LLMJudge: LLMs for Relevance Judgments

Hossein A. Rahmani[1], Emine Yilmaz[1], Nick Craswell[2], Bhaskar Mitra[3], Paul Thomas[4], Charles L. A. Clarke[5], Mohammad Aliannejadi[6], Clemencia Siro[6] and Guglielmo Faggioli[7]

[1]University College London, London, UK
[2]Microsoft, Seattle, US
[3]Microsoft, Montréal, Canada
[4]Microsoft, Adelaide, Australia
[5]University of Waterloo, Ontario, Canada
[6]University of Amsterdam, Amsterdam, The Netherlands
[7]University of Padua, Padua, Italy

1. Introduction

The LLMJudge challenge[1] is organized as part of the LLM4Eval[2] workshop [1] at SIGIR 2024. Test collections are essential for evaluating information retrieval (IR) systems. The evaluation and tuning of a search system is largely based on relevance labels, which indicate whether a document is useful for a specific search and user. However, collecting relevance judgments on a large scale is costly and resource-intensive. Consequently, typical experiments rely on third-party labelers who may not always produce accurate annotations. The LLMJudge challenge aims to explore an alternative approach by using LLMs to generate relevance judgments. Recent studies have shown that LLMs can generate reliable relevance judgments for search systems. However, it remains unclear which LLMs can match the accuracy of human labelers, which prompts are most effective, how fine-tuned open-source LLMs compare to closed-source LLMs like GPT-4, whether there are biases in synthetically generated data, and if data leakage affects the quality of generated labels. This challenge will investigate these questions, and the collected data will be released as a package to support automatic relevance judgment research in information retrieval and search.

2. Related Work

Automatic relevance judgment has recently received significant attention in the Information Retrieval (IR) community. In earlier studies, Faggioli et al. [2] studied different levels of human and LLMs collaboration for automatic relevance judgement. They suggested the need for humans to support and collaborate with LLMs for a human-machine collaboration judgment. Thomas et al. [3] leverage LLMs capabilities in judgement at scale, in Microsoft Bing. They used real searcher feedback to consider an LLM and prompt in a way that matches the small sample of searcher preferences. Their experiments show that LLMs can be as good as human annotators in indicating the best systems. They also comprehensively investigated various prompts and prompt features for the task and revealed that LLM performance on judgments can varies with simple paraphrases of prompts. Recently, Rahmani et al. [4] have studied fully synthetic test collection using LLMs. In their study, they not only generated synthetic queries but also synthetic judgment to build a full synthetic test collation for retrieval evaluation. They have shown that LLMs are able to generate a synthetic test collection that results in system ordering performance similar to evaluation results obtained using the real test collection.

LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval, 18 July 2024, Washington DC, United States

[1]https://llm4eval.github.io/challenge/
[2]https://llm4eval.github.io/

Table 1
Statistics of LLMJudge Dataset

	Dev	Test
# queries	25	25
# passage	7,224	4,414
# qrels	7,263	4,423
# irrelevant (0)	4,538	2,005
# related (1)	1,403	1,233
# highly relevant (2)	625	808
# perfectly relevant (3)	697	377

3. LLMJudge Task Design

The challenge will be, given the query and document as input, how they are relevant. Here, we use four-point scale judgments to evaluate the relevance of the query to document as follows:

- **[3] Perfectly relevant**: The passage is dedicated to the query and contains the exact answer.
- **[2] Highly relevant**: The passage has some answers for the query, but the answer may be a bit unclear, or hidden amongst extraneous information.
- **[1] Related**: The passage seems related to the query but does not answer it.
- **[0] Irrelevant**: The passage has nothing to do with the query.

The task is, by providing the datasets that include queries, documents, and query-document files to participants, to ask LLMs to generate a score [0, 1, 2, 3] indicating the relevance of the query to the document.

4. LLMJudge Data

The LLMJudge challenge dataset is built upon the passage retrieval task dataset of the TREC 2023 Deep Learning track[3] (TREC-DL 2023) [5]. Table 1 shows the statistics of the LLMJudge challenge datasets. We divide the data into development and test sets. The test set is used for the generation of judgment by participants, while the development set could be used for few-shot or fine-tuning purposes. The datasets, sample prompt, and the quick starter for automatic judgment can be found at the following repository: https://github.com/llm4eval/LLMJudge

5. Evaluation

Participants' results will then be evaluated in two methods after submission:

- automated evaluation metrics on human labels in the test set hidden from the participants;
- system ordering evaluation of multiple search systems on human judgments and LLM-based judgments

6. Submissions and Results

In order to evaluate the quality of the generated labels, we used Cohen's κ to see the labeler's agreement with LLMJudge test data at query-document level and the Kendall's κ to check the labeler's agreement with LLMJudge test data on system ordering, i.e., the runs that submitted to TREC DL 2023. In total, we had 39 submissions (i.e., the 39 labelers) from 7 groups from National Institute of Standards and

[3] https://microsoft.github.io/msmarco/TREC-Deep-Learning.html

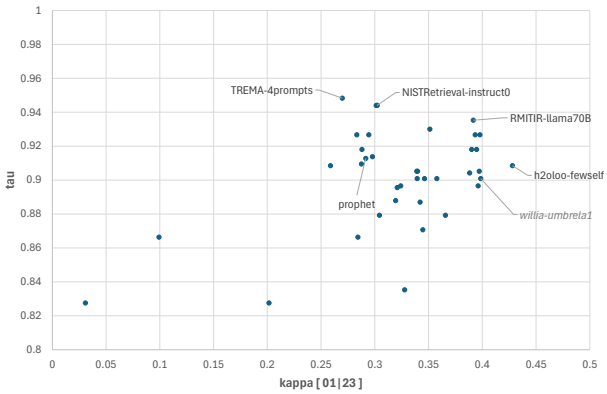

Figure 1: Scatter plot of Cohen's κ and Kendall's τ for submitted labelers

Technology (NIST), RMIT University, The University of Melbourne, University of New Hampshire, University of Waterloo, Included Health, and University of Amsterdam.

Figure 1 shows the performance of submitted labelers on the `LLMJudge` test set. The x-axis represents Cohen's κ, and the y-axis shows the labelers' agreement on system ordering. Labelers exhibit low variability in Kendall's τ but greater variability in Cohen's κ. Most labelers cluster within a narrow range of τ values, indicating consistent system rankings but more variation in inter-rater reliability, as measured by Cohen's κ. This suggests that while labelers generally agree on rankings, their exact labels are less consistent, leading to the observed variability in κ.

Acknowledgment

The challenge is organized as a joint effort by the University College London, Microsoft, the University of Amsterdam, the University of Waterloo, and the University of Padua. The views expressed in the content are solely those of the authors and do not necessarily reflect the views or endorsements of their employers and/or sponsors. This work is supported by the Engineering and Physical Sciences Research Council [EP/S021566/1], the EPSRC Fellowship titled "Task Based Information Retrieval" [EP/P024289/1], CAMEO, PRIN 2022 n. 2022ZLL7MW.

References

[1] H. A. Rahmani, C. Siro, M. Aliannejadi, N. Craswell, C. L. A. Clarke, G. Faggioli, B. Mitra, P. Thomas, E. Yilmaz, Llm4eval: Large language model for evaluation in ir, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 3040–3043. URL: https://doi.org/10.1145/3626772.3657992. doi:10.1145/3626772.3657992.

[2] G. Faggioli, L. Dietz, C. L. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, et al., Perspectives on large language models for relevance judgment, in: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, 2023, pp. 39–50.

[3] P. Thomas, S. Spielman, N. Craswell, B. Mitra, Large language models can accurately predict searcher preferences, arXiv preprint arXiv:2309.10621 (2023).

[4] H. A. Rahmani, N. Craswell, E. Yilmaz, B. Mitra, D. Campos, Synthetic test collections for retrieval evaluation, arXiv preprint arXiv:2405.07767 (2024).

[5] N. Craswell, B. Mitra, E. Yilmaz, H. A. Rahmani, D. Campos, J. Lin, E. M. Voorhees, I. Soboroff, Overview of the trec 2023 deep learning track, in: Text REtrieval Conference (TREC), NIST, TREC, 2024. URL: https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2023-deep-learning-track/.