MIRA: An LLM-Assisted Benchmark for Multi-Category Integrated Retrieval

Mehmet Deniz Türkmen
GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
deniz.tuerkmen@gesis.org

Suchana Datta
University College Dublin, Ireland
suchana.datta@ucd.ie

Dwaipayan Roy
Indian Institute of Science Education and Research, Kolkata, India
dwaipayan.roy@iiserkol.ac.in

Daniel Hienert
GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
daniel.hienert@gesis.org

Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
philipp.mayr@gesis.org

Derek Greene
University College Dublin, Ireland
derek.greene@ucd.ie

Abstract

Users increasingly expect modern search systems to offer a unified interface that seamlessly retrieves information from diverse data sources and formats. However, current information retrieval (IR) evaluation benchmarks have not kept pace with this development, primarily due to the lack of test collections that represent the diversity of contemporary search domains. We address this critical gap with MIRA, a novel benchmark based on a large-scale social science search platform. MIRA is designed for category-aware ranking across heterogeneous categories – `Publications`, `Research Data`, `Variables`, and `Instruments & Tools` – within a single, unified evaluation framework. The proposed collection is distinctive in several ways: (1) it is built upon real user queries, providing a more realistic basis for evaluation; (2) it covers scholarly items from four distinct categories, enabling multi-faceted evaluation; and (3) it leverages a Large Language Model to generate topic descriptions and narratives, as well as for relevance assessment with respect to these topics, substantially reducing the labor and cost of test collection generation. We release this resource to benefit the community by providing a foundational testbed for the research on multi-faceted, category-aware, integrated, or cross-category information retrieval.[1]

CCS Concepts

• **Information systems** → **Test collections**; **Relevance assessment**; **Evaluation of retrieval results**.

Keywords

Test Collection; Information Retrieval Evaluation; Relevance Judgments; Large Language Models

1 Introduction

In recent years, the paradigm of information retrieval (IR) has shifted steadily from siloed, single-type search experiences (e.g., "find a document") towards more integrated, multi-faceted user journeys in which users expect to pose a query and retrieve across heterogeneous content types [19]. Current IR evaluation benchmarks often do not reflect this complexity, primarily due to a shortage of test collections that encompass such a multi-categorical environment [12]. This limitation hinders the comprehensive assessment of IR systems in real-world scenarios [61]. It is particularly evident in digital libraries and scholarly search platforms, where a single information need may be satisfied by a combination of research papers, underlying datasets, measurement variables, or specialized software tools [31]. Users entering a query might reasonably expect to see relevant research artifacts – all in the same search interface. However, existing IR evaluation benchmarks have remained largely monolithic, targeting a single item type – such as passages, documents, or answers – and rarely mirroring the cross-category integration characteristic of modern search environments. The shift towards a multi-faceted information landscape requires IR systems capable of effectively navigating and ranking items across multiple categories simultaneously.

The advancement of IR systems is intrinsically linked to the availability of high-quality, realistic evaluation benchmarks [57]. Historically, the IR community relied on Cranfield-style test collections – comprising documents, topics, and relevance judgments – to drive progress in a controlled and comparable manner [13]. While invaluable, many established benchmarks, such as the TREC ad-hoc tracks [62] and CLEF campaigns [48], have often been constrained to a single predominant document type, such as newswire articles or web pages. This creates a critical gap between the simplified environments used for evaluation and the complex, multi-categorical reality of contemporary search platforms. Without test collections that mirror this integrated landscape, it is impossible to robustly evaluate and compare the performance of next-generation IR systems designed for it.

The challenge of creating such multi-categorical benchmarks is non-trivial. The process is traditionally labor-intensive and expensive, requiring significant human effort to create search topics and, most critically, to annotate relevance judgments [68]. This cost is multiplied when a collection spans disparate item types, as the relevance criteria for a research dataset, for example, differ substantially from those for a scholarly publication [35]. Consequently, the community lacks resources to study cross-category retrieval, result

[1] https://github.com/suchanadatta/MIRA-LLM-Assisted-Benchmark.git

fusion, and the nuanced user behavior associated with complex search tasks spanning multiple information types.

Recent advancements in Large Language Models (LLMs) offer a promising opportunity to address the scalability challenges of test collection creation. LLMs have demonstrated strong capabilities in text understanding and generation, and their application to IR tasks has become an active area of research. Prior work has explored using LLMs for tasks, such as query expansion [64] and synthetic document generation [51]. More relevantly, initial investigations have begun to assess the potential of LLMs as relevance judges, with studies comparing their annotations to those of human assessors [20]. While these studies caution against directly replacing human judgment in all scenarios, they do highlight the potential of LLMs as a powerful means to augment and scale the curation process, particularly for constructing novel benchmarks where traditional methods are prohibitively expensive.

In this paper, we address a critical gap in IR evaluation by introducing a novel, multi-categorical test collection built from the user logs of a large-scale social science search platform. Our benchmark is designed for the new paradigm of category-aware integrated search, encompassing four distinct scholarly categories — `Publications`, `Research Data`, `Variables`, and `Instruments & Tools` — within a single, unified evaluation framework. The proposed test collection is distinctive in several ways as follows:

(1) It provides comprehensive coverage of four distinct scholarly item categories, enabling a multi-faceted evaluation of integrated IR systems.
(2) It is grounded in real user queries, ensuring that the resulting evaluation topics reflect authentic information needs.
(3) It innovatively leverages an LLM to streamline the generation of topic descriptions and narratives, and to annotate item relevance. This approach significantly enhances the efficiency of the test collection generation process, reducing labor and costs, and offers a viable model for future benchmark development.

2 Related Work

Our work lies at the intersection of several key areas in IR research: the evolution of test collections, the specific challenges of multi-categorical search, and the emerging use of LLMs to automate and scale evaluation processes. This section reviews the relevant literature in these domains.

Cranfield Paradigm and Evolving Test Collections. The foundation of modern IR evaluation is the Cranfield paradigm, which introduced the use of test collections – a *document corpus*, *topics*, and *relevance judgments* – to compare the retrieval systems in a controlled manner [13]. This model has driven major progress through large-scale campaigns such as the TREC [5], CLEF [1], NTCIR [4] , and FIRE [2], producing invaluable resources for evaluating ad-hoc retrieval [62], web search [16], complex question-answering [46] and many other tasks. However, as search applications diversify, the limits of single-category collections have become evident. Benchmarks like TREC-CAR [17] and MS MARCO [15] mark progress but still focus on retrieving a single information type (e.g., text passages). Creating such collections remains resource-intensive, requiring extensive human effort for topic design and relevance assessment, which constrains the development of new resources [57, 68]. Recent work has explored ways to improve the static test-collection model to better reflect the dynamics of real-world search systems [32].

Test Collections for Multi-Categorical Search. Scholarly search demands multi-categorical retrieval, with information needs spanning publications, datasets, code, and other research objects [35]. Existing benchmarks address only limited aspects: ACL OCL includes papers, authors, and venues but is confined to computational linguistics [56]; TREC Chemical retrieves documents and patents but not heterogeneous item types [41]; and MMDocIR targets multimodal document elements rather than distinct object categories [18]. Dataset search has been examined separately through log analyses and tailored relevance criteria (such as [31], [11], and [36]), yet these efforts treat datasets in isolation. Our work builds on this understanding of domain-specific relevance and extends it by integrating multiple, distinct scholarly categories into a single benchmark. To the best of our knowledge, no publicly available benchmark combines four heterogeneous research object types within a unified evaluation framework derived from real user logs.

Leveraging Real User Logs. Constructing test collections from real user logs enhances ecological validity by capturing authentic, under-specified information needs [6]. Foundational analyses, such as those of the AOL log [47], revealed the complex and often under-specified nature of genuine information needs. This approach is vital in scholarly and professional search, where log studies of digital libraries [22, 25, 29, 60] and data repositories [31] reveal iterative expert behaviors. The main challenge, however, is to obtain reliable large-scale relevance judgments. A common solution is to use implicit feedback as a relevance proxy [30, 65], exemplified by MS MARCO [15], which paired Bing queries with clicked passages to produce weak labels at scale. This paradigm extends to domains such as clinical search [59] and legal question answering in COLIEE [34], proving its effectiveness for complex, professional retrieval. Following this tradition, we derive topics from logs of a large-scale social science search platform, grounding our multi-categorical benchmark in authentic researcher challenges and advancing evaluation beyond synthetic settings toward realistic, integrated retrieval assessment.

Large Language Models for IR Evaluation. LLMs have transformed IR, advancing document ranking, system evaluation, and addressing the long-standing scalability constraints of the Cranfield paradigm [7, 23, 50, 58, 66, 67]. A key research direction examines LLMs as relevance assessors capable of mirroring human judgment. Findings are mixed: Faggioli et al. [20] demonstrate that while LLMs align well with humans on clear topics, performance is highly sensitive to prompt design and task complexity.

Beyond direct relevance assessment, LLMs are increasingly used to automate test collection development – generating synthetic topics and narratives to expand collections or simulate new scenarios [9, 63]. Their instruction-following capability also aids in creating assessment guidelines and judgment explanations, improving assessor training and consistency [21, 40]. We combine these approaches by employing an LLM to generate topic descriptions, narratives, and initial relevance annotations across four scholarly

categories, mitigating labor demands while leveraging scalability and established best practices.

Following [26], we used dedicated user interactions (e.g., `views`, `downloads`) as initial relevance indicators. However, both LLM-based and human analyses revealed frequent mismatches with topical relevance (e.g., viewing irrelevant records). We therefore adopted a hybrid approach in which the LLM assisted human assessors in producing final explicit judgments, rather than treating implicit signals as direct proxies.

3 GESIS Search: An Integrated Search System

Our test collection is derived from the user logs of GESIS Search [3], an integrated digital library and search platform for the social sciences [26]. GESIS Search is designed to support the entire research life-cycle — from study design and data collection to publication — it provides a unified interface to diverse scholarly resources, enabling users to find and interlink datasets, publications, survey variables, and methodological instruments. The platform serves social scientists and related researchers, attracting about 30,000 unique users with 250,000 page impressions per month.

A key feature of GESIS Search is its rich, interlinked metadata model. Information items are not isolated; they are connected through semantic relationships, allowing users to navigate the scholarly ecosystem. For instance, a user can discover a publication and then directly navigate to the underlying research dataset it analyzes, or examine the specific survey variables and questionnaires used to collect the data. This network of links facilitates exploratory search and enhances scientific transparency and reproducibility. The platform supports this with advanced search filters, enabling users to refine their searches across different resource types, study collections, and recommendations to related content. Furthermore, the ability to directly download datasets and materials empowers researchers to reuse existing data for their own research questions.

The content in GESIS Search is organized into six primary categories that represent the core information needs of empirical social science research. The scale and specificity of these collections make it an ideal foundation for building a multi-categorical IR test collection. The categories are as follows:

- **`Research Data`**: Approximately 7,700 quantitative social science research datasets, with temporal coverage from 1945 to 2026 and a primary geographic focus on Germany and Europe.
- **`Variables`**: A fine-grained collection of approximately 1.4 million survey variables extracted from questionnaires. Roughly 200,000 of these are high-quality variables, accompanied by detailed metadata including full question texts, answer categories, and frequency tables.
- **`Instruments & Tools`**: Around 600 methodological resources to aid in survey design and data collection. This includes pretested questionnaires, validated response scales, and analysis syntax files for statistical software, alongside guides for collecting survey and digital behavioral data.
- **`Publications`**: A corpus of approximately 260,000 publications sourced from open-access repositories and literature that is explicitly linked to research data, supporting literature reviews and meta-analyses.
- **`GESIS Library`**: A specialized collection of approximately 122,000 publications from the GESIS internal library, with a strong focus on empirical social research and applied computer science.
- **`GESIS Webpages`**: Roughly 2,700 informational pages from the GESIS website, detailing consulting services, training workshops, and other support offerings for social scientists.

For the purpose of constructing our multi-categorical benchmark, we focus on the four core scientific resource types: **`Publications`**, **`Research Data`**, **`Variables`**, and **`Instruments & Tools`** that are also represented in the GESIS KG [8] with their semantic relationships. This selection captures the essential, interlinked dimensions of social science research.

4 MIRA: A Multi-Category Integrated Retrieval Assessment Dataset

To address the gap in evaluating integrated search systems, we introduce **MIRA**, a **M**ulti-Category **I**ntegrated **R**etrieval **A**ssessment test collection. The collection is constructed from real-world user interactions and data from the GESIS Search platform, providing a realistic and multi-faceted benchmark for scholarly IR. This section details its core components: the document collection, topics, and relevance judgments – and outlines its availability. The metadata is available under a Creative Commons Attribution 4.0 license, while the accompanying software is distributed under a GNU General Public License v3.0. All resources are publicly available in the project's GitHub repository, including prompt templates, generation parameters, pooling strategies, relevance judgments, and hyperparameter settings, along with all code and a quickstart guide for using the dataset.

4.1 Collection

The MIRA collection is a snapshot of the four core scholarly resource types within GESIS Search (see Section 3), encompassing a diverse set of social science information. The collection contains metadata for 7,634 research datasets, 206,434 high-quality metadata variables, 604 instruments & tools, and 254,097 publications totaling 468,769 documents, provided as a set of JSON files.

Figure 1: Top-50 topic word cloud from topic modeling.

4.2 Topics

The topics in MIRA originate from real user queries submitted to the GESIS Search platform. We used user logs collected between 2017 and 2024, comprising 16,335,937 interactions that capture various user interactions during their search sessions. To identify

interactions that indicate user interest and potential relevance between a query and a resource, we considered three action types (taking a cue from [26]): `View record` (a click to show the detailed view of a record), `Download` (downloading data or documents), and `Export` (exporting citation information). Based on these actions, we extracted corresponding query–item pairs, resulting in 412,032 pairs across the four categories in the data.

The initial pool of multi-category queries contained significant semantic overlap, as users frequently expressed core information needs through multiple query variants. Therefore, we performed query clustering to group these variations, ensuring that the final set of topics is both comprehensive and non-redundant. To transform the raw query-item pairs into a coherent set of non-redundant evaluation topics, we performed clustering on the 412,032 pre-selected queries using `BERTopic` [24]. Query embeddings were generated with the multilingual model, *paraphrase-multilingual-MiniLM-L12-v2* [52] to handle our mixed German–English dataset, producing 384-dimensional vectors. These vectors were further reduced to five dimensions using UMAP to improve clustering accuracy and efficiency, and were then grouped using the HDBSCAN [10].

This process effectively clustered semantically-related queries (e.g., 'environment' vs. 'sustainability'), topical variants (e.g., 'migration in Germany' vs. 'migration in the UK'), and translations (e.g., 'Beruf' vs. 'job') into coherent topics. After removing known-item searches (e.g., 'ALLBUS'), we retained the 50 most frequent clusters for further analysis (see Figure 1). From the refined clusters, we selected the most frequent queries across all four categories and restricted them to those with more than two terms to ensure specificity. This produced a test collection that supports robust multi-category evaluation and clearer differentiation in retrieval effectiveness. As a result, we obtained a set of 200 topics for our test collection, out of which 145 are in German and rest 55 are in English. For each resulting topic, we create a structured representation that includes the original query, a full description, and a detailed narrative. The description and narrative are uniquely created for each category in our dataset. These descriptions and narratives are generated using an LLM, guided by a purpose-built prompt designed to produce category-aware topic expansions. The exact prompt template and generation parameters are available in the project GitHub repository to ensure full reproducibility. A sample snippet for the topic 'immigration' is shown in Figure 2.

```
<top>
<num>221</num>
<title>immigration</title>
<publication>
  <desc>Scholarly and policy publications on international
        immigration: migration flows, immigrant integration, asylum
        /refugee issues, labour migration, remittances,...</desc>
  <narr>Include empirical studies, reviews, policy analyses and
        theoretical works that address international migration,
        immigrant integration and assimilation, labor market
        outcomes of immigrants, refugee/asylum issues,...</narr>
</publication>
<research_data>
  <desc>Datasets relevant to immigration research: population
        registers, censuses and microdata with country-of-birth/
        citizenship, labour force surveys, asylum and...</desc>
  <narr>Include population registers, immigration and
        naturalization administrative data, household and labor
        force surveys with migrant identifiers (country of birth,
        citizenship, year of arrival),...</narr>
</research_data>
<variables>
  <desc>Variable metadata for migration concepts: country/region
        of birth, citizenship, legal status, reason for migration,
        year/age at arrival, duration, language proficiency,...</
        desc>
  <narr>Include variables such as migrant status (immigrant,
        emigrant, refugee, asylum-seeker), country of birth,
        citizenship, parental country of birth,...</narr>
</variables>
<instruments_tools>
  <desc>Survey modules, question batteries, scales and tools
        designed to measure migration status, integration,
        acculturation, legal status, and migration...</desc>
  <narr>Include survey modules and question batteries for
        migration (e.g., migrant background question sets used in
        EU surveys), language proficiency assessments, integration
        scales (social, economic integration indices),...</narr>
</instruments_tools>
</top>
```

Figure 2: A topic ('immigration') description from the MIRA dataset.

4.3 Relevance Judgments

A notable feature of the MIRA dataset is its scale and the methodology used for its relevance judgments. For each of the 200 selected topics, we identified a pool of candidate documents to be judged. This pool consists of documents from all four categories that received a `view record`, `download` or `export` user interaction after the query was issued, providing a strong implicit signal of potential relevance (as reported in [26]) and a tractable set for annotation. To reduce potential exposure bias and improve pool completeness, we further augmented the candidate set using BM25 and ColBERT with a supervised expanded form of the query, thereby incorporating retrieval-based candidates in addition to interaction-derived documents. A detailed discussion of the pooling techniques employed is provided in the project GitHub repository.

Relevance judgments were then automatically generated using an LLM on the selected interactions. For our study, we employed OpenAI's `gpt-5-mini`, which was also used to generate descriptions and narratives for the topics. We developed a structured prompting strategy that provided the LLM with the topic description and the document metadata, instructing it to assess their relevance. Judgments were made on a graded relevance scale from 0 to 4, defined as follows: '0' (Not Relevant), '1' (Marginally Relevant), '2' (Fairly Relevant), '3' (Highly Relevant), and '4' (Perfectly Relevant). Here, we opted against a Likert scale [39], as it measures subjective perception rather than objective topical alignment, and the $0-4$ scale is a TREC-adopted standard that integrates directly with IR evaluation metrics like nDCG [28]. Following this strategy, we finally obtain a pool of 85,158 LLM-annotated relevance judgments from four distinct categories for 200 topics. The average number of relevant documents per category is 114.4, 121.94, 48.44, and 14.84 for `publications`, `research data`, `variables`, and `instruments & tools`, respectively.

To validate the LLM-generated judgments, four human experts independently annotated all pooled documents from twenty randomly selected topics (10% of the total), using the same graded relevance scale (0–4). Agreement between the human and the LLM judgments was measured using quadratic-weighted Cohen's κ [14], yielding $\kappa = 0.86$, which indicates substantial agreement [44]. In cases of disagreements, differences were predominantly within a single relevance level (i.e., $|score_{LLM}(Q, D) - score_{Human}(Q, D)| =$

1). These discrepancies typically occurred in borderline or subjective cases where either assessment was defensible, suggesting that the LLM's judgments fall within an acceptable range of human interpretation.

Unlike traditional TREC-style evaluations, our proposed dataset is designed to support category-aware search intents, making the document category a key evaluation criterion. Therefore, each judgment in the MIRA dataset is enriched with a 'category' field, specifying the resource type of the judged document. This unique feature enables not only overall performance analysis but also a detailed examination of retrieval effectiveness by category, which is important for understanding an integrated system's strengths and weaknesses across different types of resources.

5 Benchmark Analysis and Validation

5.1 Baselines

To establish a comprehensive performance baseline for the MIRA dataset, we evaluated a range of retrieval models spanning key IR paradigms. Our evaluation encompasses a range of techniques, from traditional term-based matching to advanced neural architectures. Specifically, we employed (i) the classic BM25 [54, 55] as a strong lexical baseline, (ii) the Relevance-based Language Model, RLM [27, 37] to represent traditional pseudo-relevance feedback and query expansion, (iii) the late-interaction neural model, ColBERT [33] for efficient contextualized retrieval, and (iv) the SOTA model inspired by sequence-to-sequence transformer, MonoT5 [45].

We evaluate models in a category-aware setting, assessing effectiveness within each document category to analyze performance differences across heterogeneous resource types and identify category-specific strengths and weaknesses. We intentionally exclude LLM-based retrievers from our baseline comparison, as the significant inference latency and resource costs associated with these models [38, 42, 43] preclude their deployment in real-time search scenarios, including the operational context of the GESIS Search platform. The hyperparameter settings for baseline models, as well as topic modeling, are available in the GitHub repository to support reproducibility.

5.2 Evaluation

We report four complementary metrics: (i) P@10 captures early precision, reflecting user expectations for first-page results in real-time systems, (ii) nDCG@10 accommodates our graded relevance scale by rewarding highly relevant documents, i.e. more than marginally relevant ones [28], (iii) MAP serves as our primary overall effectiveness measure across the full ranked list, and (iv) GMAP addresses query-level performance variability; unlike arithmetic mean, it is sensitive to low-performing queries and penalizes catastrophic failures on specific information needs [53]. Given the number of topics and multiple pairwise comparisons across categories and metrics, we apply a conservative significance threshold of $p < 0.0001$ with paired t-tests to reduce the likelihood of Type I errors.

The results are summarized in Table 1. The variation in performance across categories showcases the uniqueness of our integrated, category-sensitive search task, where each category provides a separate search space for the same user query. This means that for a particular user query, retrieval outcomes are likely to uncover multiple facets of the query depending on its search intent and the specific category, and hence our proposed benchmark provides a multi-faceted evaluation paradigm. The results in Table 1 clearly show that `Publications` exhibit the highest retrieval effectiveness in all four evaluation metrics.

Table 1 presents a clear, statistically validated performance hierarchy that holds consistently across all categories and metrics. The classic bag-of-words based BM25 provides a strong lexical baseline, while RLM shows significant improvements over BM25 in several cases (marked as † in Table 1), particularly for `Publications` and selected `Research Data` categories. Crucially, ColBERT and MonoT5 both achieve significant advancements over RLM across nearly all metrics and categories ('*'-ed in Table 1), highlighting the efficacy of neural approaches. While ColBERT – employing late interaction – demonstrates slightly more consistent superiority, particularly across all metrics in `Publications` and `Instruments & Tools`, MonoT5 leads in `Research Data` and partially in `Variables`, attaining peak scores in this category (nDCG@10: 0.5736, MAP: 0.4337, and GMAP: 0.3297). These results confirm that state-of-the-art neural models effectively capture global query–document relevance across diverse collection types in the MIRA dataset.

These results demonstrate two key points. *First*, the MIRA dataset can discriminate between retrieval models of varying sophistication, revealing statistically significant performance gaps that align with expected trends in the field. *Second*, the task of multi-categorical retrieval in a scholarly domain remains challenging, with the highest MAP score below 0.6.[1]

MIRA, thus, offers a rich basis for developing and evaluating more advanced integrated and category-aware search systems. The observed performance differences across categories highlight the importance of category-aware modeling. Differences in document length, metadata richness, and vocabulary distribution suggest that ranking functions optimized for one category may not transfer directly to others. MIRA therefore provides a controlled setting for studying category-sensitive retrieval strategies.

6 Discussion

We now contextualize the MIRA benchmark by addressing three key aspects: (i) a methodological limitation regarding implicit relevance signals and their impact on our assessment approach, (ii) a formal distinction between our dataset and classic TREC ad-hoc collections, and (iii) the novel applications and research tasks enabled by this resource.

Limitations of Implicit Relevance Signals. As reported in [26], user interactions, such as `view record`, `download`, or `export`, are valuable implicit signals for inferring relevance. However, it is important to note that these signals may not be definitive. Our analysis confirms that these interactions do not always correlate perfectly with topical relevance. For instance, a user may view a record to quickly ascertain its irrelevance, or download an item for administrative purposes rather than for its topical content. This potential for noise and misinterpretation motivated our decision

[1] Note that, this performance ceiling is notably higher than on traditional TREC ad-hoc collections, where MAP scores typically plateau between 0.25 – 0.40 [33, 45, 49]. We attribute this to the rich, structured metadata in scholarly objects and the presence of explicit categorical signals in real user queries. Crucially, however, substantial headroom remains for cross-category and category-aware retrieval challenges.

Table 1: Performance comparison of both statistical and neural retrieval models on the MIRA dataset. Results are reported using P@10, nDCG@10, MAP, and GMAP. The best performance across the metric-category pair is highlighted in bold. Both neural models are significantly better (paired t-test with $p < 0.0001$) than BM25. Superscript † indicates a significant difference between BM25 and RLM. Significant improvements over RLM by the two neural models are highlighted with '*'.

	Publications				Research Data				Variables				Instruments & Tools			
	P@10	nDCG	MAP	GMAP	P@10	nDCG	MAP	GMAP	P@10	nDCG	MAP	GMAP	P@10	nDCG	MAP	GMAP
BM25	0.6120	0.6091	0.5098	0.5260	0.5045	0.5321	0.4023	0.3058	0.4905	0.4408	0.5057	0.1648	0.4190	0.4711	0.4540	0.2152
RLM	0.6242†	0.6213†	0.5200†	0.5365†	0.5146	0.5427†	0.4103	0.3119†	0.5003	0.4496	0.5158†	0.1681	0.4274	0.4805	0.4631	0.2195
ColBERT	**0.6523***	**0.6492***	**0.5434***	**0.5607***	**0.5377***	0.5639*	0.4263*	0.3241*	0.5198*	0.4672*	**0.5354***	**0.1745***	**0.4436***	**0.4988***	**0.4807***	**0.2243***
MonoT5	0.6365*	0.6335	0.5302*	0.5470*	0.5247*	**0.5736***	**0.4337***	**0.3297***	**0.5288***	**0.4752**	0.5249*	0.1711*	0.4349*	0.4890	0.4713*	0.2234*

Table 2: Comparisons between Classic TREC Ad-hoc datasets and the MIRA dataset.

Features	TREC Ad-Hoc	MIRA
Document Homogeneity	Homogeneous (e.g., newswire articles, web pages).	Intentionally heterogeneous (`Research Data`, `Variables`, `Publications`, `Instruments & Tools`).
Relevance Criteria	Uniform across the collection (textual aboutness).	Category-dependent (e.g., reusability for data, measurability for variables, methodological guidance for tools).
Topic Origin	Expert-created, simulated.	Real user queries from a live search platform.
Judgment Scope	Focused on a single, primary document type.	Encompasses multiple, distinct document types for a single topic.
Core Evaluation Goal	Ranking effectiveness within a single corpus.	Integrated ranking effectiveness across multiple corpora.

to rely on explicit, LLM-assisted human assessments to establish the ground-truth relevance judgments in the MIRA benchmark, thereby ensuring greater accuracy.

Distinction from Classic TREC Ad-Hoc Datasets. The proposed MIRA dataset represents a significant evolution from classic TREC ad-hoc test collections, which were largely designed for evaluating retrieval over a uni-categorical document corpora. In contrast, MIRA items are fundamentally multi-categorical, integrating four distinct scholarly resource types – `Publications`, `Research Data`, `Variables`, and `Instruments & Tools` – within a single benchmark, thereby reflecting the reality of modern, integrated search platforms. Unlike traditional benchmark collections (e.g., TREC) that rely mostly on manually curated evaluation topics, MIRA derives its topics from real user interactions. In addition, while TREC depends on fully human-generated relevance judgments, MIRA employs a scalable LLM-assisted framework to produce graded relevance assessments across heterogeneous resource categories. The key differences are summarized in Table 2.

Application of the MIRA dataset. This new resource opens up numerous avenues for future research. Primarily, it serves as a benchmark for evaluating integrated, category-aware, and cross-category retrieval systems, aiming to produce a single, effective ranking across heterogeneous item types. It can also be used to study category-specific retrieval models, investigating, for instance, whether a specialized model for `Variables` outperforms a generic one. Furthermore, it enables researchers to explore additional tasks, such as *result diversification* (ensuring that top results span multiple relevant categories), *fairness in ranking* (analyzing whether certain categories are systematically under-represented), and *query performance prediction* across different corpora. Finally, its foundation on a dual-lingual platform invites research on code-switching and cross-lingual retrieval.

The MIRA dataset enables the integration of offline and online evaluation in IR. Continuous evaluation [32] assesses and improves search systems in dynamic environments where users, topics, and documents evolve. Platforms like GESIS Search exemplify this need. With its automated creation workflow, MIRA can be continuously expanded with new topics, documents, and user interactions, forming a living evaluation corpus that stays aligned with system and user evolution. By automating dataset generation from user logs and metadata enrichment, our approach enables continuous evaluation datasets to be produced for any domain or search system.

7 Conclusion and Future Work

The modern search experience is integrated, yet IR benchmarks have lagged behind, constrained by a lack of collections that mirror this reality. We introduce MIRA, a novel benchmark designed to address the critical evaluation gap in multi-categorical information retrieval. The MIRA dataset directly confronts this challenge by providing a unified framework encompassing four distinct scholarly categories – `Publications`, `Research Data`, `Variables`, and `Instruments & Tools` – all grounded in real user queries from the GESIS Search platform. Our primary contribution is a resource that is both realistic in origin and novel in its construction, using LLMs to efficiently produce topic narratives and scalable, graded relevance judgments spanning all categories.

The MIRA dataset provides a comprehensive benchmark for evaluating integrated, category-aware and cross-category retrieval systems, spanning over multiple heterogeneous resource types. It supports research on category-aware ranking models and enables analysis of performance differences across heterogeneous resource types, along with investigations into result diversification, fairness-aware ranking, and query performance prediction. Notably, its automatic generation process lays the groundwork for a continuous evaluation paradigm, in which evolving systems and their user base are captured within a living evaluation corpus.

Looking ahead, we plan to expand the MIRA dataset by increasing the number of judged topics and increasing the depth of the pool of relevance assessments, potentially through continued LLM-assisted methods complemented by human-in-the-loop validation.

We also envision organizing a shared task or benchmark challenge to foster community engagement and establish baseline performance metrics. We believe MIRA provides a foundational step towards evaluating the next generation of search systems and will inspire further work in complex, multi-faceted, and federated information retrieval.

Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 407518790, and partly funded by Horizon Europe grant OMINO (grant number 101086321). Also, this publication received partial support from the European Research Council (ERC) under the Horizon 2020 research and innovation program (Grant agreement No. 884951), and Research Ireland to the Insight Centre for Data Analytics under grant no. 12/RC/2289_P2.

References

[1] [n. d.]. CLEF. https://www.clef-initiative.eu/. Accessed: 2026-02-12.
[2] [n. d.]. FIRE. https://fire.irsi.org.in/fire/2025/home. Accessed: 2026-02-12.
[3] [n. d.]. GESIS Search. https://search.gesis.org/. Accessed: 2026-02-12.
[4] [n. d.]. NTCIR. https://research.nii.ac.jp/ntcir/index-en.html. Accessed: 2026-02-12.
[5] [n. d.]. TREC. https://trec.nist.gov/. Accessed: 2026-02-12.
[6] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Inf. Process. Manage.* 48, 6 (Nov. 2012), 1053–1066. https://doi.org/10.1016/j.ipm.2012.01.004
[7] Negar Arabzadeh and Charles L. A. Clarke. 2025. Benchmarking LLM-based Relevance Judgment Methods. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 3194–3204. https://doi.org/10.1145/3726302.3730305
[8] Debanjali Biswas, Endri Gupta, Ran Yu, and Benjamin Zapilko. 2025. GESIS Knowledge Graph (GESIS KG). (Version 2.0.0) [Data set]. GESIS, Cologne. https://doi.org/10.7802/2969
[9] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Unsupervised Dataset Generation for Information Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2387–2392. https://doi.org/10.1145/3477495.3531863
[10] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2013), Part II (Lecture Notes in Computer Science, Vol. 7819)*, Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu (Eds.). Springer, 160–172. https://doi.org/10.1007/978-3-642-37456-2_14
[11] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2019. Dataset search: a survey. *The VLDB Journal* 29, 1 (Aug. 2019), 251–272. https://doi.org/10.1007/s00778-019-00564-x
[12] J Chen, N Wang, C Li, B Wang, S Xiao, H Xiao, H Liao, D Lian, and Z Liu. [n. d.]. Air-bench: Automated heterogeneous information retrieval benchmark (2024).
[13] Cyril Cleverdon. 1997. *The Cranfield tests on index language devices.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 47–59.
[14] Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70, 4 (1968), 213–220. https://doi.org/10.1037/h0026256
[15] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. MS MARCO: Benchmarking Ranking Models in the Large-Data Regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1566–1576. https://doi.org/10.1145/3404835.3462804
[16] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Ellen M. Voorhees, and Ian Soboroff. 2021. TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2369–2375. https://doi.org/10.1145/3404835.3463249
[17] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC Complex Answer Retrieval Overview. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017 (NIST Special Publication, Vol. 500-324)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).
[18] Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. 2025. MMDocIR: Benchmarking Multimodal Retrieval for Long Documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 30959–30993.
[19] Alessandro Epasto, Jon Feldman, Silvio Lattanzi, Stefano Leonardi, and Vahab Mirrokni. 2014. Reduce and aggregate: similarity ranking in multi-categorical bipartite graphs. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14)*. Association for Computing Machinery, New York, NY, USA, 349–360. https://doi.org/10.1145/2566486.2568025
[20] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval* (Taipei, Taiwan) *(ICTIR '23)*. Association for Computing Machinery, New York, NY, USA, 39–50. https://doi.org/10.1145/3578337.3605136
[21] Naghmeh Farzi and Laura Dietz. 2025. Criteria-Based LLM Relevance Judgments. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)* (Padua, Italy) *(ICTIR '25)*. Association for Computing Machinery, New York, NY, USA, 254–263.
[22] Yaming Fu, Elizabeth Lomas, and Charles Inskip. 2021. Library log analysis and its implications for studying online information seeking behavior of cultural groups. *The Journal of Academic Librarianship* 47, 5 (Sept. 2021), 102421. https://doi.org/10.1016/j.acalib.2021.102421
[23] Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. LLM-based NLG Evaluation: Current Status and Challenges. *Computational Linguistics* 51, 2 (06 2025), 661–687. https://doi.org/10.1162/coli_a_00561 arXiv:https://direct.mit.edu/coli/article-pdf/51/2/661/2513169/coli_a_00561.pdf
[24] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 [cs.CL] https://arxiv.org/abs/2203.05794
[25] Daniel Hienert. 2017. User interests in German social science literature search: a large scale log analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 7–16.
[26] Daniel Hienert, Dagmar Kern, Katarina Boland, Benjamin Zapilko, and Peter Mutschke. 2019. A digital library for research data and related information in the social sciences. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 148–157.
[27] Nasreen Abdul Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference (NIST Special Publication, Vol. 500-261)*. National Institute of Standards and Technology (NIST).
[28] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418
[29] Steve Jones, Sally Jo Cunningham, Rodger McNab, and Stefan Boddie. 2000. A transaction log analysis of a digital library. *International Journal on Digital Libraries* 3, 2 (Aug. 2000), 152–169. https://doi.org/10.1007/s007999900022
[30] Seikyung Jung, Jonathan L. Herlocker, and Janet Webster. 2007. Click data as implicit relevance feedback in web search. *Information Processing & Management* 43, 3 (May 2007), 791–807. https://doi.org/10.1016/j.ipm.2006.07.021
[31] Emilia Kacprzak, Laura M. Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. 2017. A Query Log Analysis of Dataset Search. In *Web Engineering*, Jordi Cabot, Roberto De Virgilio, and Riccardo Torlone (Eds.). Springer International Publishing, Cham, 429–436.
[32] Jüri Keller. 2025. Continuous Evaluation in Information Retrieval Across Methods and Time. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 4205. https://doi.org/10.1145/3726302.3730124
[33] Omar Khattab and Matei Zaharia. 2020. *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT*. Association for Computing Machinery, New York, NY, USA, 39–48.
[34] Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2023. COLIEE 2022 Summary: Methods for Legal Document Retrieval and Entailment. In *New Frontiers in Artificial Intelligence*, Yasufumi Takama, Katsutoshi Yada, Ken Satoh, and Sachiyo Arai (Eds.). Springer Nature Switzerland, Cham, 51–67.
[35] Laura M. Koesten, Emilia Kacprzak, Jenifer F. A. Tennison, and Elena Simperl. 2017. The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1277–1289. https://doi.org/10.1145/3025453.3025838

[36] Nikolay Kolyada, Martin Potthast, and Benno Stein. 2025. A Test Collection for Dataset Retrieval. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part III* (Lucca, Italy). Springer-Verlag, Berlin, Heidelberg, 372–380. https://doi.org/10.1007/978-3-031-88714-7_36

[37] Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA) *(SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 120–127.

[38] Guoyao Li, Ran He, Shusen Jing, Kayhan Behdin, Yubo Wang, Sundara Raman Ramachandran, Chanh Nguyen, Jian Sheng, Xiaojing Ma, Chuanrui Zhu, Sriram Vasudevan, Muchen Wu, Sayan Ghosh, Lin Su, Qingquan Song, Xiaoqing Wang, Zhipeng Wang, Qing Lan, Yanning Chen, Jingwei Wu, Luke Simon, Wenjing Zhang, Qi Guo, and Fedor Borisyuk. 2026. MixLM: High-Throughput and Effective LLM Ranking via Text-Embedding Mix-Interaction. arXiv:2512.07846 [cs.IR] https://arxiv.org/abs/2512.07846

[39] Rensis Likert. 1932. A Technique for the Measurement of Attitudes. *Archives of Psychology* 140 (1932), 1–55.

[40] Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. arXiv:2304.07666 [cs.CL]

[41] Mihai Lupu, Jiashu Zhao, Jimmy X. Huang, Harsha Gurulingappa, Juliane Fluck, Marc Zimmermann, Igor V. Filippov, and John Tait. 2011. Overview of the TREC 2011 Chemical IR Track. In *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011 (NIST Special Publication, Vol. 500-296)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec20/papers/CHEM.OVERVIEW.pdf

[42] Guangyuan Ma, Yongliang Ma, Xuanrui Gou, Zhenpeng Su, Ming Zhou, and Songlin Hu. 2025. LightRetriever: A LLM-based Text Retrieval Architecture with Extremely Faster Query Inference. *CoRR* abs/2505.12260 (2025). https://doi.org/10.48550/ARXIV.2505.12260 arXiv:2505.12260

[43] Xueguang Ma, Xi Victoria Lin, Barlas Oguz, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2025. DRAMA: Diverse Augmentation from Large Language Models to Smaller Dense Retrievers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 30170–30186. https://doi.org/10.18653/v1/2025.acl-long.1457

[44] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK. http://nlp.stanford.edu/IR-book/information-retrieval-book.html

[45] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy J. Lin. 2019. Multi-Stage Document Ranking with BERT. *ArXiv* abs/1910.14424 (2019).

[46] María-Dolores Olvera-Lobo and Juncal Gutiérrez-Artacho. 2015. *Question Answering Track Evaluation in TREC, CLEF and NTCIR*. Springer International Publishing, 13–22. https://doi.org/10.1007/978-3-319-16486-1_2

[47] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006 (ACM International Conference Proceeding Series, Vol. 152)*, Xiaohua Jia (Ed.). ACM, 1. https://doi.org/10.1145/1146847.1146848

[48] Carol Peters, Giorgio Maria Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Penas, and Giovanna Roda (Eds.). 2010. *Multilingual information access evaluation I - text retrieval experiments* (2010 ed.). Springer, Berlin, Germany.

[49] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. *CoRR* abs/2101.05667 (2021).

[50] 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. 2024. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1504–1518. https://doi.org/10.18653/v1/2024.findings-naacl.97

[51] Hossein A. Rahmani, Xi Wang, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Paul Thomas. 2025. SynDL: A Large-Scale Synthetic Test Collection for Passage Retrieval. In *Companion Proceedings of the ACM on Web Conference 2025* (Sydney NSW, Australia) *(WWW '25)*. Association for Computing Machinery, New York, NY, USA, 781–784. https://doi.org/10.1145/3701716.3715311

[52] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 3982–3992.

[53] Stephen Robertson. 2006. On GMAP: and other transformations. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* (Arlington, Virginia, USA) *(CIKM '06)*. Association for Computing Machinery, New York, NY, USA, 78–83. https://doi.org/10.1145/1183614.1183630

[54] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (apr 2009), 333–389.

[55] S. E. Robertson and S. Walker. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR '94*. Springer London, London, 232–241.

[56] Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL Corpus: Advancing Open Science in Computational Linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10348–10361. https://doi.org/10.18653/v1/2023.emnlp-main.640

[57] Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375. https://doi.org/10.1561/1500000009

[58] Tefko Saracevic. 2008. Effects of Inconsistent Relevance Judgments on Information Retrieval Test Results: A Historical Perspective. *Library Trends* 56, 4 (March 2008), 763–783. https://doi.org/10.1353/lib.0.0000

[59] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. 2017. A Test Collection for Evaluating Retrieval of Studies for Inclusion in Systematic Reviews. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 1237–1240. https://doi.org/10.1145/3077136.3080707

[60] Yang Sun, Huajing Li, Isaac G. Councill, Jian Huang, Wang-Chien Lee, and C. Lee Giles. 2008. Personalized ranking for digital libraries based on log analysis. In *Proceedings of the 10th ACM Workshop on Web Information and Data Management* (Napa Valley, California, USA) *(WIDM '08)*. Association for Computing Machinery, New York, NY, USA, 133–140. https://doi.org/10.1145/1458502.1458524

[61] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[62] Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press.

[63] Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024. SOTOPIA-π: Interactive Learning of Socially Intelligent Language Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12912–12940. https://doi.org/10.18653/v1/2024.acl-long.698

[64] Yu Xia, Junda Wu, Sungchul Kim, Tong Yu, Ryan A. Rossi, Haoliang Wang, and Julian McAuley. 2025. Knowledge-Aware Query Expansion with Large Language Models for Textual and Relational Retrieval. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 4275–4286. https://doi.org/10.18653/v1/2025.naacl-long.216

[65] Junte Zhang and Jaap Kamps. 2010. *A Search Log-Based Approach to Evaluation*. Springer Berlin Heidelberg, 248–260. https://doi.org/10.1007/978-3-642-15464-5_26

[66] Dengya Zhu, Shastri L. Nimmagadda, Kok Wai Wong, and Torsten Reiners. 2023. *Relevance Judgment Convergence Degree—A Measure of Assessors Inconsistency for Information Retrieval Datasets*. Springer International Publishing, Cham, 149–168. https://doi.org/10.1007/978-3-031-32418-5_9

[67] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2025. Large Language Models for Information Retrieval: A Survey. *ACM Trans. Inf. Syst.* (Sept. 2025). https://doi.org/10.1145/3748304

[68] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 307–314. https://doi.org/10.1145/290941.291014