Large Language Model Relevance Assessors Agree With One Another More Than With Human Assessors

Maik Fröbe
Friedrich-Schiller-Universität Jena
Germany, Jena

Andrew Parry
University of Glasgow
UK, Glasgow

Ferdinand Schlatt
Friedrich-Schiller-Universität Jena
Germany, Jena

Sean MacAvaney
University of Glasgow
UK, Glasgow

Benno Stein
Bauhaus Universität Weimar
Weimar, Germany

Martin Potthast
University of Kassel, hessian.AI, and ScaDS.AI
Germany, Kassel

Matthias Hagen
Friedrich-Schiller-Universität Jena
Germany, Jena

Abstract

Relevance judgments can differ between assessors, but previous work has shown that such disagreements have little impact on the effectiveness rankings of retrieval systems. This applies to disagreements between humans as well as between human and large language model (LLM) assessors. However, the agreement between different LLM assessors has not yet been systematically investigated. To close this gap, we compare eight LLM assessors on the TREC DL tracks and the retrieval task of the RAG track with each other and with human assessors. We find that the agreement between LLM assessors is higher than between LLMs and humans and, importantly, that LLM assessors favor retrieval systems that use LLMs in their ranking decisions: our analyses with 30-50 retrieval systems show that the system rankings obtained by LLM assessors overestimate LLM-based re-rankers by 9 to 17 positions on average.

CCS Concepts

• **Information systems → Evaluation of retrieval results**.

Keywords

IR Evaluation; Inter-Annotator Agreement; Reproducibility

ACM Reference Format:
Maik Fröbe, Andrew Parry, Ferdinand Schlatt, Sean MacAvaney, Benno Stein, Martin Potthast, and Matthias Hagen. 2025. Large Language Model Relevance Assessors Agree With One Another More Than With Human Assessors. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3726302.3730218

1 Introduction

In Cranfield experiments, sets of information needs, documents, and relevance judgments are compiled to evaluate the effectiveness of retrieval systems [37]. Collecting the relevance judgments from human assessors is typically the most expensive and time-consuming part of an experiment; a TREC track usually requires six trained assessors working for 2-4 weeks [29]. In this respect, decades of research on collecting relevance judgments have shown that, despite disagreements, a single judgment per query–document pair on a given topic suffices for reliable systems rankings (see Section 2).

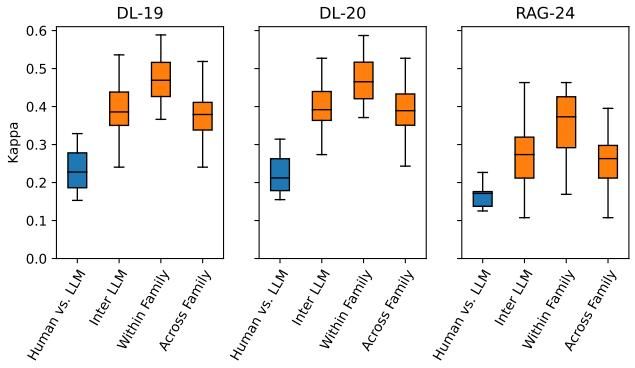
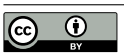


Figure 1: Agreement between judgement sets created by LLM and Human annotators measured by Fleiss' κ. Blue denotes the comparison between humans and LLMs; orange denotes the comparison between LLMs and other LLMs.

Following the introduction of large language models (LLMs), recent work has shown that their relevance judgments agree quite well with those of humans [1, 11, 12, 32, 33], suggesting that LLM assessors might complement or replace human assessors.[1] LLM assessors offer several advantages over human assessors, namely that LLMs are cheaper and faster, not affected by context switching, and that they potentially allow for greater diversity, e.g., by simulating panels. However, while a single human assessor may be sufficient to collect reliable relevance judgments, this has not been established for LLM-based relevance judgments. It is also not clear which LLM is best suited for the task and whether the use of different LLMs (e.g., in independent experiments) leads to comparable results.

In this paper, we address these issues for the first time by investigating the inter-annotator agreement of different LLMs for relevance judgments. Just like the study of human agreement in past IR research, understanding LLM agreement will inform the design of potential LLM-assisted evaluation paradigms. As a first

[1] This, in turn, has sparked a discussion about the risks and limitations of fully automating relevance judgments and its impact on retrieval evaluation [1, 11–13, 18, 29, 32–34].



SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730218

step in this direction, we conduct experiments on the TREC Deep Learning Tracks 2019/2020 [6, 7] and the retrieval task of the TREC RAG Track 2024 [33] with eight LLM assessors. We compare the LLMs with each other and with human relevance judgments.

Our results show that LLM assessors consistently agree with themselves more than with human assessors (Figure 1). On the one hand, this is positive since we do not need to use multiple models to produce equally discriminative evaluations. On the other hand, this highlights the disadvantage recently observed by Clarke and Dietz [4] (later confirmed by Balog et al. [1]) that retrieval pipelines with LLMs might be preferred by LLM assessors. We evaluate this empirically and confirm that relevance assessments of LLMs overestimate the effectiveness of LLM re-rankers. Therefore, caution is needed in evaluations with LLM relevance assessments, as the goal of most retrieval systems should be to satisfy human users and not LLMs. Our code and data is publicly available.[2]

2 Related Work

We review related work on the inter-annotator agreement for the human vs. human and the human vs. LLM scenario.

A Short History on Relevance Judgments in IR. Relevance judgments and their properties have been studied for decades [8, 17, 26]. Relevance is highly subjective [8, 27] and affected by many variables [31]. This subjectivity is reflected in the inconsistency of relevance judgments across human assessors. The resulting disagreement has negligible impact on systems rankings, though, because they rely on aggregated values and disagreements often occur on outlier topics [8, 17, 26, 36]. These observations guided the decision that a single human annotator is usually assigned per topic [36].

Inter-annotator agreement of humans. The original Cranfield experiments did exhaustive relevance judgments for all documents and all information needs [5]. As this exhaustive judgment is only possible for small corpora, pooling the results of all retrieval systems in the comparison allowed to scale to large corpora under the assumption that unpooled documents are not relevant [36]. The assumption that unpooled documents are not relevant forms a trade-off between coverage of relevant documents and budget. A single primary annotator is assigned to each topic to reduce the costs [36]. This decision was validated by measuring inter-annotator agreement in relevance assessments under several retrieval evaluation measures and retrieval scenarios [3, 17, 35]. Relevance is subjective, and multiple factors influence relevance judgements depending on the intent and ambiguity of information needs [8, 26]. Nevertheless, while humans disagree on the relevance of documents due to this subjectivity, downstream system rankings and not impacted by this disagreement, i.e., system rankings are highly correlated even when there is disagreement for individual relevance decisions [3, 17, 35].

Inter-annotator agreement between LLMs and humans. Automated evaluation of relevance has been in the interest of the community for more than two decades [21, 30, 38]. However, this research direction received renewed interest due to improvements in LLMs [28], ranging to applications of LLM assessors in diverse languages [9], image retrieval [39], and automatically generated answers across diverse modalities [10, 15, 16, 23, 39]. Faggioli et al. [11] explored different levels of support of human assessors with LLM assessors, showing that the agreement between LLM relevance assessments and humans is good and that the impact of the disagreement on system rankings is negligible. MacAvaney and Soldaini [18] also argued that automated judgments should support human annotations and should not replace them. These studies and subsequent work [24, 34] have observed that evaluation with LLM annotations produces stable and system rankings that are highly correlated with that of human annotations. Upadhyay et al. [33] used automated relevance judgments in the first Retrieval-Augmented-Generation track at TREC 2024, arguing that automated judgments can replace human judgments. This was contested by Clarke and Dietz [4] and Balog et al. [1], showing that there exists a bias in language models so that runs produced by LLMs are preferred by LLM assessors. With our work, we aim to expand those observations and verify if this also happens on a larger scale. Faggioli et al. [12] give a perspective on how the human-like aspects of language models may lead them to fail to annotate relevance correctly; we consider a quantitative assessment similar to those done when Cranfield testing was an emerging paradigm may shed light on how language models compare to human annotators.

3 Automated Relevance Assessments with LLMs

We implement automated LLM relevance assessors so that they take a query and a passage as prompted input to return a relevance grade. While initiatives such as the LLMJudge shared task [25] highlight how diverse LLM assessment pipelines can be, we go for a pipeline as simplistic as possible that we keep constant and only change the LLM while keeping all other components identical. We do this so that we focus our experiments on the LLM, i.e., ensuring that we use the same data, preprocessing, and prompt but only switch out the LLM. We use `ir_datasets` [19] for access to queries and document texts, using the default text of queries (the title field for the TREC Deep Learning tracks) and the default text of the passages as implemented in `ir_datasets`. We embed the query and the passage into a prompt and submit this prompt to a LLM, asking for relevance judgments between 0 (not relevant) and 3 (highly relevant) that we parse from the generated response. We use the prompt and the response parser of UMBRELA [34].

Many labs do not have the resources to locally run LLMs. Thus, many use REST-API-based systems served on massive compute. As such, we use a selection of these systems to allow our investigation to best reflect real use cases. We use multiple models from many industry providers. Overall, we incorporate 8 LLMs from 4 different providers for our experiments.[3] We use 2 models from the Claude family, the smaller Haiku and the larger Sonnet model in its 3.0 versions. From the Gemini family, we use Gemini-1.5-flash and Gemini-1.5-flash-8b. From the GPT family, we use GPT-4o and GPT-4o-mini. In addition to closed source models, we use Llama-3 and Llama-3.1 with 70 billion parameters, which provides a helpful comparison as a model that can be served on academic budgets.

[2] https://github.com/webis-de/SIGIR-25

[3] We use: Claude-3-haiku, Claude-3-sonnet, Llama-3, Llama-3.1, Gemini-1.5-flash, Gemini-1.5-flash-8b, GPT-4o, and GPT-4o-mini.

Table 1: Distribution of relevance judgments for the TREC Deep Learning tracks in 2019 (DL-19), in 2020 (DL-20), and for the 2024 Retrieval-Augmented Generation track (RAG-24) for the eight LLM assessors and human annotations (Official).

Annotator	TREC-DL-19				TREC-DL-20				TREC-RAG-24			
	0	1	2	3	0	1	2	3	0	1	2	3
Claude-3-haiku	.283	.173	.227	.317	.380	.192	.209	.219	.136	.293	.297	.274
Claude-3-sonnet	.262	.186	.298	.254	.305	.193	.295	.207	.143	.355	.324	.178
Gemini-1.5-flash	.278	.290	.292	.140	.373	.297	.230	.100	.095	.324	.521	.060
Gemini-1.5-flash-8b	.454	.181	.316	.049	.549	.151	.273	.028	.352	.178	.464	.005
GPT-4o	.355	.360	.173	.112	.455	.336	.122	.087	.136	.506	.268	.090
GPT-4o-mini	.294	.409	.179	.118	.376	.419	.127	.078	.090	.496	.323	.091
Llama-3	.250	.218	.413	.118	.331	.225	.351	.093	.106	.214	.560	.121
Llama-3.1	.217	.249	.431	.103	.312	.262	.342	.084	.087	.211	.608	.094
Official	.557	.173	.195	.075	.683	.170	.090	.057	.373	.311	.230	.086

Table 2: Comparison of the eight LLM assessors with human relevance judgments as agreement (Fleiss' κ, Jaccard similarity) and impact on system rankings (Kendall's τ, Spearman's ρ when possible) on Deep Learning 2019 (DL-19), 2020 (DL-20), and Retrieval-Augmented Generation 2024 (RAG-24).

Annotator	DL-19				DL-20				RAG-24	
	κ	Jacc.	τ	ρ	κ	Jacc.	τ	ρ	κ	Jacc.
Claude-3-haiku	.165	.378	.901	.980	.197	.457	.898	.983	.138	.339
Claude-3-sonnet	.153	.369	.874	.973	.155	.391	.928	.990	.172	.382
Gemini-1.5-flash	.260	.462	.859	.966	.227	.484	.937	.992	.175	.391
Gemini-1.5-flash-8b	.329	.563	.901	.982	.276	.585	.861	.970	.179	.420
GPT-4o	.316	.522	.895	.977	.314	.578	.939	.992	.227	.441
GPT-4o-mini	.266	.470	.901	.980	.258	.513	.937	.992	.171	.395
Llama-3	.193	.408	.862	.968	.171	.421	.910	.987	.137	.348
Llama-3.1	.195	.401	.838	.954	.181	.422	.943	.993	.125	.341

4 Evaluation

We study the relevance grades generated by 8 LLM assessors with respect to their (1) relevance distributions, (2) inter-annotator agreement, and (3) the impact on system rankings.

4.1 Experimental Setup

We use the 2019 and 2020 editions of the TREC Deep Learning tracks [6, 7] on the MS MARCO passage collection [20] and the 2024 TREC Retrieval-Augmented Generation track on version 2.1 of MS MARCO. Both corpora contain passages extracted from documents and queries mined from the Bing Search Engine. The relevance judgments for the 2019 and 2020 Deep Learning tracks are publicly available, and because of the popularity of MS MARCO, potentially included in the training procedure of LLMs, whereas the relevance judgments for the 2024 edition were not public during our experiments and only available to TREC participants. We use all runs submitted to the 2019/2020 Deep Learning tracks (37 runs in 2019 and 59 in 2020) for our experiments on system ranking correlations, as those runs are available password protected whereas the 2024 runs are not available at the time of the experiments.[4]

Measures. We use several measures of agreement from broader literature and IR-specific measures of agreement. We report Fleiss' κ implemented in nltk [2] and the agreement as implemented in trectools [22] to compare relevance judgments. For comparing system rankings, we use Kendall's τ and Spearman's ρ. In this case, we measure correlation not of the annotations themselves but for all systems that participated in the TREC Deep Learning track when evaluating for nDCG@10. This allows us to validate that system order is stable, which is important as annotators may disagree; ultimately, if system A is considered better than system B in all cases, we can disregard annotation discrepancy.

4.2 Results

LLMs are optimistic annotators. We first investigate how, on densely annotated corpora, an LLM annotates relevance compared to a human in terms of grade distributions. Table 1 shows the distribution of relevance judgments of the official human judgments and our eight LLM assessors. Like prior work of Faggioli et al. [11], we observe that LLMs tend to label more documents as relevant (≥ 0). We also observe this trend to overestimation for highly relevant documents (label 3), as most LLMs assess many documents as highly relevant, with Claude-Haiku being the most optimistig with between 21.9 % (for DL-20) and 31.7 % (for DL-19) compared to 5.7 % and 7.5 % for human judgments. Though we generally observe overestimation of relevance, Gemini 1.5-flash-8b is an outlier that most closely follows the human distribution that non-relevant documents are most frequent. In terms of model families, we find that the Claude variants tend to overestimate perfect relevance, with the smaller Haiku variant showing greater overestimation. In addition, the GPT variants show similar distributions with an overestimation (relative to humans) of related texts (Label 1). We primarily see trends in estimation based on model family rather than model size; possibly due to a similar training distribution.

LLMs disagree with humans but system rankings are reliable. Decades of research have shown that human relevance judgments are subjective but that the resulting disagreements have little impact on systems rankings [3, 17, 36]. This has also been observed for LLM-based relevance assessors [11, 18, 24, 33, 34]. We reproduce both in our experiments for all our eight LLM assessors.

Table 2 shows the inter-annotator agreement between the official human relevance judgments and our LLM judgments and their impact on system rankings (not reported for RAG-24 as no runs are available at the time of the experiments). The agreement to human relevance judgments is in line with prior work, with a Fleiss´ κ between 0.138 (for Claude-Haiku on RAG-24) and 0.329 (for Gemini-1.5-flash-8b on DL-19). Furthermore, the impact on system rankings is negligible in all cases (1.0 shows perfect, 0.0 random correlation, Voorhees [36] reported a Kendall's τ between 0.85 and 0.9), with a minimum Kendall's τ of 0.838 (for Llama-3.1 on DL-19) and a minimum Spearman's ρ of 0.954 (again for Llama-3.1 on DL-19).

LLM–LLM agreement is greater than LLM–Human agreement. We compare the inter-annotator agreement of (1) human vs. LLM judgments and (2) LLM judgments vs. LLM judgments. Figure 1 shows that the median LLM–LLM agreement is between 0.1 and 0.16 higher than the median human–LLM agreement. The agreement is even higher for LLMs from the same model family. This high agreement suggests that using a diverse set of models for creating relevance judgments does not provide a diverse set of judgments, especially for models from the same family.

[4] DL 19: https://trec.nist.gov/results/trec28/deep.passages.input.html
DL 20: https://trec.nist.gov/results/trec29/deep.passages.input.html

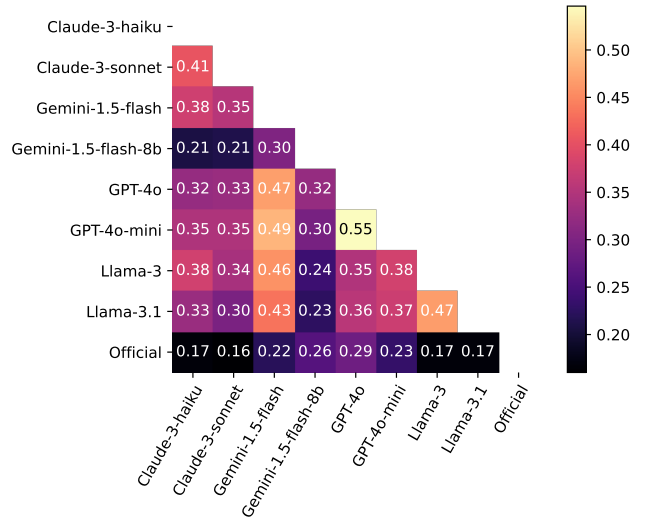


Figure 2: Fleiss' κ agreement between LLM assessors and human relevance judgments macro-averaged over the 2019 and 2020 TREC Deep Learning and the 2024 TREC RAG tracks.

Figure 2 provides a more fine-grained view of the agreement of LLMs with one another and the human judgments macro-averaged over the three collections. We find a comparatively low LLM–human agreement illustrated by the dark bottom row in the heatmap. The Claude and Llama models have an especially low agreement with human judgments, while the Gemini and GPT models show a higher agreement. This consistent greater agreement with each other over humans is concerning when coupled with the arguments of Clarke and Dietz [4]. Since LLMs are increasingly used in retrieval pipelines, and they agree with each other more than with human judgments, retrieval pipelines with LLMs might be preferred by LLM assessors and their effectiveness might be overestimated.

LLM evaluators favor LLM re-rankers. We use all eight LLM assessors as re-rankers under different evaluation scenarios to validate if the high agreement of LLM-based assessors causes that LLM-based retrieval pipelines are overestimated by LLM assessors. This is re-ranker setup is common, as expensive models can not run on the complete corpus [14]. We re-rank all runs submitted to the 2019 and 2020 Deep Learning track (no runs available for TREC RAG at the time of experiments) with all eight LLM assessors. We use nDCG@10 as an evaluation measure as this was the primary measure in the Deep Learning tracks, and we use the system ranking and nDCG@10 scores from the official human relevance judgments as ground truth. For every run, we use one LLM as a re-ranker and another LLM as a relevance assessor, iterating over all combinations of runs, re-ranker LLMs, and relevance assessor LLMs. For every triple of a to-be-re-ranked run, an LLM re-ranker, and an LLM assessor, we first re-rank the run with the LLM and then evaluate all systems (i.e., the re-ranked run and all other, original runs) twice to create two system rankings. First, the ground truth system ranking from human judgments, and second, the system ranking generated by the LLM assessor (that is different from the LLM re-ranker). We then compare the position of the LLM re-ranker in the ground truth system ranking with the position of the LLM re-ranker in the system ranking derived from the LLM assessor. If LLM-based relevance assessors favor LLM re-rankers, we would observe that the position of the re-ranked system is higher in the LLM-evaluated system ranking than in the human evaluated system ranking.

Table 3: Overview how LLM-based evaluators overestimate LLM-based re-rankers (mean, 25 %, and 75 % quantiles) in system rankings (Δ Rank) and nDCG@10 scores (Δ Score) when re-ranking pre-trained language models (NNLM), neural networks (NN), traditional (Trad), or all approaches submited to the 2019 and 2020 TREC Deep Learning tracks. We compare when re-rankers and evaluators come from the same LLM family (within) or from different LLM families (accross).

Systems		DL 19						DL 20					
		Δ Rank			Δ Score			Δ Rank			Δ Score		
		25 %	Avg.	75 %	25 %	Avg.	75 %	25 %	Avg.	75 %	25 %	Avg.	75 %
Within	NNLM	9.0	10.5	13.0	.146	.176	.200	9.0	17.5	27.0	.166	.200	.244
	NN	9.8	11.2	13.2	.147	.181	.202	10.2	17.5	24.0	.164	.203	.241
	Trad	8.8	10.2	13.0	.147	.173	.196	10.8	18.8	28.0	.176	.208	.248
	All	9.0	10.5	13.0	.146	.175	.199	9.0	17.8	27.0	.166	.202	.246
Accross	NNLM	3.0	8.8	12.0	.132	.178	.202	6.0	17.2	27.0	.154	.205	.240
	NN	4.0	9.5	13.0	.138	.183	.207	8.5	17.4	26.2	.159	.208	.243
	Trad	3.0	8.9	12.0	.121	.176	.201	9.0	18.8	28.0	.167	.214	.251
	All	3.0	9.0	12.0	.129	.178	.204	7.0	17.6	27.0	.156	.207	.242

Table 3 shows how LLM-based evaluators overestimate LLM-based re-rankers across the three classes of retrieval systems submitted to the deep learning tracks (NNLM systems that use pre-trained language models, NN systems that use some neural networks without pre-trained models, traditional models such as BM25, and all systems). We report by how many ranks the LLM-based re-ranker is overestimated in the LLM-assessed system ranking (Δ Rank) and the overestimation in the nDCG@10 scores (Δ Score) on average and for the 25 % respectively 75 % quantiles. LLM-based re-rankers are especially overestimated by LLM assessors from the same family; by 10.5 positions on average for DL-19 (with 37 runs), and 17.8 positions on average for DL-20 (59 runs). Even across LLM families, LLM-re-rankers are substantially overestimated by 9.0 positions on average for DL-19 and 17.6 positions on average for DL-20. This highlights that LLM-based relevance assessors substantially overestimate LLM-based retrieval pipelines, even when the evaluation LLM differs from the one used in the retrieval system.

5 Conclusion and Future Work

We studied the inter-annotator agreement of LLMs for relevance assessments. Our experiments showed that the inter-annotator agreement is substantially higher between LLM assessments than between humans and LLMs. Furthermore, we showed that this higher agreement impacts evaluations of retrieval systems, as LLM-based assessors overestimate the effectiveness of retrieval pipelines that use LLMs for internal relevance decisions. Interesting directions for future work could be to identify ways to substantially increase or decrease the agreement of LLM assessors, depending on the use-case. Furthermore, it would be interesting to validate if switching the LLM within a topic yields more realistic automated assessments as LLMs can easily switch context.

References

[1] Krisztian Balog, Donald Metzler, and Zhen Qin. 2025. Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. *CoRR* abs/2503.19092 (2025). doi:10.48550/ARXIV.2503.19092 arXiv:2503.19092

[2] Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, Nicoletta Calzolari, Claire Cardie, and Pierre Isabelle (Eds.). The Association for Computer Linguistics. doi:10.3115/1225403.1225421

[3] Robert Burgin. 1992. Variations in relevance judgments and the evaluation of retrieval performance. *Inf. Process. Manage.* 28, 5 (July 1992), 619–627. doi:10.1016/0306-4573(92)90031-T

[4] Charles L. A. Clarke and Laura Dietz. 2024. LLM-based relevance assessment still can't replace human relevance assessment. arXiv:2412.17156 [cs.IR] https://arxiv.org/abs/2412.17156

[5] C. Cleverdon, J. Mills, and M. Keen. 1966. *Factors Determining the Performance of Indexing Systems. Volume I. Design. Part 2. Appendices.* Technical Report PB169574. Association of Special Libraries and Information Bureau, Cranfield (England). https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB169574.xhtml Num Pages: 261.

[6] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2020 Deep Learning Track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020 (NIST Special Publication, Vol. 1266)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *CoRR* abs/2003.07820 (2020). arXiv:2003.07820

[8] Carlos A. Cuadra and Robert V. Katter. 1967. Opening the Black Box of Relevance. *Journal of Documentation* 23, 4 (April 1967), 291–303. doi:10.1108/eb026436

[9] Gabriel de Jesus and Sérgio Sobral Nunes. 2024. Exploring Large Language Models for Relevance Judgments in Tetun. In *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) co-located with 10th International Conference on Online Publishing (SIGIR 2024), Washington D.C., USA, July 18, 2024 (CEUR Workshop Proceedings, Vol. 3752)*, Clemencia Siro, Mohammad Aliannejadi, Hossein A. Rahmani, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz (Eds.). CEUR-WS.org, 19–30.

[10] Niklas Deckers, Maik Fröbe, Johannes Kiesel, Gianluca Pandolfo, Christopher Schröder, Benno Stein, and Martin Potthast. 2023. The Infinite Index: Information Retrieval on Generative Text-To-Image Models. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR 2023, Austin, TX, USA, March 19-23, 2023*, Jacek Gwizdka and Soo Young Rieh (Eds.). ACM, 172–186. doi:10.1145/3576840.3578327

[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] 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. 2024. Who Determines What Is Relevant? Humans or AI? Why Not Both? *Commun. ACM* 67, 4 (2024), 31–34. doi:10.1145/3624730

[13] Naghmeh Farzi and Laura Dietz. 2024. EXAM++: LLM-based Answerability Metrics for IR Evaluation. In *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) co-located with 10th International Conference on Online Publishing (SIGIR 2024), Washington D.C., USA, July 18, 2024 (CEUR Workshop Proceedings, Vol. 3752)*, Clemencia Siro, Mohammad Aliannejadi, Hossein A. Rahmani, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz (Eds.). CEUR-WS.org, 31–50. https://ceur-ws.org/Vol-3752/paper3.pdf

[14] Maik Fröbe, Andrew Parry, Harrisen 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 Tonellotto (Eds.). Springer, 453–471. doi:10.1007/978-3-031-88708-6_29

[15] Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Evaluating Generative Ad Hoc Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 1916–1929. doi:10.1145/3626772.3657849

[16] Jia-Hong Huang, Hongyi Zhu, Yixian Shen, Stevan Rudinac, Alessio M. Pacces, and Evangelos Kanoulas. 2024. A Novel Evaluation Framework for Image2Text Generation. In *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) co-located with 10th International Conference on Online Publishing (SIGIR 2024), Washington D.C., USA, July 18, 2024 (CEUR Workshop Proceedings, Vol. 3752)*, Clemencia Siro, Mohammad Aliannejadi, Hossein A. Rahmani, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz (Eds.). CEUR-WS.org, 51–65.

[17] M.E. Lesk and G. Salton. 1968. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval* 4, 4 (Dec. 1968), 343–359. doi:10.1016/0020-0271(68)90029-6

[18] Sean MacAvaney and Luca Soldaini. 2023. One-Shot Labeling for Automatic Relevance Estimation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2230–2235. doi:10.1145/3539618.3592032

[19] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified Data Wrangling with ir_datasets. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2429–2436. doi:10.1145/3404835.3463254

[20] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[21] Rabia Nuray and Fazli Can. 2006. Automatic ranking of information retrieval systems using data fusion. *Inf. Process. Manage.* 42, 3 (May 2006), 595–614. doi:10.1016/j.ipm.2005.03.023

[22] João R. M. Palotti, Harrisen Scells, and Guido Zuccon. 2019. TrecTools: an Open-source Python Library for Information Retrieval Practitioners Involved in TREC-like Campaigns. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 1325–1328. doi:10.1145/3331184.3331399

[23] Zackary Rackauckas, Arthur Câmara, and Jakub Zavrel. 2024. Evaluating RAG-Fusion with RAGElo: an Automated Elo-based Framework. In *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) co-located with 10th International Conference on Online Publishing (SIGIR 2024), Washington D.C., USA, July 18, 2024 (CEUR Workshop Proceedings, Vol. 3752)*, Clemencia Siro, Mohammad Aliannejadi, Hossein A. Rahmani, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz (Eds.). CEUR-WS.org, 92–112.

[24] Hossein A. Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. 2024. Synthetic Test Collections for Retrieval Evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 2647–2651. doi:10.1145/3626772.3657942

[25] Hossein A. Rahmani, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, Paul Thomas, Charles L. A. Clarke, Mohammad Aliannejadi, Clemencia Siro, and Guglielmo Faggioli. 2024. LLMJudge: LLMs for Relevance Judgments. In *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) co-located with 10th International Conference on Online Publishing (SIGIR 2024), Washington D.C., USA, July 18, 2024 (CEUR Workshop Proceedings, Vol. 3752)*, Clemencia Siro, Mohammad Aliannejadi, Hossein A. Rahmani, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz (Eds.). CEUR-WS.org, 1–3.

[26] Alan M. Rees and Douglas G. Schultz. 1967. *A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. Final Report to the National Science Foundation. Volume I.* Technical Report. Clearinghouse for Federal Scientific and Technical Information, Springfield, Va.

[27] Linda Schamber, Michael B. Eisenberg, and Michael S. Nilan. 1990. A Re-Examination of Relevance: Toward a Dynamic, Situational Definition*. *Information Processing & Management* 26, 6 (Jan. 1990), 755–776. doi:10.1016/0306-

4573(90)90050-C

[28] Clemencia Siro, Mohammad Aliannejadi, Hossein A. Rahmani, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz (Eds.). 2024. *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) co-located with 10th International Conference on Online Publishing (SIGIR 2024), Washington D.C., USA, July 18, 2024.* CEUR Workshop Proceedings, Vol. 3752. CEUR-WS.org.

[29] Ian Soboroff. 2024. Don't Use LLMs to Make Relevance Judgments. *CoRR* abs/2409.15133 (2024). doi:10.48550/ARXIV.2409.15133 arXiv:2409.15133

[30] Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking retrieval systems without relevance judgments *(SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 66–73. doi:10.1145/383952.383961

[31] Mortimer Taube. 1965. A Note on the Pseudo-Mathematics of Relevance. *American Documentation* 16, 2 (1965), 69–72. doi:10.1002/asi.5090160204

[32] 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, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 1930–1940. doi:10.1145/3626772.3657707

[33] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2024. A Large-Scale Study of Relevance Assessments with Large Language Models: An Initial Look. *CoRR* abs/2411.08275 (2024). doi:10.48550/ARXIV.2411.08275 arXiv:2411.08275

[34] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. *CoRR* abs/2406.06519 (2024). doi:10.48550/ARXIV.2406.06519 arXiv:2406.06519

[35] Ellen Voorhees. 2000. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. 36 No. 5 (2000-01-01 2000).

[36] Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 315–323. doi:10.1145/290941.291017

[37] Ellen M. Voorhees. 2019. The Evolution of Cranfield. In *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, Nicola Ferro and Carol Peters (Eds.). The Information Retrieval Series, Vol. 41. Springer, 45–69. doi:10.1007/978-3-030-22948-1_2

[38] Shengli Wu and Fabio Crestani. 2002. Data fusion with estimated weights. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (McLean, Virginia, USA) *(CIKM '02)*. Association for Computing Machinery, New York, NY, USA, 648–651. doi:10.1145/584792.584908

[39] Jheng-Hong Yang and Jimmy Lin. 2024. Toward Automatic Relevance Judgment using Vision-Language Models for Image-Text Retrieval Evaluation. In *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) co-located with 10th International Conference on Online Publishing (SIGIR 2024), Washington D.C., USA, July 18, 2024 (CEUR Workshop Proceedings, Vol. 3752)*, Clemencia Siro, Mohammad Aliannejadi, Hossein A. Rahmani, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz (Eds.). CEUR-WS.org, 113–123.