On the Viability of Exploiting Large Language Models for Misinformation Annotation

Pablo Landrove[1], Marcos Fernández-Pichel[1], and David E. Losada[1]

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain
pablo.landrove@rai.usc.es
{marcosfernandez.pichel, david.losada}@usc.es

Abstract. This paper investigates the potential of LLMs for automatically annotating the usefulness, supportiveness, and credibility of search results. These aspects, while essential to the construction of misinformation benchmarks, are expensive and difficult to obtain at scale. Our comparative study suggests that, under certain conditions, LLMs can provide reasonable estimates of usefulness and supportiveness. In contrast, credibility judgments generated by LLMs show almost no agreement with human assessments. This raises concerns for the exploitation of LLMs to assist in the construction of collections that require annotations that go beyond relevance.

Keywords: Information Retrieval, LLMs, Automatic Labeling, Misinformation

1 Introduction

In Information Retrieval (IR), the exploitation of LLMs for automatic annotation of documents has mainly focused on relevance labeling. Over the past two years, there has been intense activity in this area [5,10,19,4]. Some discrepancies remain regarding the impact of LLMs in the generation of synthetic qrels, but it is generally accepted that LLMs represent promising tools to support the construction of IR benchmarks [4]. Although some authors express reluctance to use LLMs to produce qrels [17,6], other studies advocate full replacement of human assessors or, at least, the use of LLMs to reduce human labeling workload [22,19]. Related to this, some authors have proposed LLM-assisted assessments that combine manual and automatic annotations [18], or generation of relevance judgments through multi-criteria methods [11].

However, little attention has been devoted to exploring LLM-based annotation of aspects beyond relevance. A few attempts have been made (e.g., addressing usefulness or utility judgments [24,9]), but further work is needed in this direction. In this paper, we explore the feasibility of using LLMs to annotate aspects that are more difficult to estimate than relevance. In particular, we focus on variables such as credibility, which play a crucial role in benchmarks addressing misinformation.

Our investigation is oriented towards search scenarios in the health domain, where the construction of test collections requires not only determining the relevance of a document but also its usefulness to the user, its stance regarding a specific treatment (supportive or non-supportive), and its credibility. Following prior work [19,23,12], we design different prompt variants adapted to the generation of judgments of credibility, supportiveness, and usefulness. Some of these aspects are inherently subjective and, thus, it is essential to understand the limits of LLMs in producing estimates that resemble those made by humans.[1] Our experimental results suggest that LLMs can reasonably estimate usefulness and supportiveness but, in contrast, the credibility estimates seem useless.

2 Related Work

Human relevance judgments are costly and time-consuming, thus making the creation of comprehensive test collections infeasible. LLMs are increasingly being explored as a solution to the long-standing challenge of building relevance assessments. LLMs offer the potential to assess large volumes of documents cheaply, thereby enabling scalable and cost-efficient test collection development. Researchers have investigated various approaches, including zero-shot, one-shot, few-shot, and fine-tuning methods, utilizing both commercial LLMs and open-source models [10,19,16,12,1,15,21,23]. A consistent finding across studies is that LLM-generated judgments often yield highly correlated system rankings with human judgments, even if the agreement on individual relevance labels is substantially lower. However, some authors have already raised concerns about the potential risks of automatic evaluation, such as biases, circularity, and unreliable identification of the best-performing systems [17,6].

Although the primary focus in the literature has been annotations of relevance, some authors have explored the role of LLMs in estimating other types of variables. For example, Zhang et al. [24], in the context of RAG for Question Answering, examined the ability of LLMs to estimate the potential usefulness of retrieved passages. Similarly, Dewan et al. [9] investigated the prediction of document usefulness, defined as the contribution to achieving the user's goals within a search session. Our work seeks to join this line of research that leverages LLMs to annotate additional criteria.

We have designed a prompt whose template incorporates multiple criteria and, thus, our work is also related to recent studies that generate relevance judgments through the aggregation of diverse criteria [11]. We are not interested in estimating relevance but rather in annotating other variables that are critical for misinformation benchmarks. Nonetheless, we share with these studies the underlying philosophy of encouraging the LLM to reason based on multiple criteria (instead of requesting a single generic label for documents).

[1] We release all the judgments and code derived from this research: `https://github.com/PabloLandro/llms-viability-misinfo-label`

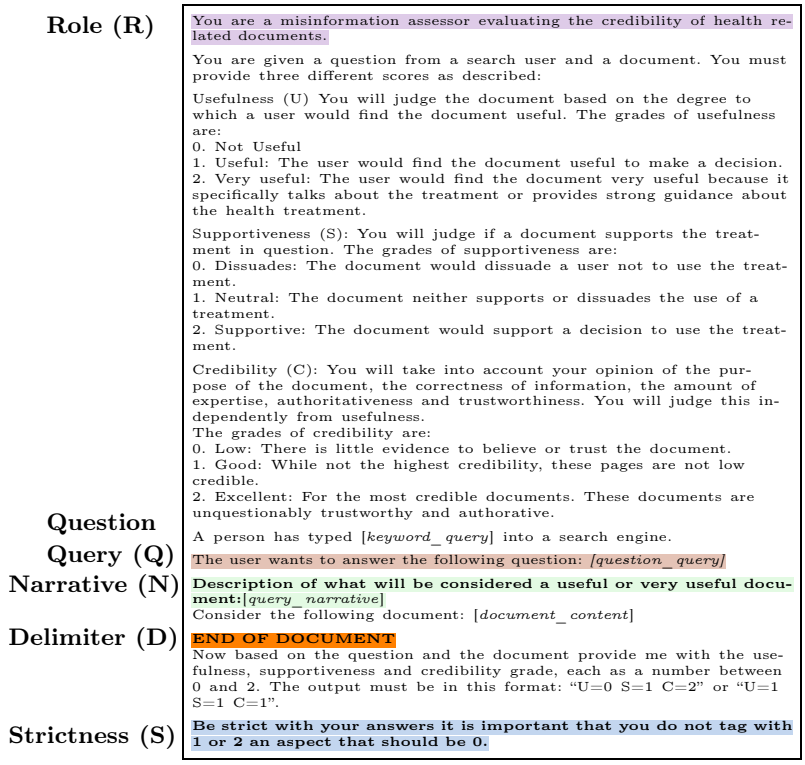

Fig. 1: Template for annotating usefulness, supportiveness and credibility with LLMs. Shaded parts are optional (included in some prompt variants) and italicized texts are placeholders. The best performance for each metric is bolded.

3 Experimental Design

Inspired by Thomas et al. [19], we have designed a prompt template to systematically assess the impact of different instructions (see Figure 1). We adapted its structure including elements that have being proven useful in the literature [19,20]. The template includes a system role (R), a query in question form (Q), a narrative (N), a delimiter to clearly specify the end of the input document (D), and a final instruction asking the LLM to be strict when providing the scores (S). This final element was not considered in previous studies, but we decided to include it due to the tendency of LLMs toward leniency. The three target variables (usefulness, supportiveness and credibility) are estimated following a 3-graded scheme (0, 1 or 2). We experimented with gpt-4o-mini, and the temperature was set to 0.7 and top_p to 1.

Our experiments were carried out on the test collections built under the TREC Health Misinformation (HM) 2019, 2021 and 2022 tracks [2,8,7]. The

main goal of these tracks was to develop retrieval systems that promote credible and correct health information over misinformation. Thus, multi-criteria judgments (e.g., usefulness of the document, supportiveness regarding a specific treatment, and perceived credibility of the document) were produced. There may be some overlap between these variables; however, we deliberately adopted this labeling strategy because these were the variables annotated by humans in our reference collections. From these individual judgments, a preference of helpful and harmful documents was generated (the most helpful documents are useful, credible, and support a correct treatment, while the most harmful documents are useful, credible and support an incorrect treatment). These derived qrels were then used to compute a compatibility measure that combines helpfulness and harmfulness.

Following prior work [19,3], we report Mean Absolute Error (MAE) and Cohen's κ between the (graded) ground-truth judgments and the LLMs' estimates. Given the LLM-based judgments, we also ranked all the systems participating in these evaluation campaigns (using the official metric, compatibility helpful-harmful) and then we compared this ranking with that obtained with the official judgments (using Rank-biased Overlap, RBO).[2]

The cost of evaluating all prompt combinations with all judged documents would be prohibitive and, thus, we followed a sampling approach to repeatedly select 1,000 documents from TREC HM 2021.[3] This sampling was repeated 20 times and we report averages and standard deviations. Once all prompt variations were evaluated, we proceeded to transfer the best performing prompt to judge all pooled documents in the other TREC HM collections.

4 Results

Table 1 (MAE and κ) and Figure 2 (RBO) compare the effectiveness of different prompts following the sampling approach described above. We can observe that:

- MAEs and kappas show substantial variation depending on the specific prompt. Such sensitivity to the input prompt features is a commonly acknowledged phenomenon in LLM research.
- The LLM estimates of usefulness and supportiveness appear somewhat promising, reaching κ in some cases above 0.5, which is considered close to human performance [4]. In contrast, credibility estimates are ineffective. To put these results in perspective, it is important to note that [13] examined the agreement of credibility judgments among humans, finding only moderate agreement (median κ 0.44). Nevertheless, such level of human agreement remains much higher than the one reported in the table between the LLM and the golden truth judgments.

[2] In RBO, we set the persistence parameter to 0.95, which means that in average we are considering the top-20 ranked systems from the evaluation campaign.

[3] The sampled documents and the LLM-based judgments are available at `https://github.com/PabloLandro/llms-viability-misinfo-label`.

Table 1: Performance of different prompts. R, Q, N, D, and S mean the inclusion of role, question query, narrative, delimiter and strictness, respectively.

Prompt features					**Usefulness**		**Credibility**		**Supportiveness**	
R	Q	N	D	S	MAE	κ	MAE	κ	MAE	κ
–	–	–	–	–	0.44±0.01	0.11±0.01	0.46±0.02	0.00±0.00	0.45±0.27	0.13±0.05
–	–	–	D	–	0.38±0.02	0.16±0.02	0.37±0.03	0.00±0.00	0.43±0.21	0.10±0.04
–	–	–	D	S	0.38±0.03	0.16±0.03	0.37±0.03	0.01±0.01	0.38±0.24	0.15±0.05
–	Q	–	–	–	0.26±0.02	0.45±0.03	0.36±0.03	0.02±0.02	0.31±0.19	0.45±0.22
–	Q	–	D	–	0.24±0.02	0.49±0.03	0.37±0.02	0.01±0.01	0.30±0.19	0.47±0.21
–	Q	–	D	S	0.19±0.03	0.50±0.09	0.35±0.04	0.01±0.01	0.24±0.18	0.48±0.29
–	Q	N	–	–	0.25±0.02	0.47±0.04	0.37±0.03	0.01±0.01	0.32±0.23	0.49±0.28
–	Q	N	D	–	0.20±0.02	0.46±0.08	0.35±0.02	0.02±0.01	0.28±0.22	0.45±0.28
–	Q	N	D	S	0.21±0.02	0.45±0.05	0.35±0.04	0.02±0.02	0.26±0.20	0.48±0.29
–	Q	N	–	S	0.23±0.02	0.37±0.05	0.35±0.04	0.01±0.01	0.25±0.19	0.49±0.28
–	Q	–	–	S	0.24±0.02	0.49±0.04	0.37±0.02	0.01±0.01	0.28±0.18	**0.50±0.23**
–	–	N	–	–	0.29±0.02	0.38±0.03	0.37±0.03	0.00±0.00	0.37±0.26	0.38±0.25
–	–	N	D	–	0.28±0.02	0.41±0.04	0.37±0.02	0.00±0.00	0.37±0.26	0.38±0.26
–	–	N	D	S	0.25±0.03	0.31±0.05	0.35±0.03	0.01±0.01	0.27±0.22	0.41±0.36
–	–	N	–	S	0.27±0.02	0.41±0.03	0.37±0.02	0.01±0.01	0.35±0.24	0.41±0.24
R	–	–	–	–	0.35±0.02	0.22±0.03	0.37±0.02	0.01±0.01	0.41±0.21	0.09±0.06
R	–	–	D	–	0.34±0.02	0.24±0.03	0.36±0.03	0.02±0.01	0.41±0.21	0.10±0.05
R	–	–	D	S	0.27±0.03	0.15±0.02	0.35±0.03	0.01±0.02	0.24±0.19	0.44±0.36
R	Q	–	–	–	0.23±0.02	0.51±0.04	0.36±0.03	0.02±0.02	0.32±0.18	0.38±0.20
R	Q	–	D	–	0.18±0.02	0.52±0.04	**0.34±0.03**	**0.04±0.02**	0.24±0.19	0.48±0.30
R	Q	–	D	S	0.17±0.02	0.57±0.05	0.34±0.03	0.03±0.03	0.24±0.20	0.47±0.28
R	Q	N	–	–	0.20±0.02	0.46±0.05	0.35±0.04	0.01±0.02	0.27±0.21	0.46±0.25
R	Q	N	D	–	0.17±0.02	0.55±0.05	0.35±0.04	0.02±0.02	0.27±0.22	0.45±0.31
R	Q	N	D	S	**0.16±0.02**	**0.59±0.07**	0.34±0.03	0.03±0.02	0.29±0.23	0.44±0.25
R	Q	N	–	S	0.19±0.03	0.51±0.04	0.35±0.03	0.02±0.02	0.27±0.21	0.45±0.29
R	Q	–	–	S	0.19±0.03	0.51±0.07	0.35±0.05	0.03±0.02	**0.23±0.17**	0.49±0.27
R	–	N	–	–	0.27±0.02	0.42±0.04	0.37±0.03	0.01±0.01	0.34±0.22	0.37±0.19
R	–	N	D	–	0.23±0.02	0.35±0.05	0.35±0.03	0.02±0.02	0.29±0.24	0.41±0.34
R	–	N	D	S	0.22±0.02	0.42±0.07	0.35±0.03	0.02±0.03	0.30±0.24	0.40±0.36
R	–	N	–	S	0.24±0.03	0.35±0.06	0.35±0.03	0.01±0.02	0.29±0.23	0.42±0.33
R	–	–	–	S	0.35±0.02	0.24±0.03	0.37±0.03	0.01±0.02	0.41±0.20	0.10±0.04
–	–	–	–	S	0.38±0.03	0.15±0.02	0.37±0.03	0.01±0.01	0.41±0.21	0.11±0.06

- For the most promising cases (usefulness and supportiveness), some common elements of the best-performing prompts are the presence of the system role, the question query, and the strictness instruction. The narrative and delimiter were effective for usefulness prediction, but seem dispensable for supportiveness prediction.
- In terms of system rankings, only a few prompt configurations led to RBO results comparable to those achieved for relevance estimation in [19].

Next, we selected the best-performing prompt in terms of RBO (i.e., QD) and used it to annotate all judged documents in all HM collections. This yielded RBOs of 0.72 (2019), 0.15 (2021), and 0.86 (2022). The ranking similarities in 2019 and 2022 are similar to those reported by previous studies [19], but the 2021 correlations are low. This effect may be attributed to the greater difficulty of the 2021 search task, as reflected in the overall lower performance of the participants.

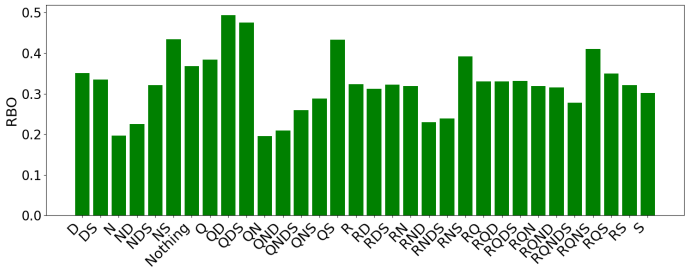

Fig. 2: Consistency on system rankings (Rank-biased Overlap, human vs LLM)

We also conducted an initial analysis of failure cases. For example, for the question "*Can crystals heal?*" from TREC HM 2022, the LLM classified as supportive a document stating that "*Crystal healing is the use of crystals or stones in the promotion of physical, mental, or emotional healing ... due to the high levels of vibrational energy in these crystals*", whereas human annotators labeled the same document as neutral regarding the treatment.

Discrepancies were also observed in credibility assessments. For instance, a 2019 topic asked whether an ankle brace can help heal Achilles tendinitis. For one document, human evaluators assigned a low credibility label, while the model rated it as highly credible. Upon inspection, the document appears to be an online sales website promoting a specific ankle brace, emphasizing its quality and its benefits for recovery from conditions such as tendinitis. It therefore appears that, in some cases, models and humans do not align in how they associate commercial or marketing-related elements with the credibility of content.

Finally, we identified cases in which the disagreement was clearly attributable to an incorrect human label. For the question "*Can cancer be inherited?*", a document stating that "*multifactorial inherited diseases ... include conditions such as diabetes, high blood pressure, and even cancer*" was labeled as unsupportive by the human annotator, despite it clearly supports the claim. In this case, the LLM's judgment aligns more closely with the document's content.

5 Limitations & Future Work

This research represents a preliminary step toward understanding the role that LLMs can play in labeling challenging criteria such as the credibility of search results. Our study should be extended, for instance, by experimenting with a broader range of LLMs from different families and sizes. We made also preliminary experiments with Llama3[4] finding similar trends to those reported here (but lower RBOs, namely 0.49 –2019–, 0.19 –2021–, and 0.51 –2022–). In the future, it will also be necessary to evaluate alternative instructions to be incorporated into

[4] We used 8B instruct version quantized to 4-bits.

the prompts, since it is well known that the phrasing used in the input expressions plays a crucial role. Future research could also focus on examining few-shot or fine-tuning methods for the automatic prediction of misinformation-related variables.

It is also necessary to adopt a critical stance on the role that LLMs should play in automatic annotation. In the literature, issues such as circularity, leniency, "narcissism" (favoring LLM-based reranking systems), or inability to identify the best-performing systems have been discussed as concerning aspects in the automatic annotation with LLMs [17,6,3]. These challenges must also be examined in the near future for annotation tasks such as those conducted in our study. Regarding leniency, we made preliminary analysis of the LLM annotations and observed that the LLM was lenient on supportiveness estimates (emitting many more 2s than those present in the human labels) but, in contrast, it was conservative on usefulness and credibility.

In some evaluation campaigns (e.g., the IR track of CLEF eHealth [14]), other aspects, such as trustworthiness and readability, have been considered. In the near future, it will be interesting to investigate how LLMs can address these criteria, which are also closely related to information reliability.

It is also worth noting that hybrid approaches, which combine a small set of human annotations with those generated by LLMs –as in the relevance annotation framework proposed by [18]–, could also be considered in the context of misinformation annotation.

6 Conclusions

Our study highlights the limitations of LLMs for annotating variables that are crucial in misinformation benchmarks. On the one hand, LLM-generated credibility judgments diverged substantially from human assessments. On the other hand, predictions of usefulness and supportiveness yielded results comparable to those reported in the literature of automatic relevance annotation with LLMs. Nevertheless, further research is needed to investigate the impact of circularity, biases, and the ability of synthetic assessments to reliably identify the best-performing systems.

Acknowledgments. The authors thank the financial support from the Agencia Estatal de Investigación (Spain) (PID2022-137061OB-C22 funded by MICIU/AEI/10.13039/501100011033), the Xunta de Galicia - Conselleria de Educación, Ciencia, Universidades e Formación Profesional (Centro de investigación de Galicia acreditación 2024-2027 ED431G-2023/04 and Reference Competitive Group accreditation ED431C 2022/19) and the European Union (European Regional Development Fund - ERDF). This research is also supported by the the project Cátedra de IA aplicada a la Medicina Personalizada de Precisión (Cátedras ENIA, TSI-100932-2023-3); Cátedras ENIA is funded by the Ministerio de Transformación Digital y Función Pública (Secretaría de Estado de Digitalización e Inteligencia Artificial); and by the NextGeneration EU-fund.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abbasiantaeb, Z., Meng, C., Azzopardi, L., Aliannejadi, M.: Can We Use Large Language Models to Fill Relevance Judgment Holes? In: 1st Workshop on Evaluation Methodologies,Testbeds and Community for Information Access Research (EMTCIR 2024) (2024)
2. Abualsaud, M., Lioma, C., Maistro, M., Smucker, M.D., Zuccon, G.: Overview of the TREC 2019 decision track (2019)
3. Alaofi, M., Thomas, P., Scholer, F., Sanderson, M.: 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. pp. 32–41. Association for Computing Machinery (2024)
4. Arabzadeh, N., Clarke, C.L.A.: Benchmarking LLM-based Relevance Judgment Methods. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3194–3204. ACM, Padua Italy (2025)
5. Balog, K., Metzler, D., Qin, Z.: 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. pp. 3865–3875. ACM, Padua Italy (2025)
6. Clarke, C., Dietz, L.: LLM-based relevance assessment still can't replace human relevance assessment. In: Proceedings of the 10th International Workshop on Evaluating Information Access (2025)
7. Clarke, C., Maistro, M., Seifikar, M., Smucker, M.: Overview of the TREC 2022 health misinformation track (2022)
8. Clarke, C., Maistro, M., Smucker, M.: Overview of the TREC 2021 health misinformation track (2021)
9. Dewan, M., Liu, J., Shah, C.: LLM-Driven Usefulness Labeling for IR Evaluation. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3055–3059. ACM, Padua Italy (2025)
10. Faggioli, G., Dietz, L., Clarke, C., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., Wachsmuth, H.: Perspectives on Large Language Models for Relevance Judgment. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval. pp. 39–50 (2023)
11. Farzi, N., Dietz, L.: Criteria-Based LLM Relevance Judgments. In: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR). pp. 254–263. Association for Computing Machinery (2025)
12. Farzi, N., Dietz, L.: Does UMBRELA work on other LLMs? In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3214–3222. ACM, Padua Italy (2025)
13. Fernández-Pichel, M., Meyer, S., Bink, M., Frummet, A., Losada, D.E., Elsweiler, D.: Improving the reliability of health information credibility assessments. In: Proceedings of the 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2023). pp. 43–50 (2023)

14. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF ehealth evaluation lab 2016 (2016)
15. MacAvaney, S., Soldaini, L.: One-Shot Labeling for Automatic Relevance Estimation. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2230–2235 (2023)
16. Rahmani, H.A., Yilmaz, E., Craswell, N., Mitra, B., Thomas, P., Clarke, C.L.A., Aliannejadi, M., Siro, C., Faggioli, G.: LLMJudge: LLMs for Relevance Judgments. In: 1st Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) (2024)
17. Soboroff, I.: Don't Use LLMs to Make Relevance Judgments. Information Retrieval Research **1**(1), 29–46 (2025)
18. Takehi, R., Voorhees, E.M., Sakai, T., Soboroff, I.: LLM-Assisted Relevance Assessments: When Should We Ask LLMs for Help? In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 95–105. ACM, Padua Italy (2025)
19. Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1930—1940. Association for Computing Machinery (2024)
20. Upadhyay, R., Viviani, M.: Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy. Discover Computing **28**(1), 27 (2025). https://doi.org/10.1007/s10791-025-09505-5, https://link.springer.com/10.1007/s10791-025-09505-5
21. Upadhyay, S., Kamalloo, E., Lin, J.: LLMs Can Patch Up Missing Relevance Judgments in Evaluation (2024), arXiv:2405.04727 [cs]
22. Upadhyay, S., Pradeep, R., Thakur, N., Campos, D., Craswell, N., Soboroff, I., Dang, H.T., Lin, J.: A Large-Scale Study of Relevance Assessments with Large Language Models: An Initial Look (2024), arXiv:2411.08275 [cs]
23. Upadhyay, S., Pradeep, R., Thakur, N., Craswell, N., Lin, J.: UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor (2024), arXiv:2406.06519 [cs]
24. Zhang, H., Zhang, R., Guo, J., De Rijke, M., Fan, Y., Cheng, X.: Are Large Language Models Good at Utility Judgments? In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1941–1951. ACM, Washington DC USA (2024)