Re-Rankers are Effective Relevance Judgment Predictors

Chuan Meng
University of Amsterdam
Amsterdam, The Netherlands
c.meng@uva.nl

Jiqun Liu
University of Oklahoma
Norman, United States
jiqunliu@ou.edu

Mohammad Aliannejadi
University of Amsterdam
Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Fengran Mo
Université de Montréal
Montréal, Canada
fengran.mo@umontreal.ca

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Abstract

Using large language models (LLMs) to predict relevance judgments has shown promising results. Prior studies treat LLM-based relevance judgment prediction as a distinct research line: either prompting an LLM to directly generate relevance judgments, or fine-tuning an LLM on human-labelled judgments to improve its prediction capability. However, predicting relevance judgments is essentially a form of *relevance prediction*, a problem extensively studied in tasks such as retrieval and re-ranking. Despite this potential overlap, existing studies have not explored reusing or adapting established best practices for relevance prediction (e.g., a re-ranker) to predict relevance judgments, resulting in potential resource waste and redundant development. In this paper, we examine adapting re-rankers to function as relevance judgment predictors. We propose two adaptation strategies: (1) using binary tokens (e.g., "true" and "false") output by a re-ranker as direct relevance judgments; and (2) converting continuous re-ranking scores into binary labels by applying a threshold. Experimental results show that re-rankers can be adapted into effective relevance judgment predictors, although they still fall short of UMBRELA, a state-of-the-art LLM-based method using a carefully designed prompt.

CCS Concepts

• **Information systems** → **Retrieval models and ranking**; **Evaluation of retrieval results**.

Keywords

Relevance prediction, Re-ranking, Relevance judgment

ACM Reference Format:
Chuan Meng, Jiqun Liu, Mohammad Aliannejadi, Fengran Mo, and Maarten de Rijke. 2025. Re-Rankers are Effective Relevance Judgment Predictors. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2024, Padua, Italy*. ACM, New York, NY, USA, 4 pages.

SIGIR '25, July 13–18, 2024, Padua, Italy

1 Introduction

Relevance judgments, which map each query to the documents that should be retrieved for it [16], play a critical role in information retrieval (IR). Accurate relevance judgments are essential for both training and evaluating ranking systems [15]. However, the manual annotation of relevance judgments is labour-intensive [2]. Recently, the IR community has witnessed a surge in the use of large language models (LLMs) for automatically predicting relevance judgments [8, 11, 17, 20], which has shown promising results [17]. Existing studies on using LLMs to predict relevance judgments generally follow two main approaches. First, the majority of studies design prompts to generate relevance judgments directly from an LLM [2, 8, 15, 17, 19]. Second, some studies explore fine-tuning open-source LLMs on human-labelled relevance judgments to further improve performance [1, 12].

Recent literature tends to treat LLM-based relevance judgment prediction as a distinct line of research [8]. However, we argue that this task can essentially be viewed as a specific instance of the broader problem of *relevance prediction* [12]. In this work, we adopt a broad definition of relevance prediction, which refers to estimating how relevant a document is to a given query; the relevance can be represented either as discrete relevance labels or as continuous relevance scores. From this perspective, both relevance judgment prediction and text ranking (e.g., retrieval and re-ranking), an extensively studied task in IR, fall under the umbrella of relevance prediction, but they differ in how relevance is represented. Specifically, relevance judgment prediction aims to assign a discrete label (e.g., relevant or irrelevant) to a query–document pair, while text ranking typically produces a continuous relevance score, with the final ranking generated by sorting documents according to these scores. Despite the conceptual overlap, limited research has explored reusing or adapting well-established text ranking methods (e.g., re-rankers [10, 13]) for predicting relevance judgments. This research gap has led to potential inefficiencies, e.g., duplicated effort and underutilisation of existing resources.

In this paper, we examine adapting text ranking methods to function as relevance judgment predictors. Because relevance judgments are often used to evaluate or train other ranking models, the accuracy of relevance judgments should be prioritsed. Amongst various text-ranking methods, re-rankers, particularly those based on LLMs, have shown strong performance in estimating query–document relevance [10, 13]. Therefore, in this work, we focus

on adapting LLM-based re-ranking methods for predicting relevance judgments. We propose two adaptation strategies. First, for all re-rankers, we convert continuous re-ranking scores into binary labels by applying a threshold. Second, for a re-ranker that use large language models (LLMs)' output logits of special tokens [13] (e.g., "true" and "false") to compute relevance scores, we directly use the final outputted token by the re-ranker as direct relevance judgments.

We experiment with two widely-used re-rankers, monoT5 [13] and RankLLaMA [10], on the TREC 2019–2022 deep learning (TREC-DL) tracks [3–7]. Experimental results show that re-rankers can be adapted into effective relevance judgment predictors, although they still fall short of UMBRELA [20], a state-of-the-art LLM-based method using carefully designed prompts [20]. In particular, the best-performing re-rankers achieve Cohen's κ scores within 0.1 of UMBRELA on most datasets.

Our main contributions are as follows:

- To the best of our knowledge, we are the first to explicitly reuse re-rankers as relevance judgment predictors.
- We propose two adaptation strategies to enable re-rankers to function as relevance judgment predictors.
- Experimental results demonstrate that re-rankers can serve as effective predictors of relevance judgments.

2 Methodology

In this section, we describe the re-rankers we include in this work and their adaptation methods, along with the experimental setup.

2.1 Re-rankers

We use two widely-used LLM-based re-rankers, namely monoT5 [13] and RankLLaMA [10]. Both are pointwise re-rankers: given a query and a candidate document, they independently assign a relevance score, and the final ranking is produced by sorting documents according to these scores. Both re-rankers are trained on the training set of MS MARCO V1.

Given a query and a document, monoT5 [13] fine-tunes T5 [14] to produce one of two special tokens, "true" or "false", depending on whether the document is relevant to the query. During inference, monoT5 applies a softmax over the logits of the "true" and "false" tokens, and uses the probability assigned to the "true" token as the relevance score.

RankLLaMA [10] fine-tunes Llama 2 [18] using LoRA [9] to directly project the representation of the end-of-sequence token to a relevance score.

We also investigate the impact of re-ranker scaling on the performance of relevance judgment prediction. For monoT5, we use the base (220M)[1], large (770M)[2], and 3B[3] variants. For RankLLaMA, we adopt the 7B[4] and 13B[5] versions. In total, our experiments cover five distinct re-rankers with varying model architectures and sizes.

2.2 Adaptation methods

We propose two strategies for adapting re-rankers to produce binary relevance judgments: *score thresholding* and *direct generation*.

[1] https://huggingface.co/castorini/monot5-base-msmarco
[2] https://huggingface.co/castorini/monot5-large-msmarco
[3] https://huggingface.co/castorini/monot5-3b-msmarco
[4] https://huggingface.co/castorini/rankllama-v1-7b-lora-passage
[5] https://huggingface.co/castorini/rankllama-v1-13b-lora-passage

Score thresholding applies to all re-rankers. We convert their continuous relevance scores into binary labels by applying a predefined threshold: a document is predicted as relevant if its score is greater than or equal to the threshold, and irrelevant otherwise.

Direct generation is specific to monoT5. Instead of using the numerical relevance score, we directly map the model's generated special token to a relevance label: if monoT5 generates the token "true", the document is labelled as "relevant"; if it generates "false", it is labelled as irrelevant.

Extending these strategies to support graded relevance labels is left for future work.

2.3 Datasets

We follow recent studies [15, 20] to use the TREC 2019–2023 Deep Learning (TREC-DL) tracks [3–7]. See Table 1 for summary statistics. TREC-DL 19–20 and TREC-DL 21–23 are based on the MS MARCO V1 and MS MARCO V2 passage ranking collections respectively. In the V1 edition, the corpus comprises 8.8 million passages while the V2 edition has over 138 million passages. The relevance judgments in the five datasets include five scales: perfectly relevant (3), highly relevant (2), related (1), and irrelevant (0). In this work, we only consider binary relevance judgments and follow prior work [3–7] to use relevance scale ≥ 2 as positive.

All relevance judgments in TREC-DL 19—21 are annotated by assessors from the U.S. National Institute of Standards and Technology (NIST). In addition to human-annotated relevance judgments, TREC-DL 22 and 23 also include automatically propagated labels, where relevance labels from human-annotated passages are assigned to their near-duplicate counterparts within the corpus. Following Upadhyay et al. [20], we remove these automatically propagated judgments in TREC-DL 22 and 23, and retain only the human-annotated ones.

Table 1: Statistics of the TREC 2019–2023 Deep Learning (TREC-DL) tracks [3–7]. Note that for TREC-DL 22 and 23, we follow Upadhyay et al. [20] in using relevance judgments after removing automatically propagated passages.

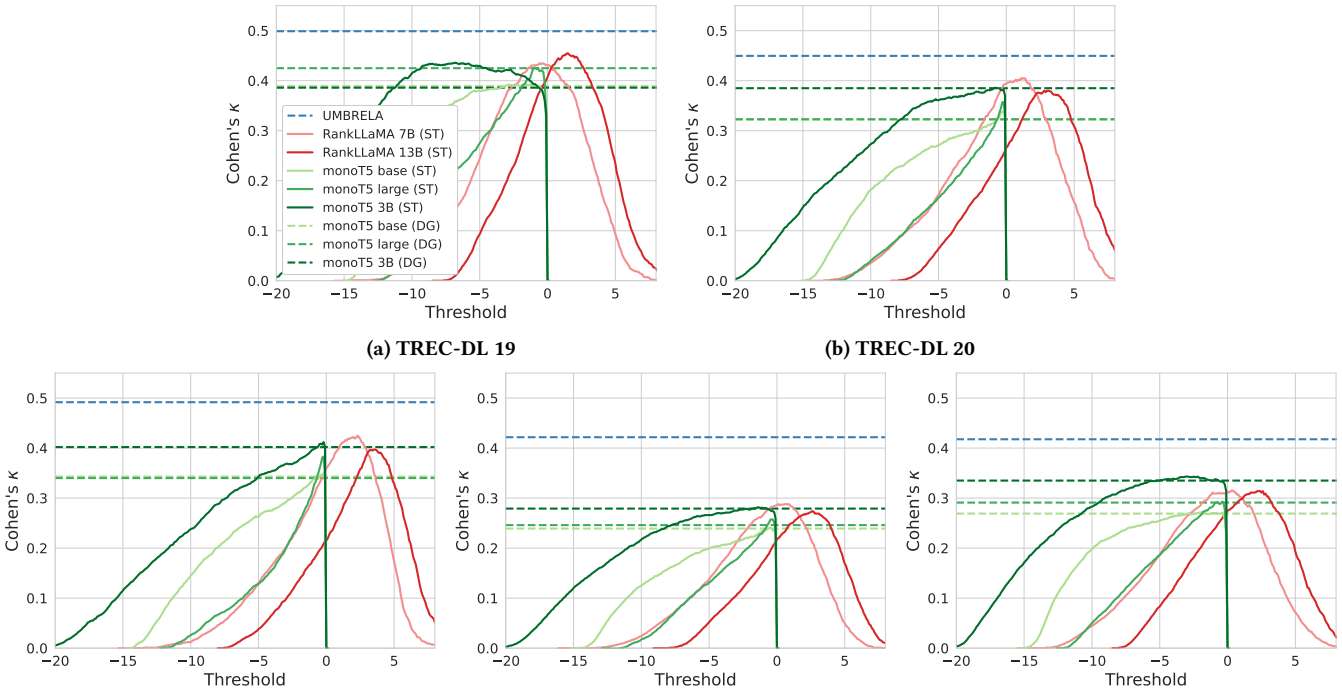
Track	# Runs	# Queries	Relevance labels (0/1/2/3)
TREC-DL 2019	36	43	5,158 / 1,601 / 1,804 / 697
TREC-DL 2020	59	54	7,780 / 1,940/ 1,020 / 646
TREC-DL 2021	62	53	4,338 / 3,063 / 2,341 / 1,086
TREC-DL 2022	60	76	12,892 / 6,192 / 3,053 / 1,385
TREC-DL 2023	35	82	11,618 / 3,774 / 1,942 / 1,544

2.4 Evaluation metric

To evaluate the performance of relevance judgment prediction, we follow prior work [8, 12, 15] and compute Cohen's κ, which measures the agreement between human-annotated relevance judgments (a.k.a., qrels) and judgments predicted by a re-ranker.

2.5 Baseline

We use a state-of-the-art relevance judgment prediction model, UMBRELA [20]. Given a query and a document, UMBRELA prompts



(a) TREC-DL 19 (b) TREC-DL 20

(c) TREC-DL 21 (d) TREC-DL 22 (e) TREC-DL 23

Figure 1: Relevance judgment agreement (Cohen's κ) between human assessors and each re-ranker, across relevance score thresholds. ST and DG stand for score thresholding and direct generation, respectively. Note that monoT5's relevance score is defined as the probability of the "true" in its original paper [13]; however, we follow the common implementation used in existing monoT5 codebases (e.g., PyTerrier), where the log probability of "true" is used as the relevance score.

GPT-4o with a zero-shot descriptive, narrative, and aspects (DNA) prompting technique [17].

3 Results

We present the results of relevance judgment prediction in Figure 1, covering monoT5 re-rankers (base, large, and 3B) and RankLLaMA re-rankers (7B and 13B). All re-rankers are evaluated using the score thresholding adaptation strategy, and monoT5 is additionally evaluated with the direct generation strategy. For comparison, we also include UMBRELA [20], a state-of-the-art LLM-based relevance judgment predictor, as a reference method. We have three main observations.

First, under the score thresholding adaptation strategy, all re-rankers can achieve at least fair agreement (Cohen's $\kappa > 0.21$), and in some cases moderate agreement (Cohen's $\kappa > 0.41$), with human assessors, when the threshold is set appropriately. Surprisingly, the gap in Cohen's κ between the best-performing re-ranker and UMBRELA is less than 0.1 on all datasets except TREC-DL 22. For example, on TREC-DL 19, RankLLaMA 13B reaches a κ score above 0.430 with a threshold around 2, compared to 0.499 for UMBRELA. Comparing monoT5 and RankLLaMA under this strategy, there is no clear winner.

Second, monoT5 with the direct generation strategy can also achieve fair or even moderate agreement with human assessors, without requiring threshold tuning. We observe that the κ gap between monoT5 (3B) and UMBRELA is below 0.1 on all datasets except TREC-DL 19 and 22. For instance, on TREC-DL 19, monoT5 large achieves a κ score of 0.425, compared to 0.499 for UMBRELA; on TREC-DL 2020, monoT5 3B achieves 0.385, compared to UMBRELA's 0.450. Overall, the best-performing monoT5 variant with the direct generation strategy achieves performance comparable to that of the best monoT5 and RankLLaMA models using score thresholding. However, in some cases, threshold tuning enables slightly higher performance.

Third, the impact of model scaling is mixed. For RankLLaMA, the 13B model shows improvement over the 7B model only on TREC-DL 19, with no consistent gains across other datasets. In contrast, monoT5 exhibits a clearer positive scaling trend: under both adaptation strategies, larger variants generally perform better, with the 3B model outperforming the base and large versions in most cases.

4 Conclusions & discussion

In this work, we have explored reusing LLM-based re-rankers as relevance judgment predictors. We have used two representative re-rankers, monoT5 and RankLLaMA, across multiple model sizes, and have proposed two adaptation strategies to enable them to function as relevance judgment predictors: score thresholding and direct token generation. Our experiments on TREC-DL 19–23 have shown that although re-rankers adapted with these strategies fall short of UMBRELA (GPT-4o), a state-of-the-art LLM-based method using carefully designed prompts, the performance gap is relatively

small. In particular, the best-performing re-rankers achieve Cohen's κ scores within 0.1 of UMBRELA on most datasets. This suggests that re-rankers can be adapted into effective relevance judgment predictors.

These findings suggest that relevance judgment prediction is not a fundamentally new task, but rather a specific case of *relevance prediction*. Relevance judgment prediction is closely related to re-ranking, which is already a well-established and widely used form of relevance prediction. This connection indicates that our community can benefit from building relevance judgment predictors *on top of* well-established relevance prediction methods, rather than designing new relevance judgment prediction models from scratch. By reusing existing models, we can reduce duplicated effort and avoid underutilisation of existing resources.

We identify the limitations of this work and outline directions for future research. First, we only used Cohen's κ to evaluate the agreement between predicted and human relevance judgments. In future work, we plan to measure how well predicted labels preserve system ranking, for example by computing the correlation between system rankings based on human and predicted labels. Second, we focus only on binary relevance judgments. As a next step, we aim to explore new adaptation strategies that enable re-rankers to output graded relevance labels.

References

[1] Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. 2025. Improving the Reusability of Conversational Search Test Collections. In *ECIR*. 196–213.

[2] 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 *SIGIR-AP*. 32–41.

[3] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2020 Deep Learning Track. In *TREC 2020*.

[4] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2019. Overview of the TREC 2019 Deep Learning Track. In *REC 2019*.

[5] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Fernando Campos, and Jimmy Lin. 2021. Overview of the TREC 2021 Deep Learning Track. In *TREC 2021*.

[6] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Fernando Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2022. Overview of the TREC 2022 Deep Learning Track. In *TREC 2022*.

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Hossein A. Rahmani, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2023. Overview of the TREC 2023 Deep Learning Track. In *TREC 2023*.

[8] Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on Large Language Models for Relevance Judgment. In *ICTIR*. 39–50.

[9] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

[10] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. In *SIGIR*. 2421–2425.

[11] Sean MacAvaney and Luca Soldaini. 2023. One-Shot Labeling for Automatic Relevance Estimation. In *SIGIR*. 2230–2235.

[12] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2025. Query Performance Prediction using Relevance Judgments Generated by Large Language Models. *TOIS* (2025).

[13] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *EMNLP*. 708–718.

[14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.

[15] Julian A Schnabel, Johanne R Trippas, Falk Scholer, and Danula Hettiachchi. 2025. Multi-stage Large Language Model Pipelines Can Outperform GPT-4o in Relevance Assessment. *arXiv preprint arXiv:2501.14296* (2025).

[16] Ian Soboroff. 2025. Don't Use LLMs to Make Relevance Judgments. *Information Retrieval Research* 1, 1 (2025), 29–46.

[17] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models Can Accurately Predict Searcher Preferences. In *SIGIR*. 1930–1940.

[18] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023).

[19] Shivani Upadhyay, Ehsan Kamalloo, and Jimmy Lin. 2024. LLMs Can Patch Up Missing Relevance Judgments in Evaluation. *arXiv preprint arXiv:2405.04727* (2024).

[20] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. *arXiv preprint arXiv:2406.06519* (2024).