

Does LLM Relevance Labelling Work for Arabic?

Marwah Alaofi
Taibah University
Medina, Saudi Arabia
maofi@taibahu.edu.sa

Fatima Haouari
The University of Sheffield
Sheffield, United Kingdom
f.haouari@sheffield.ac.uk

Abstract

Large Language Models (LLMs) are increasingly used in Information Retrieval, both within retrieval pipelines and for constructing evaluation resources. Existing studies on using LLMs for IR evaluation, however, focus almost exclusively on English, leaving their applicability to other languages, where evaluation resources are often limited and highly needed, unexplored.

We examine the use of LLMs to generate relevance labels for an Arabic test collection (ArTest). Using about 10K relevance labels, generated by three LLMs, and used to order eight automated systems and nine simulated manual systems, we show that agreement on binary labels and system ordering is generally high, with no cases of significant opposite conclusions; however, there are cases of false or missed improvements and noticeable limitations when labelling manual, highly performing systems. These findings align with results reported for English and indicate that LLMs could, with some caveats, offer a viable approach to supporting IR evaluation in languages with limited evaluation resources.¹

CCS Concepts

• Information systems → Relevance assessment; Test collections.

Keywords

Arabic information retrieval; test collections; relevance judgements; large language models

ACM Reference Format:

Marwah Alaofi and Fatima Haouari. 2026. Does LLM Relevance Labelling Work for Arabic?. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*, July 20–24, 2026, Melbourne, VIC, Australia. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3805712.3809855>

1 Introduction and Background

The evaluation and development of Information Retrieval (IR) systems have long relied on the availability, quality, and relevance of test collections for the retrieval task. Creating test collections is expensive, with relevance judgements being the most labour-intensive and costly component. In recent years, LLMs have been explored as a cost-effective approach to replace human judgements. Prior

¹Relevance labels and systems are available at: <https://github.com/MarwahAlaofi/ArTest-qrels>

studies have demonstrated their promise; however, this has been investigated predominantly in English-language settings [1, 4, 11, 27]. In other languages, where IR evaluation resources are limited, using LLMs to construct test collections has received little attention.

For Arabic, IR test collections are rare, often outdated, and subject to a range of limitations. To the best of our knowledge, Arabic IR evaluation has largely relied on two early resources developed under the TREC Cross-Language tracks of 2001 and 2002 [29, 30], comprising 25 and 50 topics, respectively. These collections were constructed for cross-lingual retrieval where queries and documents are in two different languages, with topics developed in English and subsequently translated into Arabic to enable performance comparison with a monolingual setting. Therefore, the resulting queries do not adequately reflect the information needs of the Arabic-speaking population. Also, the document collections are limited to the news domain (using Arabic Newswire Part 1). Beyond these early efforts, only one Arabic web search test collection, ArTest, has been developed [15].

Recent work on using LLMs for relevance labelling generally concludes that LLM-generated labels – at least those produced by competitive LLMs – agree with human judgements to a high degree, often comparable to human-to-human agreement, and/or that retrieval system rankings remain largely stable when evaluated using LLM labels [1, 11, 26, 27]. The community has, however, warned against LLM evaluation ‘tropes’ [10]. Recent work has raised concerns about circularity when LLMs are used for both ranking and evaluation [6, 24], bias in favour of LLM-based re-rankers [5], recency bias [12], and systematic disagreements with human judgements that are concentrated in specific semantic clusters – often for definition-seeking, policy-related, or ambiguous contexts [22]. Additional concerns include failures under adversarial conditions [3], and missing significant or overestimating insignificant improvements of retrieval systems [4, 23].

While research on the use of LLMs for relevance judgements in English – as described above – has moved beyond establishing their potential and now focuses on identifying limitations, the more fundamental question of whether LLMs can achieve a high level of agreement with human judgements remains underexplored in other languages. Exception is a study on a non-English collection with queries translated into Chinese, Russian, and Persian [26], which found that the LLM performs well across languages despite variations in the language of the prompt and information need description.

This work investigates the similarity between LLM relevance labels and human relevance judgements using intrinsic measures of per-label agreement, and extrinsic measures assessing the impact of these labels on system ordering (from best to worst in terms of performance) for an Arabic test collection. Specifically, we address the following research questions:



This work is licensed under a Creative Commons Attribution 4.0 International License. *SIGIR '26, Melbourne, VIC, Australia*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2599-9/2026/07
<https://doi.org/10.1145/3805712.3809855>

Topics	Variants			Judgements (%)			
	Min	Max	Avg	-1	0	1	2
50	1	7	3.42	5.2	69.4	12.1	13.3

Table 1: The number of topics, the minimum, maximum and average number of query variants per topic, and the distribution of relevance judgements in the the ArTest collection (over 10,529 judgements).

RQ1. To what extent do LLM- and human-generated relevance judgements agree?

RQ2. How similar are LLMs and humans in their ordering of retrieval systems – including both *automated* and *manual*?

2 Experimental Design

This section describes the test collection used to examine the use of LLMs for relevance labelling, the LLMs and their configurations, and the evaluation methods.

2.1 Test Collection

We used the ArTest test collection [15], which is built on the ArabicWeb16 corpus [25]. To the best of our knowledge, ArTest is the most recent Arabic IR test collection, constructed from original Arabic queries sourced directly from Arabic speakers, and the only one designed for the web search task. This distinguishes it from prior collections that rely on queries adapted to Arabic and with limited scope, such as news articles [29, 30], or that designed for Arabic microblog search [2].

The collection consists of 50 topics, each associated with a description, a narrative, and multiple query variants that reflect alternative formulations of the same information need. These variants were created and used by the judges to construct the document pools for relevance judgement, following the premise that query variants can contribute to pool diversity in a manner comparable to system runs [21]. Relevance judgements are provided on a four-point graded scale, from -1 to 2 , where -1 denotes empty or spam, 0 not relevant, 1 relevant, and 2 highly relevant. Table 1 provides key statistics of the collection.

2.2 Relevance Labelling

LLM Selection and Configuration. To enable comparison, we used GPT-4o, which has been used in prior studies and shown to achieve high agreement with human judges in relevance judgement for English collections [3, 28] and also for Russian, Chinese and Persian collections [26]. We also included two open-source LLMs – Llama-3.3-70B-Instruct² and GPT-oss-120B³ – to support reproducibility. We used the same parameter setup as Thomas et al. [27].

Document Pre-processing. Most prior work on using LLMs for relevance labelling has been conducted on passage-level test collections – mainly using the TREC deep learning tracks, which

use the MSMARCO corpus of passages. Studies that consider full HTML web pages do not report their pre-processing steps. We therefore experimented with two settings: providing the LLM with raw HTML web pages as-is, and providing cleaned HTML after removing scripts and non-content tags. While we initially expected that supplying full HTML would enable the LLM to make better relevance labelling by exploiting structural cues encoded in markup, this was not observed in practice. Instead, we obtained higher agreement with human judgements when using cleaned HTML web pages, and therefore adopted this setting in our experiment.

Prompt. We used the DNA prompt proposed in prior work [27] including information need **Description**, **Narrative** and **Aspect**. We provided the prompt in English, following findings that show that changing the prompt language when labelling non-English documents does not seem to affect relevance labelling results [26].

2.3 Relevance Label Evaluation

To assess the alignment between LLM-generated labels and human relevance judgements, we used standard intrinsic agreement metrics, including Cohen’s κ [7] and Krippendorff’s α [18]. These measures quantify the degree of agreement at the label level. We further evaluated agreement at ordering retrieval systems (i.e., runs) using Kendall’s τ [17] and conducted pairwise comparisons across systems.

For each system pair (S_1, S_2) , outcomes were categorised as **Active Agreement (AA)** when both LLM and human judgements agree on the direction of the comparison ($S_1 > S_2$ or $S_2 > S_1$) and both identify a statistically significant difference; **Passive Agreement (PA)** when both agree on the direction but neither finds significance; and **Mixed Agreement (MA)** when only one identifies significance. **Active**, **Passive**, and **Mixed Disagreement (AD, PD, MD)** are used similarly but for comparisons in which the LLMs and human judgements indicate opposite ranking directions [13, 20].

Both automated retrieval systems and manually constructed runs were included in this analysis to examine whether different types of systems are treated differently under LLM-based evaluation. That is, manually constructed runs may exhibit characteristics that differ substantially from automated systems, and thus may not be treated equally by automated labelling. Evaluation of system performance was conducted using the official TREC run_eval script.⁴

Given the judgement process used to construct ArTest – where judges developed topic descriptions and narratives, and were encouraged to use query variants to identify relevant documents and judge at least 200 documents without a fixed inspection depth – our automated runs contained a substantial number of missing judgements, with a minimum of 30% missing judgements per topic for BM25 (the system used by the judges). Accordingly, we restricted our corpus to documents that have relevance judgements. This decision is supported by prior findings showing that sub-sampling the top-k documents included in relevance judgements can provide a reliable estimate of system effectiveness [14]. The following describes how the two types of runs were generated:

Automated Runs. We indexed the HTML web pages – including their titles and main content – via Pyserini [19] with Arabic

²<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>

³<https://huggingface.co/openai/gpt-oss-120b>

⁴https://github.com/usnistgov/trec_eval

language processing enabled. We generated eight different runs classified as follows:

- **Lexical Models:** Using both BM25 and BM25 with RM3 pseudo-relevance feedback over topic titles. BM25 was configured with $k_1 = 0.9$ and $b = 0.4$, and RM3 used 10 feedback documents and 10 terms with an original topic title weight of 0.5.
- **Dense Models:** Web pages and topic titles were encoded using the multilingual E5 encoders (base and large) [31] with FAISS indexing in Pyserini. Web pages were ranked based on the inner-product similarity.
- **Dense Re-rankers:** Similar to the aforementioned dense models, except that ranking was applied to the top 100 documents retrieved by the lexical models (BM25 and BM25+RM3).

Manual Runs. Inspired by the synthetic oracle runs created by Balog et al. [5] to examine potential differences in how LLM labellers and human judges evaluate synthetic oracle runs relative to LLM-based re-rankers, we constructed a set of *simulated* manual runs that approximate rankings produced by human judges under different levels of *effort* and *patience*. Effort is modelled through the extent to which query variants are used to locate relevant documents, while patience is modelled by the depth to which judges inspect ranked documents. We considered three inspection depths, $d \in \{30, 20, 10\}$, corresponding to decreasing levels of patience.

For each inspection depth d , the document pool was generated using BM25 across all available query variants (i.e., maximum effort), mirroring the judgement process used in the ArTest. We also considered settings in which we minimise effort by leaving out one or two query variants, i.e., leave out $l \in \{0, 1, 2\}$. Given that the average number of query variants per topic is approximately three, excluding two variants represents the most restrictive setting. We note that Topic 4 has only a single variant; therefore, no variants were excluded.

The resulting runs were constructed by ordering the pooled documents according to their human relevance judgements in decreasing order. We refer to these runs as *manual runs*, parameterised by the inspection depth d and the number of left out variants l . These manual runs outperform the automated runs under human judgement.

Finally, we note that Topics 2 and 5 have no query variants and were therefore excluded from both the manual and automated runs for fair comparison.

3 Results and Discussion

This section presents and discusses the results in response to the two research questions.

3.1 Per-Label Agreement (RQ1)

Table 2 reports the agreement between LLM labels and human relevance judgements, the probability of LLMs for labelling a document as relevant and the precision of the generated binary labels: relevant (1) and non-relevant (0).

The observed agreement, as measured by κ , is high for both GPT models when compared with typical inter-annotator agreement between human judges, which has been reported as 0.52 by Cormack

LLM	κ	α	$P(L = 1)$	$Prec(L = 0)$	$Prec(L = 1)$
GPT-4o	0.58	0.35	0.20	0.88	0.76
GPT-oss-120B	0.58	0.57	0.31	0.92	0.63
Llama-3.3-70B	0.38	0.27	0.58	0.99	0.43

Table 2: Agreement between LLM-generated labels and human relevance judgements, measured using Cohen’s κ and Krippendorff’s α , together with the probability (P) of assigning a document a relevant label (L) and label-wise precision (Prec). All metrics were computed using binary labels, where -1 and 0 were mapped to 0 and 1 and 2 were mapped to 1 , except for Krippendorff’s α , which was computed over graded labels.

et al. [8], 0.41 by Hersh et al. [16], and within the range 0.24–0.58 by Damessie et al. [9].

This finding is consistent with prior work using GPT-4o on English test collections, which reported comparable agreement levels (e.g., 0.42–0.50 on the Deep Learning (DL) tracks from 2019–2023 [28], and 0.50–0.60 across different topic, document, and prompt languages on the TREC NeuCLIR tracks from 2022–2023 [26]). In contrast, Llama exhibits considerably lower agreement, although it still falls within the range reported by Damessie et al. [9].

When considering agreement over graded relevance labels using α , GPT-oss-12B exhibits substantially higher agreement than other models. This suggests that, while models may show similar levels of agreement under binary relevance, they differ in their ability to make finer-grained distinctions. Notably, GPT-oss-12B appears better able to differentiate – or at least to agree with human judges – between spam and non-relevant documents. In contrast, GPT-4o shows a higher tendency to falsely label non-relevant documents as spam. This is shown in the confusion matrices shown in Figure 1.

Overall, the α values for Llama and GPT-4o are low relative to previously reported inter-annotator agreement levels (0.41–0.69 [9]), with GPT-4o performing substantially worse than reported in prior studies conducted on English test collections (e.g., 0.62 [3] on DL 2021 and 2022). However, we cannot determine whether this degraded performance is attributable to language effects, the increased noise introduced by HTML web pages, or differences in the labelling scale. Note that previous studies did not experiment with a scale of relevance that included spam. These factors may have influenced our comparisons; for example, if previous studies included a spam label, they might have resulted in comparable α .

In terms of the LLMs’ probability of labelling a document as relevant (0 or 1), Llama is the most liberal, with a probability of 0.58, compared with approximately 0.25 for human judges (see Table 1). This behaviour is known for LLMs [1, 3, 26]. The GPT models are less liberal and more comparable to human judgements, with probabilities of 0.31 for GPT-oss and 0.20 for GPT-4o. This behaviour is reflected in their precision values: high for non-relevant labels and moderate to low for relevant labels.

3.2 Per-System Ordering Agreement (RQ2)

Figure 2 shows the ordering of systems according to human judgements and different LLM labels, and Table 3 shows Kendall’s τ and the proportions of pairwise system comparisons across automated,

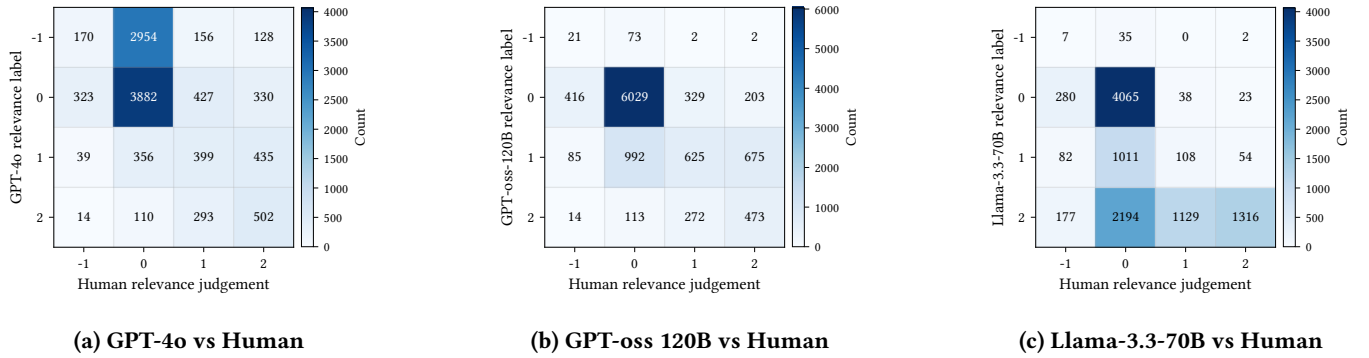


Figure 1: Confusion matrices comparing human relevance judgements with LLM relevance labels.

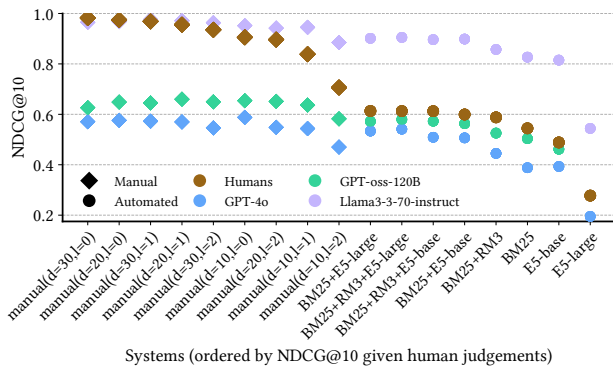


Figure 2: NDCG@10 of retrieval systems evaluated using human relevance judgments and labels generated by the three LLMs.

Table 3: System ordering agreement using Kendall’s τ and the proportions of Active, Passive, and Mixed Agreement (AA, PA, MA) and Disagreement (AD, PD, MD) for pairwise system comparisons give only automated (aut), manual (man) and all systems.

LLM	τ	AA	PA	MA	AD	PD	MD
GPT-4o (aut)	0.86	0.57	0.18	0.14	0.00	0.11	0.00
GPT-4o (man)	0.56	0.25	0.08	0.44	0.00	0.03	0.19
GPT-4o (all)	0.79	0.46	0.06	0.38	0.00	0.05	0.06
GPT-oss-120B (aut)	0.86	0.43	0.29	0.21	0.00	0.71	0.00
GPT-oss-120B (man)	0.06	0.19	0.08	0.25	0.00	0.03	0.44
GPT-oss-120B (all)	0.72	0.60	0.08	0.18	0.00	0.02	0.12
Llama-3.3-70B (aut)	0.86	0.43	0.32	0.18	0.00	0.04	0.04
Llama-3.3-70B (man)	0.67	0.39	0.09	0.19	0.00	0.03	0.14
Llama-3.3-70B (all)	0.82	0.63	0.09	0.19	0.00	0.01	0.07

manual and all system pairs, i.e., 28, 36, and 136 pairs, using an independent t -test.

It is clear from Figure 2 that Llama and GPT models exhibit two different behaviours: scores are overestimated under Llama, particularly for automated systems, and underestimated under the GPT models, most noticeably for manual systems. According to

human judgements, the manual system with the highest inspection depth ($d = 30$) and using all variants ($l = 0$) is the top-ranked system. Under LLM labelling, however, the ordering differs: the top system according to Llama is ranked third by humans, the top system according to GPT-oss is ranked fourth, and the top system according to GPT-4o is ranked sixth. The agreement in ordering the top automated systems, however, is higher: the top-ranked system under human judgements is ranked second by all LLMs, and the top- and second-ranked systems show no statistically significant difference under human judgements; therefore, choosing either system is valid (a case of PD).

Considering pairwise comparisons across all systems (Table 3), Llama achieves the highest agreement with human judgements in system ordering, as measured by Kendall’s τ and the pairwise agreement metrics. Agreement scores are consistently higher when only automated systems are considered, while performance degrades – sometimes severely, as in the case of GPT-oss – when evaluation is restricted to manual systems.

We observe no cases of AD, in which the LLM and human judgements reach opposite ranking conclusions (e.g., $S_1 > S_2$ vs. $S_1 < S_2$) and both identify a statistically significant difference. We have cases of PA and PD, but they are of less concern, as no significant difference is identified and selecting either system is therefore valid. We do, however, observe cases of MA and MD in which significance is identified by either humans or LLMs. In most cases, LLMs fail to detect improvements instead of falsely identifying non-significant differences as significant. In particular, we find that about 92% of MA and MD cases (across all systems) in Llama and GPT-oss occur when humans detect statistically significant differences that LLMs fail to identify; this increases to about 96.6% for GPT-4o.

4 Conclusions and Future Work

Arabic IR test collections are limited. If LLMs can reliably support their construction, they offer a promising and cost-effective means to accelerate resource development and, consequently, advance Arabic IR research.

We found that agreement on binary labels and on system ordering is generally high, aligning with previous results on English test collections. However, agreement on graded relevance appears lower in our evaluation. We note that differences in document format, relevance scales, and collection characteristics may have

contributed to the observed difference. We also found that LLMs perform less reliably when evaluating manually constructed systems. As prior work has not examined using LLMs to judge such systems, it remains unclear whether this limitation is specific to Arabic or reflects a broader issue.

Our use of ‘simulated’ manual runs offers a complementary perspective for evaluating LLM-generated labels and may be adopted in efforts to assess the reliability of LLMs as labellers. As these runs were constructed by sampling documents from the top-k BM25 results retrieved in response to query variants, we expect them to differ from other automated runs where documents were retrieved in response to information need titles alone. We believe that this setup may have resulted in documents that are much harder for LLMs to label.

This study does not aim to identify the best-performing LLM, prompt, or parameter configuration for the task. Rather, it investigates whether LLMs, configured in line with prior literature, can support Arabic IR evaluation at a level comparable to that reported for English. Our findings indicate that this is largely the case, as we do not observe sufficiently strong evidence of differences.

Future work will investigate the conditions under which LLM and human judgements differ, particularly for documents of manual systems.

Acknowledgments

The authors thank Falk Scholer, Mark Sanderson, and Paul Thomas for their time and insightful discussions related to this work.

References

- [1] Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. 2024. Can We Use Large Language Models to Fill Relevance Judgment Holes? arXiv:2405.05600 [cs.LG]. <https://arxiv.org/abs/2405.05600>
- [2] Hind A. Al-Merkehi, Maram Hasanain, and Tamer Elsayed. 2016. EveTAR: A New Test Collection for Event Detection in Arabic Tweets. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 689–692. doi:10.1145/2911451.2914681
- [3] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2024. LLMs can be Fooled into Labelling a Document as Relevant: best café near me; this paper is perfectly relevant. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2024, Tokyo, Japan, December 9-12, 2024*, Tetsuya Sakai, Emi Ishita, Hiroaki Ohshima, Faegheh Hasibi, Jiaxin Mao, and Joemon M. Jose (Eds.). ACM, 32–41. doi:10.1145/3673791.3698431
- [4] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2026. On the Use of LLMs for Relevance Labelling. *ACM Trans. Inf. Syst.* 44, 4, Article 77 (April 2026), 31 pages. doi:10.1145/3788872
- [5] Krisztian Balog, Donald Metzler, and Zhen Qin. 2025. Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. 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, 3865–3875. doi:10.1145/3726302.3730348
- [6] Charles Clarke and Laura Dietz. 2025. LLM-based Relevance Assessment Still Can't Replace Human Relevance Assessment. In *Proceedings of the Eleventh International Workshop on Evaluating Information Access, EVIA 2025, a Satellite Workshop of the NTCIR-18 Conference, Tokyo, Japan, June 10, 2025*. National Institute of Informatics (NII). doi:10.20736/0002002105
- [7] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
- [8] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient construction of large test collections. 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, 282–289. doi:10.1145/290941.291009
- [9] Tadele Tedla Damessie, Thao P. Nghiem, Falk Scholer, and J. Shane Culpepper. 2017. Gauging the Quality of Relevance Assessments using Inter-Rater Agreement. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 1089–1092. doi:10.1145/3077136.3080729
- [10] Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges. 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, 218–229. doi:10.1145/3731120.3744588
- [11] 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, ICTIR 2023, Taipei, Taiwan, 23 July 2023*, Masaharu Yoshioka, Julia Kiseleva, and Mohammad Aliannejadi (Eds.). ACM, 39–50. doi:10.1145/3578337.3605136
- [12] Hanpei Fang, Sijie Tao, Nuo Chen, Kai-Xin Chang, and Tetsuya Sakai. 2025. Do Large Language Models Favor Recent Content? A Study on Recency Bias in LLM-Based Reranking. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (China) (SIGIR-AP 2025)*. Association for Computing Machinery, New York, NY, USA, 85–94. doi:10.1145/3767695.3769493
- [13] Nicola Ferro and Mark Sanderson. 2022. How Do You Test a Test? A Multifaceted Examination of Significance Tests. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (Virtual Event, AZ, USA) (WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 280–288. doi:10.1145/3488560.3498406
- [14] Maik Fröbe, Andrew Parry, Harrison Scells, Shuai Wang, Shengyao Zhuang, Guido Zuccon, Martin Potthast, and Matthias Hagen. 2025. Corpus Subsampling: Estimating the Effectiveness of Neural Retrieval Models on Large Corpora. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 15572)*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonello (Eds.). Springer, 453–471. doi:10.1007/978-3-031-88708-6_29
- [15] Maram Hasanain, Yassmine Barkallah, Reem Suwaileh, Mucahid Kutlu, and Tamer Elsayed. 2020. ArTest: The First Test Collection for Arabic Web Search with Relevance Rationales. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (Virtual Event, China) (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2017–2020. doi:10.1145/3397271.3401223
- [16] William Hersh, Chris Buckley, T. J. Leone, and David Hickam. 1994. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In *SIGIR '94*, Bruce W. Croft and C. J. van Rijsbergen (Eds.). Springer London, London, 192–201.
- [17] Maurice G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1-2 (06 1938), 81–93. doi:10.1093/biomet/30.1-2.81
- [18] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).
- [19] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2356–2362.
- [20] Alistair Moffat, Falk Scholer, and Paul Thomas. 2012. Models and metrics: IR evaluation as a user process. In *Proceedings of the Seventeenth Australasian Document Computing Symposium (Dunedin, New Zealand) (ADCS '12)*. Association for Computing Machinery, New York, NY, USA, 47–54. doi:10.1145/2407085.2407092
- [21] Alistair Moffat, Falk Scholer, Paul Thomas, and Peter Bailey. 2015. Pooled Evaluation Over Query Variations: Users Are as Diverse as Systems. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, 1759–1762*. doi:10.1145/2806416.2806606
- [22] Samaneh Mohtadi and Gianluca Demartini. 2026. Query–Document Dense Vectors for LLM Relevance Judgment Bias Analysis. In *Proceedings of the 48th European Conference on Information Retrieval (ECIR 2026)*. to appear.
- [23] David Otero, Javier Parapar, and Álvaro Barreiro. 2025. Limitations of Automatic Relevance Assessments with Large Language Models for Fair and Reliable Retrieval Evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne (Eds.). ACM, 2545–2549. doi:10.1145/3726302.3730221
- [24] Ian Soboroff. 2025. Don't Use LLMs to Make Relevance Judgments. *Information Retrieval Research* 1, 1 (Mar. 2025), 29–46. doi:10.54195/irrr.19625
- [25] Reem Suwaileh, Mucahid Kutlu, Nihal Fathima, Tamer Elsayed, and Matthew Lease. 2016. ArabicWeb16: A New Crawl for Today's Arabic Web. In *Proceedings*

- of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 673–676. doi:10.1145/2911451.2914677
- [26] Paul Thomas, Douglas W. Oard, Eugene Yang, Dawn J. Lawrie, and James Mayfield. 2025. System Comparison Using Automated Generation of Relevance Judgements in Multiple Languages. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne (Eds.). ACM, 2812–2816. doi:10.1145/3726302.3730252
- [27] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models can Accurately Predict Searcher Preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) (SIGIR '24). Association for Computing Machinery, New York, NY, USA, 1930–1940. doi:10.1145/3626772.3657707
- [28] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: Umbrella is the (Open-Source Reproduction of the) Bing RElevance Assessor. arXiv:2406.06519 [cs.IR] <https://arxiv.org/abs/2406.06519>
- [29] Ellen M. Voorhees. 2001. Overview of TREC 2001. In *Proceedings of The Tenth Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA, November 13-16, 2001 (NIST Special Publication, Vol. 500-250)*, Ellen M. Voorhees and Donna K. Harman (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec10/papers/overview_10.pdf
- [30] Ellen M. Voorhees. 2002. Overview of TREC 2002. In *Proceedings of The Eleventh Text REtrieval Conference, TREC 2002, Gaithersburg, Maryland, USA, November 19-22, 2002 (NIST Special Publication, Vol. 500-251)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). <http://trec.nist.gov/pubs/trec11/papers/OVERVIEW.11.pdf>
- [31] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533* (2022).