'domain-specific performance is a persistent challenge' for LLMs. To mitigate issues with data scarcity researchers often relied on synthetic data generation; however, studies note that LLM-generated data do not perfectly align with the real distribution of text data, which can degrade performance [44, 45]. [44], mentioned that 'LLM-modified (expanded) queries led to a performance degradation when evaluated with human-generated queries'. Considering LLM-based re-rankers,

prominent limitation is the risk that evaluation queries may leak into training data in advance, so it becomes difficult to verify results [17], [15], [46].

Bias and Hallucinations. LLM-based rankers tend to hallucinate [17], [41], [15], [26, 27]. [26], states that LLMs 'can often make claims that appear plausible but are not grounded in reliable evidence'. Another recurring challenge is sensitivity of neural models to biases and adversarial vulnerabilities [47], [44, 45].

Interpretability and System Transparency. A lack of interpretability is often reported by researchers. Many neural and LLM-based ranking models function as "black boxes", making it difficult to understand the complex inner logic for ranking decisions [48, 49], [25], [4], [35], [30]. [35], states that 'conversational search engines are black-box and feature no principled or interpretable mechanism for ranking their outputs'. [30], mentions that LLM-based re-rankers 'lack explainability'. Some frameworks attempted to generate model explanations for their decisions but they still lack human evaluation and trustworthiness [17], [30].

4. DISCUSSION

4.1 Summary of Evidence

The literature analysis shows that ML-based ranking models in search systems can be organized into four main groups: classical LTR models, neural ranking models, transformer-based models and LLM-based models, as was proposed in [10]. Studies of classical LTR approaches consider integration of features that are coming as output from neural models [3], [23]. This suggests that recently classical LTR has been used as a lightweight layer on top of neural signals, particularly in re-ranking. Transformer-based and LLM-based studies represent the largest categories. However, this descriptive distribution should not be interpreted as a complete representation of the field. The apparent shift toward transformer-based and LLM-based ranking may reflect current research interest rather than demonstrated practical superiority in production systems [14], [17].

Most studies rely on publicly available datasets, both for training and evaluation of models, so that it is easy to compare efficiency between different approaches. MS MARCO is the standard baseline for quality evaluation, alongside the BEIR suite. Also there is a relatively new benchmark BRIGHT, which evaluates reasoning-intensive retrieval: a model's ability to answer long, complex questions. However, BRIGHT was used in only 4 of the 78 studies (TABLE 5), which indicates that reasoning-focused evaluation is still an emerging rather than an established practice.

Most studies use ranking and non-ranking metrics to evaluate ranking quality regardless of the ML approach used, confirming that these metrics remain primary quality indicators in the field (TABLE 6). For emerging ML approaches, like transformer-based and LLM-based, efficiency metrics are increasingly reported, reflecting their high computational cost, therefore it is important to measure their latency, computational overhead, etc. None of the 78 included studies report an online evaluation (RQ3): efficiency and accuracy metrics are calculated offline. This is a strong limitation for deployment in production, since offline gains in latency or nDCG do not necessarily translate into comparable behaviour under real infrastructure constraints[17].

For LLM-based models used in QA or RAG systems, generation-quality metrics such as BLEU and ROUGE-L are also used to assess output quality. This capability, however, is applied unevenly: generation-quality metrics appear in 6 studies (TABLE 6), while generative search accounts for 15 studies (TABLE 8). This gap suggests that most transformer-based and LLM-based ranking approaches are still evaluated with ranking metrics without assessing generation quality [17].

As this remains a relatively new domain, most proposed approaches are evaluated in research settings rather than integrated in production systems, primarily due to inference costs, latency, hallucination issues, despite reporting high offline accuracy [14], [17]. Together, these evaluation gaps suggest that the practical maturity of transformer-based and LLM-based ranking may be overstated relative to classical and neural approaches, a gap the future work should address.

4.2 Limitations

This scoping review is subject to several limitations that should be acknowledged:

- **Data-related constraints.** Only Open Access publications in English or Russian, published from 2023 onward, were included. While English predominates in machine learning research, this criterion excludes potentially valuable contributions in other languages. Furthermore, to capture emerging trends, time period scope was restricted to works published from 2023 onward. Although this temporal filter is justified by the rapid evolution of the field, it may introduce certain biases, particularly regarding transformer-based and LLM-based architectures, whose development has accelerated in recent years.
- **Database coverage.** The search was limited to IEEE Xplore and Lens.org. Although Lens.org aggregates records from multiple publishers, its coverage may not be complete. Relevant studies indexed only in ACM Digital Library, Scopus, Web of Science, SpringerLink, ScienceDirect, or ACL Anthology may therefore have been missed, resulting in potential database coverage bias.
- **Methodological considerations.** To answer specific research questions (e.g., RQ2–RQ5) it was important to construct distinct analytical groups and to develop study-specific taxonomy. Consequently, this taxonomy introduces a degree of subjectivity, which may inherently affect the adequacy of the subsequent findings.
- **Reviewer-related bias.** The review was primarily conducted by a single reviewer, which may have introduced reviewer-related selection and extraction bias. Although ambiguous cases were discussed with a colleague and resolved through consensus, independent duplicate screening and data extraction were not performed.
- **Absence of formal critical appraisal.** No formal critical appraisal was conducted, as the review aimed to map the scope and characteristics of the available literature rather than evaluate the comparative methodological quality of individual studies.

4.3 Future work directions

Ranking remains a critical component of modern search systems and these systems continue to evolve rapidly, therefore several directions for future work emerge from the findings of this scoping review. These directions are proposed with the aim of addressing the research gaps identified through conducted analysis.

- **Enhancing Computational Efficiency for LLMs.** The computational overhead of utilizing LLMs significantly hinders their real-time applicability in production environments [17]. Future research must focus on reducing model complexity and optimizing inference costs.
- **Improving Model Interpretability.** Current LLMs, neural and transformer-based rankers are often reported as "black-box" models, meaning that it is difficult to explain why specific documents are prioritized, which hinders trustworthiness. Future work should focus on explanation-based solutions [30].
- **Reducing hallucinations.** LLMs are highly vulnerable to hallucinations, especially when presented with noisy or irrelevant context. A critical future direction is the development of approaches that penalize hallucinations by scoring documents on faithfulness rather than just semantic relevance [50].
- **Evaluation beyond offline benchmarks.** Future studies should complement MS MARCO, BEIR, and TREC-style evaluation with online experiments and controlled user studies which are needed to establish whether offline improvements result in better user outcomes.
- **Domain adaptation with human validation.** Synthetic queries and documents may help address the shortage of relevance judgments in specialized domains, but their value should be verified against real user queries and expert annotations. Research should identify when synthetic data improves domain transfer and when it distorts the target distribution [27], [1].

5. CONCLUSION

The results of this scoping review systematize information on the main ML approaches used for ranking in search systems and identify key trends in recent years (2023-2026). To answer RQ1, studies were assigned to groups of ML approaches that were used for ranking. It showed that transformer-based and LLM-based models are prevalent these years, also there's temporal trend showing increasing interest in transformer-based and LLM-based model families in 2025 and 2026. In response to RQ2, it was shown that most studies relied on widely known and publicly available datasets for evaluation. In response to RQ3, it was shown that most models used non-ranking and ranking metrics for evaluation, which means that they remain strong quality indicators. RQ4 revealed main domains where models are used. An examination of the limitations, challenges, and research gaps identified in this review leads to the conclusion that future work should pay particular attention to issues surrounding transformer-based and LLM-based models. Several key challenges remain, including high inference costs, latency, and dependency on third-party LLM API providers. Addressing these issues is essential for the practical deployment of such models in production systems. In addition, many studies have highlighted the "black-box" nature of transformer-based

and LLM-based architectures, underscoring the need for improved interpretability and reduced hallucination rates in future work.

References

[1] Bindle A, Singla P, Sharma S, Khakimov A, Alkanhel RI, et al. A Hybrid Large Language Model for Context-Aware Document Ranking in Telecommunication Data. IEEE Access. 2025;13:120345-120359.

[2] Mitrov G, Stanoev B, Gievska S, Mirceva G, Zdravevski E. Combining Semantic Matching, Word Embeddings, Transformers, and Llms for Enhanced Document Ranking: Application in Systematic Reviews. Big Data Cogn Cmput. 2024;8:110.

[3] Singh AM, Singh DB, Pandey AR, Siddiqui MR, Sharma AP, et al. Integrating Deep Learning Techniques in Information Retrieval: A Hybrid Approach to Relevance Optimization. J. Inf. Syst. Eng. Manage. 2025;10:311-316.

[4] Haddad R, Hlaoua L, Omri MN. Regat-Bert: Transformer-Graph Fusion for Dynamic Reranking. Procedia Comput Sci. 2025;270:1816-1825.

[5] M S. GDRMA: Graph Neural Networks for Document Retrievals With Mean Aggregation. IEEE Access. 2024;12:185706-185727.

[6] Yamkovyi K. Adaptation of Lambdamart Model to Semi-Supervised Learning. Bull Natl Tech Univ KhPI Ser Syst Anal Control Inf Technol. 2023;1:76-81.

[7] Chang S, Ahn GJ, Park S. Improving Performance of Neural IR Models by Using a Keyword-Extraction-Based Weak-Supervision Method. IEEE Access. 2024;12:46851-46863.

[8] Lin J, Chen AH, Lassance C, Ma X, Pradeep R, et al. Gosling Grows Up: Retrieval With Learned Dense and Sparse Representations Using Anserini. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025:3223-3233.

[9] Adeyemi M, Oladipo A, Pradeep R, Lin J. Zero-Shot Cross-Lingual Reranking With Large Language Models for Low-Resource Languages. In: Proceedings of the 62nd annual meeting of the Association for Computational Linguistics. 2024;2:650-656.

[10] Xu Z, Mo F, Huang Z, Zhang C, Yu P, et al. A Survey of Model Architectures in Information Retrieval. 2025. Arxiv preprint: https://arxiv.org/pdf/2502.14822v1

[11] Kumar S, Rohatgi D, Prakash N, Sahai S, Chandra S, et al. 2P-Benc: A Two-Phase Information Retrieval and Ranking System Based on the Bert Encoder. Ain Shams Eng J. 2026;17:103853.

[12] Wang X, MacDonald C, Tonellotto N, Ounis I. Colbert-Prf: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval. ACM Trans Web. 2023;17:1-39.

[13] An GT, Lee KS. Colbert-Aw: Enhancing Late Interaction Retrieval With Attribute-Aware Query Token Weighting. IEEE Access. 2026;14:43227-43237.

[14] Sharifymoghaddam S, Pradeep R, Slavescu A, Nguyen R, Xu A, et al. Rankllm: A Python Package for Reranking With Llms. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025:3681-3690.

[15] Sun W, Yan L, Ma X, Wang S, Ren P, et al. 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. Association for Computational Linguistics. 2023:14918-14937.

[16] Lawrie D, Kayi E, Mayfield J, Yang E, Yates A, et al. A Reproducibility Study of Llm Setwise Reranker With Heapsort. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025:3145-3154.

[17] Abdallah A, Piryani B, Mozafari J, Ali M, Jatowt A. How Good Are Llm-Based Rerankers? An Empirical Analysis of State-Of-The-Art Reranking Models. In: Findings of the Association for Computational Linguistics: EMNLP 2025. Association for Computational Linguistics. 2025:5693-5709.

[18] Hoffmann M, Bergmann R. Ranking-Based Case Retrieval With Graph Neural Networks in Process-Oriented Case-Based Reasoning. InThe International FLAIRS Conference Proceedings. 2023;36.

[19] Ergashev U, Dragut E, Meng W. Learning to Rank Resources With Gnn. Proceedings of the ACM Web Conference. In Proceedings of the ACM Web Conference. 2023:3247-3256.

[20] Sangamithra B, Kumar MS. An Improved Information Retrieval System Using Hybrid Rnn Lstm for Multiple Search Engines. Commun Appl Nonlinear Anal. 2024;31:167-180.

[21] Arora N. Personalized Re-Ranking of Search Engine Results Based on User Preferences and Query Specificity. International Journal for Research in Engineering Application & Management. 2026:128.

[22] https://ijsrem.com/download/curriculum-vitae-analysis-using-machine-learning-algorithm

[23] Yildiz E, Bozkir AS. Learning to Fuse Retrieval Signals: A Lightweight Meta Relevance Classifier for Rag Pipelines. Procedia Comput Sci. 2025;269:1434-1443.

[24] Abdallah A, Mozafari J, Piryani B, Jatowt A. Rank: Zero-Shot Re-Ranking With Answer Scent for Document Retrieval. In Findings of the Association for Computational Linguistics: NAACL. 2025:2950-2970.

[25] Aitim A, Satybaldiyeva R. A Comparison of Kazakh Language Processing Models for Improving Semantic Search Results. East-Eur J Enterp Technol. 2025;1:66-75.

[26] Alt G, Hirsch E, Basch S, Dagan I, Glickman O. User-Centric Evidence Ranking for Attribution and Fact Verification. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2026:7215-7237.

[27] Divya C, Xue C, Sun S. Reasoning-Enhanced Retrieval for Misconception Prediction: A Rag-Inspired Approach With Llms. In Proceedings of The First Workshop on Human–LLM Collaboration for Ethical and Responsible Science Production (SciProdLLM). 2025:38-51.

[28] Gangi Reddy R, Doo J, Xu Y, Sultan MA, Swain D, et al. First: Faster Improved Listwise Reranking With Single Token Decoding. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024:8642-8652.

[29] Liu W, Ma X, Zhu Y, Su L, Wang S, et al. Coranking: Collaborative Ranking With Small and Large Ranking Agents. In EMNLP (Findings). 2025:5098-5110.

[30] Ji Y, Li Z, Meng R, He D. Reason-To-Rank: Distilling Direct and Comparative Reasoning From Large Language Models for Document Reranking. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025:2320-2329.

[31] Zhang P, Liu Z, Xiao S, Dou Z, Nie JY. A Multi-Task Embedder for Retrieval Augmented Llms. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024:3537-3553.

[32] Zhang Z, Vu T, Moschitti A. Double Retrieval and Ranking for Accurate Question Answering. In: Findings of the Association for Computational Linguistics: Eacl 2023. In Findings of the Association for Computational Linguistics: EACL. 2023:1751-1762.

[33] Li X, Tian B, Tian X. Scientific Documents Retrieval Based on Graph Convolutional Network and Hesitant Fuzzy Set. IEEE Access. 2023;11:27942-27954.

[34] An GT, Park JM, Lee KS. Neglu: Negation-Aware Sparse Retrieval With Negative Weights for Product Search. IEEE Access. 2025;13:144492-144504.

[35] Pfrommer S, Bai Y, Gautam T, Sojoudi S. Ranking Manipulation for Conversational Search Engines. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024:9523-9552.

[36] Bibi N, Maqbool A, Rana T, Afzal F, Akgül A, et al. Enhancing Semantic Code Search With Deep Graph Matching. IEEE Access. 2023:11:52392-52411.

[37] Nam G, Yang G. Llmloc: A Structure-Aware Retrieval System for Zero-Shot Bug Localization. Electronics. 2025;14:4343.

[38] Siouffi C, Glandon P, Kubler S, Soumianarayanan V. Llm-Based Contrastive Representation Learning for Enhanced Maintenance Work Order Retrieval. Procedia CIRP. 2025;134:981-986.

[39] Fang J, Meng Z, Macdonald C. KGPR: Knowledge Graph Enhanced Passage Ranking. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023:3880-3885.

[40] Cai Y, Zhang Y, Long D, Li M, Xie P, et al. Erank: Fusing Supervised Fine-Tuning and Reinforcement Learning for Effective and Efficient Text Reranking. In Proceedings of the AAAI Conference on Artificial Intelligence . 2026;40:30121-30129.

[41] http://dx.doi.org/10.5281/zenodo.17829256

[42] Borges L, Jha R, Callan J, Martins B. Generalizable Tip-Of-The-Tongue Retrieval With Llm Re-Ranking. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024:2437-2441.

[43] Magomere J, La Malfa E, Tonneau M, Kazemi A, Hale SA. When Claims Evolve: Evaluating and Enhancing the Robustness of Embedding Models Against Misinformation Edits. In Findings of the Association for Computational Linguistics: ACL. 2025:22374-22404.

[44] Nakachi Y, Kato MP. Impact of Llm-Modified Queries and Documents in Training Data on Neural Retrieval Models. In Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. 2025:364-373.

[45] Pan Z, Fan K, Liu R, Li D. Towards Robust Neural Rankers With Large Language Model: A Contrastive Training Approach. Appl Sci. 2023;13:10148.

[46] Zhuang S, Liu B, Koopman B, Zuccon G. Open-Source Large Language Models Are Strong Zero-Shot Query Likelihood Models for Document Ranking. In Findings of the Association for Computational Linguistics: EMNLP. 2023:8807-8817.

[47] Seyedsalehi S, Le HS, Zihayat M, Bagheri E. Bias-Aware Curriculum Sampling for Fair Ranking. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025:2774-2778.

[48] Biyyala S, Tokachichu SC, Chennuri S. Ai-Powered Search Systems: Integrating Machine Learning With Search Technology for High-Scalability Applications. Int J Sci Res Comput Sci Eng Inf Technol. 2024;10:74-81.

[49] Amala KJ, Rajeswari D. Neural Learning to Rank Model With Bias Correction and Attention Enhanced Relevance Prediction. IEEE Access. 2025;13:188399-188419.

[50] Guan T, Sun S, Chen B. Faithfulness-Aware Multi-Objective Context Ranking for Retrieval-Augmented Generation. In Proceedings of the 2025 3rd International Conference on Artificial Intelligence, Systems and Network Security. 2025:119-126.

[51] Lee J, Bang J, Shim K, Yang S, Chang S. Chain-Of-Rank: Enhancing Large Language Models for Domain-Specific Rag in Edge Device. In Findings of the Association for Computational Linguistics: NAACL. 2025:5601-5608.

[52] Lim H. The Impact of Query Decomposition and Cross-Encoder Reranking in Multi-Hop Retrieval-Augmented Generation. J Youth Impact. 2026;1.

[53] Gangi Reddy R, Dasigi P, Sultan MA, et al. A Large-Scale Study of Reranker Relevance Feedback at Inference. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025:3010-3014.

[54] Wang M, Liu J, Wang J, Wang Y, Chu X. A Topicality Relevance-Aware Intent Model for Web Search. IEEE Access. 2023;11:65739-65748.

[55] Fathallah M, El-Makky N, Torki M. Alexunlp-Fmt at Climatecheck Shared Task: Hybrid Retrieval With Adaptive Similarity Graph-Based Reranking for Climate-Related Social Media Claims Fact Checking. In Proceedings of the Fifth Workshop on Scholarly Document Processing. 2025:288-292.

[56] Albishre K. Anchored Dense-Chain Pseudo-Relevance Feedback: Sequential State Refinement for Neural Retrieval. J King Saud Univ - Comput Inf Sci. 2026;38:199.

[57] Zamiri M, Kotov A. Conversational Entity Retrieval From a Knowledge Graph Using Aggregation of Fine-Grained Relevance Signals With Graph Convolutions and Self-Attention. In Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining. 2026:902-912.

[58] Yoon J, Sael L. Document Re-Ranking With Evidential Neural Networks. IEEE Access. 2025;13:161964-161972.

[59] Yadav N, Monath N, Zaheer M, McCallum A. Efficient K-Nn Search With Cross-Encoders Using Adaptive Multi-Round Cur Decomposition. InFindings of the Association for Computational Linguistics: EMNLP. 2023:8088-8103.

[60] Zuo L, Hong P, Kraus O, Plank B, Litschko R. Evaluating Large Language Models for Cross-Lingual Retrieval. In: Findings of the Association for Computational Linguistics: EMNLP 2025. 2025:11415-11429.

[61] Katsimpras G, Paliouras G. Genra: Enhancing Zero-Shot Retrieval With Rank Aggregation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024:7566-7577.

[62] Killingback J, Zeng H, Hypencoder ZH. Hypernetworks for Information Retrieval. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025:2372-2383.

[63] Iturra-Bocaz G, Vo D, Galuščáková P. Impact of Shallow vs. Deep Relevance Judgments on BERT-based Reranking Models. Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR). Acad Med. 2025:326-335.

[64] Yang B. Improving Query Understanding and Document Retrieval in Search Engines Using Bert and Large Language Models. Intell Hum Future. 2026;2:292.

[65] Panchendrarajan R, Frade R, Zubiaga A. ClaimCatchers at SemEval-2025 Task 7: Sentence Transformers for Claim Retrieval. Published online July 2025.

[66] Krasakis AM, Yates A, Kanoulas E. Constructing Set-Compositional and Negated Representations for First-Stage Ranking. In Proceedings of the 34th ACM International Conference on Information and Knowledge Management. 2025:1406-1416.

[67] Askari A, Abolghasemi A, Pasi G, Kraaij W, Verberne S. Injecting the Score of the First-Stage Retriever as Text Improves Bert-Based Re-Rankers. Discov Comput. 2024;27:15.

[68] Dudek JM, Kong W, Li C, Zhang M, Bendersky M. Learning Sparse Lexical Representations Over Specified Vocabularies for Retrieval. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023:3865-3869.

[69] http://dx.doi.org/10.5281/zenodo.17822947

[70] Song T, Zhao Y, Zhang S, Zhao C, Cohan A. Limrank: Less Is More for Reasoning-Intensive Information Reranking. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025:20636-20650.

[71] Xu S, Pang L, Xu J, Shen H, Cheng X. List-Aware Reranking-Truncation Joint Model for Search and Retrieval-Augmented Generation. In Proceedings of the ACM Web Conference. 2024:1330-1340.

[72] He S, Xie W, Qiao Y, Carlson P, Yang T. Low-Cost Document Retrieval With Dense Pseudo-Query Encoding. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025:2987-2993.

[73] Khamnuansin D, Chalothorn T, Chuangsuwanich E. Mrrank: Improving Question Answering Retrieval System Through Multi-Result Ranking Model. In Findings of the Association for Computational Linguistics: ACL. 2024:4750-4762.

[74] Yoon S, Kim J, Kwon D, Anand A, Hwang SW. On Listwise Reranking for Corpus Feedback. In Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining. 2026:1273-1277.

[75] Coman A, Barlacchi G, De Gispert A. Strong and Efficient Baselines for Open Domain Conversational Question Answering. In Findings of the Association for Computational Linguistics: EMNLP. 2023:6305-6314.

[76] Wang B, Zhou G, Xi Y, Yu W. Trirank: A Cognitive-Inspired Three-Layer Adaptive Reranking Framework. In Proceedings of the 4th International Conference on Artificial Intelligence and Intelligent Information Processing. 2025:794-798.

[77] Ezerceli Ö, El Hussieni M, Taş S, Bayraktar R, Terzioğlu FB, et al. Turkcolbert: A Benchmark of Dense and Late-Interaction Models for Turkish Information Retrieval. Procedia Comput Sci. 2026;275:393-400.

[78] Pan M, Zhou S, Li T, Liu Y, Pei Q, Huang AJ, et al. Utilizing Passage-Level Relevance and Kernel Pooling for Enhancing Bertbased Document Reranking. Comput Intell. 2024;40:e12656.

[79] Yang Y, Qiao Y, He S, Yang T. Weighted Kl-Divergence for Document Ranking Model Refinement. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024:2698-2702.

6. Appendix A

Ref.	Title	Year	Model family	Ranking stage	Dataset	Metrics	Application domain
[1]	A Hybrid Large Language Model for Context-Aware Document Ranking in Telecommunication Data	2025	Transformer-based	re-ranking	Proprietary	Non-ranking, Ranking, Generation quality	Task-specific search
[2]	Combining Semantic Matching, Word Embeddings, Transformers, and LLMs for Enhanced Document Ranking: Application in Systematic Reviews	2024	Transformer-based	re-ranking	Proprietary	Non-ranking, Efficiency	General and Web search
[3]	Integrating Deep Learning Techniques in Information Retrieval: A Hybrid Approach to Relevance Optimization	2025	Classical LTR	end-to-end	MS MARCO, TREC DL	Non-ranking, Ranking	General and Web search
[4]	ReGAT-BERT: Transformer-Graph Fusion for Dynamic Reranking	2025	Transformer-based	re-ranking	MS MARCO	Ranking, Non-ranking	General and Web search
[5]	GDRMA: Graph Neural Networks for Document Retrievals With Mean Aggregation	2024	Neural Ranking	end-to-end	Other	Non-ranking, Ranking, Efficiency	General and Web search

Ref.	Title	Year	Model family	Ranking stage	Dataset	Metrics	Application domain
[6]	ADAPTATION OF LAMBDAMART MODEL TO SEMI-SUPERVISED LEARNING	2023	Classical LTR	re-ranking	Other	Ranking	General and Web search
[7]	Improving Performance of Neural IR Models by Using a Keyword-Extraction-Based Weak-Supervision Method	2024	Transformer-based	re-ranking	MS MARCO	Non-ranking, Ranking	General and Web search
[8]	Gosling Grows Up: Retrieval with Learned Dense and Sparse Representations Using Anserini	2025	Transformer-based	first-stage	MS MARCO, BEIR	Ranking, Non-ranking, Efficiency	General and Web search
[9]	Zero-Shot Cross-Lingual Reranking with Large Language Models for Low-Resource Languages	2024	LLM-based	re-ranking	Other	Ranking, Generation quality	General and Web search
[11]	2P-BEnc: A two-phase information retrieval and ranking system based on the BERT encoder	2026	Transformer-based	end-to-end	MS MARCO, BEIR	Non-ranking, Ranking, Efficiency	General and Web search
[12]	ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval	2023	Transformer-based	end-to-end	TREC DL, TREC Robust04	Non-ranking, Ranking, Efficiency	E-commerce search
[13]	ColBERT-AW: Enhancing Late Interaction Retrieval With Attribute-Aware Query Token Weighting	2026	Transformer-based	first-stage	Other	Ranking, Non-ranking	Task-specific search
[14]	RankLLM: A Python Package for Reranking with LLMs	2025	LLM-based	re-ranking	TREC DL, BEIR	Ranking, Non-ranking	QA and generative search
[15]	Is ChatGPT Good at Search- Investigating Large Language Models as Re-Ranking Agents	2023	LLM-based	re-ranking	TREC DL, BEIR	Ranking, Efficiency	General and Web search
[16]	A Reproducibility Study of LLM Setwise Reranker with Heapsort	2025	LLM-based	re-ranking	BEIR, TREC DL, TREC Robust04	Ranking, Efficiency	General and Web search
[17]	How Good are LLM-based Rerankers- An Empirical Analysis of State-of-the-Art Reranking Models	2025	LLM-based	re-ranking	TREC DL, BEIR, MS MARCO	Non-ranking, Ranking, Efficiency	General and Web search
[18]	Ranking-Based Case Retrieval with Graph Neural Networks in Process-Oriented Case-Based Reasoning	2023	Neural Ranking	re-ranking	Proprietary	Non-ranking, Ranking, Efficiency	Task-specific search
[19]	Learning To Rank Resources with GNN	2023	Neural Ranking	first-stage	Other	Non-ranking, Ranking	General and Web search
[20]	An Improved Information Retrieval System using Hybrid RNN LSTM for Multiple Search Engines	2024	Neural Ranking	re-ranking	Other	Non-ranking	General and Web search
[21]	Personalized Re-Ranking of Search Engine Results Based on User Preferences and Query Specificity	2026	Neural Ranking	re-ranking	Other	Non-ranking, Ranking	General and Web search
[22]	CURRICULUM VITAE ANALYSIS USING MACHINE LEARNING ALGORITHM	2024	Classical LTR	re-ranking	Proprietary	Non-ranking	General and Web search
[23]	Learning to Fuse Retrieval Signals: A Lightweight Meta Relevance Classifier for RAG Pipelines	2025	Classical LTR	re-ranking	NQ, HotpotQA	Non-ranking, Ranking	QA and generative search
[24]	ASRank: Zero-Shot Re-Ranking with Answer Scent for Document Retrieval	2025	LLM-based	re-ranking	MS MARCO, BEIR, TREC DL	Non-ranking, Ranking, Efficiency	General and Web search
[25]	A comparison of Kazakh language processing models for improving semantic search results	2025	Transformer-based	end-to-end	Proprietary	Non-ranking, Ranking	Task-specific search
[26]	User-Centric Evidence Ranking for Attribution and Fact Verification	2026	LLM-based	re-ranking	Other	Ranking, Non-ranking, Efficiency	Task-specific search
[27]	Reasoning-Enhanced Retrieval for Misconception Prediction: A RAG-Inspired Approach with LLMs	2025	LLM-based	re-ranking	Proprietary	Non-ranking, Ranking, Efficiency	Task-specific search
[28]	FIRST: Faster Improved Listwise Reranking with Single Token Decoding	2024	LLM-based	re-ranking	MS MARCO, BEIR	Ranking, Non-ranking, Efficiency	QA and generative search
[29]	CoRanking: Collaborative Ranking with Small and Large Ranking Agents	2025	LLM-based	re-ranking	TREC DL, BEIR, BRIGHT	Ranking, Efficiency	Task-specific search
[30]	Reason-to-Rank: Distilling Direct and Comparative Reasoning from Large Language Models for Document Reranking	2025	LLM-based	re-ranking	MS MARCO, BEIR, BRIGHT	Ranking, Generation quality, Efficiency	General and Web search
[31]	A Multi-Task Embedder For Retrieval Augmented LLMs	2024	Transformer-based	first-stage	MS MARCO, NQ	Non-ranking, Ranking, Generation quality	QA and generative search
[32]	Double Retrieval and Ranking for Accurate Question Answering	2023	Transformer-based	end-to-end	HotpotQA	Ranking, Efficiency	QA and generative search
[33]	Scientific Documents Retrieval Based on Graph Convolutional Network and Hesitant Fuzzy Set	2023	Neural Ranking	end-to-end	Other	Ranking, Efficiency	Scientific search
[34]	NeGLU: Negation-Aware Sparse Retrieval With Negative Weights for Product Search	2025	Transformer-based	first-stage	Other	Ranking, Non-ranking	E-commerce search
[35]	Ranking Manipulation for Conversational Search Engines	2024	LLM-based	re-ranking	Proprietary	Ranking, Efficiency	E-commerce search
[36]	Enhancing Semantic Code Search With Deep Graph Matching	2023	Transformer-based	end-to-end	Other	Ranking, Efficiency	Task-specific search

Ref.	Title	Year	Model family	Ranking stage	Dataset	Metrics	Application domain
[37]	LLMLoc: A Structure-Aware Retrieval System for Zero-Shot Bug Localization	2025	Transformer-based	re-ranking	Other	Non-ranking, Ranking, Efficiency	Task-specific search
[38]	LLM-based Contrastive Representation Learning for Enhanced Maintenance Work Order Retrieval	2025	LLM-based	first-stage	Proprietary	Ranking, Non-ranking	Task-specific search
[39]	KGPR: Knowledge Graph Enhanced Passage Ranking	2023	Transformer-based	re-ranking	MS MARCO, TREC DL	Ranking, Efficiency	General and Web search
[40]	ERank: Fusing Supervised Fine-Tuning and Reinforcement Learning for Effective and Efficient Text Reranking	2026	LLM-based	re-ranking	MS MARCO, TREC DL, BRIGHT, BEIR	Ranking, Efficiency	Task-specific search
[41]	Adaptive Generative Reranking: Leveraging LLMs for Dynamic Information Synthesis	2025	LLM-based	re-ranking	Other	Non-ranking, Ranking	General and Web search
[42]	Generalizable Tip-of-the-Tongue Retrieval with LLM Re-ranking	2024	LLM-based	re-ranking	Other	Non-ranking, Ranking	Scientific search
[43]	When Claims Evolve: Evaluating and Enhancing the Robustness of Embedding Models Against Misinformation Edits	2025	Transformer-based	end-to-end	Proprietary	Ranking	Task-specific search
[44]	Impact of LLM-Modified Queries and Documents in Training Data on Neural Retrieval Models	2025	Transformer-based	first-stage	MS MARCO, NQ, HotpotQA, TREC DL	Ranking	General and Web search
[45]	Towards Robust Neural Rankers with Large Language Model: A Contrastive Training Approach	2023	Transformer-based	re-ranking	Other	Ranking, Non-ranking	QA and generative search
[46]	Open-source Large Language Models are Strong Zero-shot Query Likelihood Models for Document Ranking	2023	LLM-based	re-ranking	MS MARCO, BEIR, TREC Robust04	Ranking	General and Web search
[47]	Bias-Aware Curriculum Sampling For Fair Ranking	2025	Transformer-based	re-ranking	MS MARCO	Ranking	QA and generative search
[48]	AI-Powered Search Systems : Integrating Machine Learning with Search Technology for High-Scalability Applications	2024	Classical LTR	end-to-end	Other	Ranking, Efficiency	General and Web search
[49]	Neural Learning to Rank Model With Bias Correction and Attention Enhanced Relevance Prediction	2025	Neural Ranking	re-ranking	Other	Ranking, Efficiency	General and Web search
[50]	Faithfulness-Aware Multi-Objective Context Ranking for Retrieval-Augmented Generation	2025	LLM-based	re-ranking	MS MARCO, NQ, HotpotQA	Non-ranking, Ranking, Efficiency	General and Web search
[51]	Chain-of-Rank: Enhancing Large Language Models for Domain-Specific RAG in Edge Device	2025	LLM-based	re-ranking	HotpotQA	Efficiency	General and Web search
[52]	The Impact of Query Decomposition and Cross-Encoder Reranking in Multi-Hop Retrieval-Augmented Generation	2026	Transformer-based	re-ranking	MS MARCO, HotpotQA	Non-ranking, Efficiency	QA and generative search
[53]	A Large-Scale Study of Reranker Relevance Feedback at Inference	2025	Transformer-based	end-to-end	MS MARCO, BEIR	Non-ranking, Ranking, Efficiency	General and Web search
[54]	A Topicality Relevance-Aware Intent Model for Web Search	2023	Transformer-based	re-ranking	Other	Ranking	General and Web search
[55]	AlexUNLP-FMT at ClimateCheck Shared Task: Hybrid Retrieval with Adaptive Similarity Graph-based Reranking for Climate-related Social Media Claims Fact Checking	2025	Transformer-based	end-to-end	Other	Non-ranking, Ranking	Scientific search
[56]	Anchored dense-chain pseudo-relevance feedback: sequential state refinement for neural retrieval	2026	Transformer-based	re-ranking	BEIR	Non-ranking, Ranking, Efficiency	General and Web search
[57]	Conversational Entity Retrieval from a Knowledge Graph using Aggregation of Fine-grained Relevance Signals with Graph Convolutions and Self-Attention	2026	Neural Ranking	re-ranking	Other	Non-ranking, Ranking, Efficiency	General and Web search
[58]	Document Re-Ranking With Evidential Neural Networks	2025	Transformer-based	re-ranking	MS MARCO, NQ, HotpotQA	Ranking, Non-ranking	Task-specific search
[59]	Efficient k-NN Search with Cross-Encoders using Adaptive Multi-Round CUR Decomposition	2023	Transformer-based	end-to-end	BEIR	Non-ranking, Efficiency	QA and generative search
[60]	Evaluating Large Language Models for Cross-Lingual Retrieval	2025	LLM-based	re-ranking	BEIR, MS MARCO	Ranking	General and Web search
[61]	GENRA: Enhancing Zero-shot Retrieval with Rank Aggregation	2024	LLM-based	end-to-end	TREC DL, BEIR	Ranking, Non-ranking, Generation quality, Efficiency	Task-specific search
[62]	Hypencoder: Hypernetworks for Information Retrieval	2025	Transformer-based	first-stage	MS MARCO, TREC DL, TREC Robust04	Ranking, Non-ranking, Efficiency	General and Web search
[63]	Impact of Shallow vs. Deep Relevance Judgments on BERT-based Reranking Models	2025	Transformer-based	re-ranking	MS MARCO, TREC DL	Ranking	General and Web search

Ref.	Title	Year	Model family	Ranking stage	Dataset	Metrics	Application domain
[64]	Improving Query Understanding and Document Retrieval in Search Engines Using BERT and Large Language Models	2026	LLM-based	re-ranking	MS MARCO	Ranking, Non-ranking, Efficiency	Task-specific search
[65]	ClaimCatchers at SemEval-2025 Task 7: Sentence Transformers for Claim Retrieval	2025	Transformer-based	end-to-end	Other	Non-ranking	QA and generative search
[66]	Constructing Set-Compositional and Negated Representations for First-Stage Ranking	2025	Transformer-based	first-stage	Other	Ranking, Non-ranking	Scientific search
[67]	Injecting the score of the first-stage retriever as text improves BERT-based re-rankers	2024	Transformer-based	re-ranking	MS MARCO, TREC DL	Ranking	General and Web search
[68]	Learning Sparse Lexical Representations Over Specified Vocabularies for Retrieval	2023	Transformer-based	first-stage	BEIR, MS MARCO	Non-ranking, Ranking, Efficiency	E-commerce search
[69]	Contextualized Reranking: Beyond Pointwise Relevance in Neural Information Retrieval	2025	Transformer-based	re-ranking	MS MARCO	Ranking	Task-specific search
[70]	LimRank: Less is More for Reasoning-Intensive Information Reranking	2025	LLM-based	re-ranking	MS MARCO, BRIGHT	Ranking, Non-ranking, Efficiency	General and Web search
[71]	List-aware Reranking-Truncation Joint Model for Search and Retrieval-augmented Generation	2024	LLM-based	re-ranking	NQ	Ranking, Efficiency	QA and generative search
[72]	Low-Cost Document Retrieval with Dense Pseudo-Query Encoding	2025	LLM-based	re-ranking	MS MARCO, TREC DL, BEIR	Non-ranking, Ranking, Efficiency	QA and generative search
[73]	MrRank: Improving Question Answering Retrieval System through Multi-Result Ranking Model	2024	Neural Ranking	re-ranking	MS MARCO, BEIR	Non-ranking, Ranking, Efficiency, Generation quality	QA and generative search
[74]	On Listwise Reranking for Corpus Feedback	2026	LLM-based	re-ranking	MS MARCO, TREC DL, TREC Robust04	Ranking, Efficiency	General and Web search
[75]	Strong and Efficient Baselines for Open Domain Conversational Question Answering	2023	Transformer-based	end-to-end	Other	Non-ranking, Efficiency	QA and generative search
[76]	TriRank: A Cognitive-Inspired Three-Layer Adaptive Reranking Framework	2025	Transformer-based	re-ranking	MS MARCO	Ranking	QA and generative search
[77]	TurkColBERT: A Benchmark of Dense and Late-Interaction Models for Turkish Information Retrieval	2026	Transformer-based	end-to-end	BEIR, MS MARCO	Non-ranking, Ranking, Efficiency	General and Web search
[78]	Utilizing passage-level relevance and kernel pooling for enhancing BERT-based document reranking	2024	Transformer-based	re-ranking	TREC Robust04, MS MARCO	Non-ranking, Ranking, Efficiency	General and Web search
[79]	Weighted KL-Divergence for Document Ranking Model Refinement	2024	Transformer-based	end-to-end	MS MARCO, TREC DL, BEIR	Ranking	General and Web search