Can We Use Large Language Models to Fill Relevance Judgment Holes?

Zahra Abbasiantaeb[1], Chuan Meng[1], Leif Azzopardi[2] and Mohammad Aliannejadi[1]

[1]University of Amsterdam, Amsterdam, The Netherlands
[2]University of Strathclyde, Glasgow, Scotland, UK

Abstract

Incomplete relevance judgments limit the reusability of test collections. When new systems are compared against previous systems used to build the pool of judged documents, they often do so at a disadvantage due to the "*holes*" in test collection (i.e., pockets of un-assessed documents returned by new systems). In this paper, we aim to extend test collections by employing Large Language Models (LLM) to fill these holes by leveraging and grounding existing judgments. We explore this problem in the context of *Conversational Search (CS)* using TExt Retrieval Conference (TREC) Interactive Knowledge Assistance Track (iKAT) collection, where information needs are highly dynamic and the responses (and, the results retrieved) are much more varied (leaving bigger holes). While previous work has shown that automatic judgments from LLMs result in highly correlated rankings, we find it substantially lower correlates when human plus automatic judgments are used (regardless of LLM, one/two/few shot, or fine-tuned). We further find that, depending on the LLM employed, new runs will be highly favored (or penalized), and this effect is magnified proportionally to the size of the holes. Instead, one should generate the LLM annotations on the whole document pool to achieve more consistent rankings with human-generated labels. Further work is needed to prompt engineer and fine-tune LLMs to align with human judgment, thereby, improving the methods' accuracy.

Keywords

Conversational search, Large language models, Judgments

1. Introduction

Building reusable test collections in a cost-efficient manner to evaluate current and future systems has been a long-standing challenge in the field of information retrieval (IR) [1]. The predominant strategy for creating such collections has been through the use of pooling [2, 3] – where a subset of documents, taken from various systems, is assessed for relevance. This is a compromise away from the "ideal test collection" with complete relevance assessments which is infeasible and impractical. While the pooling strategy is fairly robust [4, 5, 6, 7], it leads to various evaluation biases (e.g., [8, 5]) where systems that did not contribute to the pool, can be significantly disadvantaged. This is because documents that have not been judged are considered irrelevant. Therefore, the fewer judged/assessed documents returned in a ranking, the lower the retrieval performance ceiling [9]. However, the fewer the judgments required to compare systems the cheaper the test collection.

While researchers have tried to address these trade-offs in various ways, either by proposing new metrics and methodologies for compensating for the un-assessed documents (e.g., [5, 10, 9]) and/or developing new pooling and judgment strategies (e.g., [11, 12, 13]), "holes" in the pools still remain [14, 15].

EMTCIR '24: The First Workshop on Evaluation Methodologies, Testbeds and Community for Information Access Research, December 12, 2024, Tokyo, Japan
*Corresponding author.
z.abbasiantaeb@uva.nl (Z. Abbasiantaeb); c.meng@uva.nl (C. Meng); leif.azzopardi@strath.ac.uk (L. Azzopardi); m.aliannejadi@uva.nl (M. Aliannejadi)
https://zahraabbasiantaeb.github.io (Z. Abbasiantaeb); https://chuanmeng.github.io (C. Meng); https://www.strath.ac.uk/staff/azzopardileifdr/ (L. Azzopardi); https://aliannejadi.com (M. Aliannejadi)
0000-0002-4046-3419 (Z. Abbasiantaeb); 0000-0002-1434-7596 (C. Meng); 0000-0002-6900-0557 (L. Azzopardi); 0000-0002-9447-4172 (M. Aliannejadi)

However, with the advances in the development of powerful Large Language Models (LLMs) and other neural-based models, new opportunities arise for building scalable, robust, and reusable test collections at a lower cost. LLMs offers the possibility to: (1) assess large volumes of documents reasonably cheaply, especially compared to human judgments, (2) do so in a consistent and independent but potentially biased manner, (3) if the LLM and prompt are fixed and shared, then judgments can be collected at different times under the same conditions, and, (4) typically at a higher quality than "typical" crowd workers.

Indeed, recent studies [16, 17, 18, 19] have shown the effectiveness of using LLMs to automatically generate relevance judgments in the scenario of ad-hoc search. These works demonstrate that LLM-based judgments exhibit a high correlation with human judgments. Thomas et al. [18], Faggioli et al. [19] prompted commercial LLMs (e.g., GPT-3.5, GPT-3.5/4) to generate relevance judgments. However, commercial LLMs come with limitations like non-reproducibility, non-deterministic outputs, and potential data leakage between pre-training and evaluation data, impeding their utility in scientific research [20]. MacAvaney and Soldaini [15] and Khramtsova et al. [17] prompted an open-source LLMs, Flan-T5 [21], for generating relevance judgments. While open-source LLMs are less effective, they do offer the potential for the development of reproducible and reusable test collections at scale. This led to efforts by Meng et al. [16], who fine-tuned an open-source LLM, Llama [22] using parameter-efficient fine-tuning (PEFT) [23] to better condition the LLM for performing the task of assigning relevance judgments. They found that more complete test collections could be produced, with high quality, at a lower cost.

While this prospect is very appealing, it is fraught with new, unexplored challenges. Of interest, in this work is the notion of grounding. Training systems on the judgments of LLMs, and then evaluating those systems on subsequent test collections, based on judgments from LLMs, creates a potentially dangerous cycle that may amplify and re-enforce existing biases inherent in LLMs. Grounding the LLMs based judgment given human judgments provides a mechanism to condition the LLMs to be more aligned with human annotators, reducing the risk of AI falling into a negative feedback loop [24]. To this end, MacAvaney and Soldaini [15] focus on a setting wherein the LLM is given one relevant example to help ground the subsequent judgments. In this paper, we draw up this direction in the context of Conversational Search (CS) and aim to build/augment test collections with grounded LLM based judgments.

Conversational search (CS) is defined as responding to the user's information needs in the context of the conversation [25, 26, 27, 28]. In CS the user's information need depends on the query, the context of the conversation, and the user's personal preferences [29]. This results in highly dynamic, non-linear conversational trajectories – where a user's information need could be answered quite differently depending on the system's interpretation because the needs are evolving and change in response to the information presented. To meet these changing information needs, systems are likely to pull in a wider range of documents. This could result in many more unassessed documents, creating bigger gaps when evaluating new systems, which would significantly reduce the reusability of test collections [30]. So, in this work, we explore whether LLMs can be used to augment and extend CS test collections which are grounded by human annotations, in order to evaluate new, future systems.

In this paper, we leverage both commercial and open-source LLMs in zero- and few-shot, as well as fine-tuning manners, to automatically generate relevance judgments in the CS scenario. Specifically, for commercial LLMs, we use the GPT-3.5 model with different prompts. We try one-shot, two-shot, and zero-shot prompts. For open-source LLMs, we consider three setups: (1) we directly use the Llama checkpoint released by Meng et al. [16], which has undergone fine-tuning based on human-labeled relevance judgments on MS MARCO, (2) we directly prompt Llama-3 [31] in a one-shot way, and (3) we first use partial human-labeled relevance judgments in a CS dataset to fine-tune Llama-3 and then test it using the rest of the relevance judgments in the dataset.

We use the TREC iKAT 2023 [29] dataset which is a personalized CS benchmark. In our experiments, we re-create the relevance judgments of the TREC iKAT 2023 benchmark using various prompts and techniques. We compare the generated judgments with the official TREC iKAT 2023 relevance labels in terms of various metrics. In particular, we are interested in answering the following research questions:

Table 1
Distribution of the relevance scores in train, test, and validation sets.

	0	1	2	3	4	Total
Train	1551	217	1289	661	134	3852
Test	389	60	295	150	23	917
Validation	299	48	219	121	22	709

RQ1 How do different LLMs compare in predicting relevance judgments in CS datasets?

RQ2 How do LLM-generated assessments compare to human-generated assessments in both absolute label prediction and relative ranking of retrieval models in CS datasets?

RQ3 How are new models with different levels of holes ranked using LLM-generated assessments? Can we rely on LLM-generated labels to compare a new model with existing models in CS datasets?

To answer the RQs, we conduct a set of experiments where for **RQ1**, we create a training and test set of relevance labels and compare Llama-1, Llama-3, and GPT-3.5 in zero-shot, few-shot, and fine-tuning settings. For **RQ2**, we use GPT-3.5 to generate relevance labels on the official TREC iKAT 2023 pool and use it to rank the official TREC runs. To answer **RQ3**, we conduct multiple experiments where at each experiment, we remove the relevance labels of each run from the pool, mimicking the case where that model is not included in the original pool. We then generate the relevance labels using GPT-3.5 and use those labels to assess the new model.

Our results show that ranking of the retrieval models in CS datasets using human- and LLM-generated annotations are highly correlated, although they have a low agreement in binary- and graded-level, in line with the findings of Faggioli et al. [19] on ad-hoc search. In addition, the correlation of different IR metrics converges as we add more retrieval systems to the comparison pool. We show that by fine-tuning the open-source LLMs we can achieve a higher agreement between LLM-generated and human judgments. However, higher agreement in terms of binary and graded judgments does not necessarily result in a higher correlation in the ranking of the runs. Moreover, we show that in the case of adding a new retrieval model, filling the holes with a zero-shot Llama model results in less significant shifts in the ranking of the corresponding retrieval model, compared to using one-shot GPT-3.5, perhaps due to higher agreement of the ratings, and because the one-shot GPT-3.5 model is biased to predict higher relevance scores.

2. Methodology

2.1. The choice of LLMs

As a commercial closed-source LLM, we consider the GPT-3.5 (gpt-3.5-turbo-0125) model with values of 0 and 1 for the *temperature* (*tmp*) and *top_p=1*. The value of 0 for temperature means that the model outputs the tokens with the highest probability and has no randomness. For open-source LLMs, we consider the following three setups: (1) we directly use the Llama-1 (7B) checkpoint released by Meng et al. [16],[1] which has undergone fine-tuning using the human-labeled relevance judgments from the development set of MS MARCO [32]; (2) we directly prompt Llama-3 (8B) [31] in a one-shot way; specifically, besides the original version of Llama-3 (8B), we also consider its instruction-tuned version (Llama-3-inst); (3) we first use partial human-labeled relevance judgments in a CS dataset to fine-tune the two variants Llama-3, and then test them using the rest of the relevance judgments in the dataset.

2.2. Dataset

We use the TREC iKAT 2023 [29] benchmark in our experiments. In this benchmark, the relevance of each query–document pair is assessed and represented with a score in the range of 0–4. Using

[1] https://github.com/ChuanMeng/QPP-GenRE

the GPT-3.5 model, we judge the relevance of all query–document pairs from the TREC iKAT 2023 collection. For fine-tuning the Llama-3 model [31], we divide the TREC iKAT 2023 benchmark into train, test, and validation sets. First, we randomly remove 20680 irrelevant documents ($score < 2$) from the existing pool to ensure that the number of relevant ($score >= 2$) and irrelevant documents for each query are the same. Second, we randomly split the documents for each query between train, test, and validation sets. We keep the portion of train, test, and validation sets as 70%, 15%, and 15%. The train, test, and validation sets include 3852, 917, and 709 query-passage pairs, respectively. The distribution of the labels for the train, test, and validation set are shown in Table 1. We ensure that all user queries appear in the training set and have at least one positive and one negative document.

2.3. Retrieval models

To assess the correlation of the generated pools, we use the baselines and runs submitted to TREC iKAT 2023. There are in total 28 baselines and runs submitted to the TREC iKAT 2023. We use the output of retrieval for these runs and baselines which are released by the organizers.

2.4. Metrics

To assess the ranking-level performance of our proposed models for relevance judgment, we rank the retrieval models two times, once based on their performance using the main pool and second based on the generated pool. We compute and report Kendall's Tau (τ) metrics to assess the correlation between the two rankings. To assess the agreement between proposed models for relevance judgment and humans, we report Cohen's Kappa agreement at binary and graded levels. We convert the predicted graded scores (0-4) to a binary label by considering scores 2-4 as relevant and scores 0-1 as irrelevant.

2.5. GPT-3.5 prompt design

We designed three different prompts, inspired by relevant work. We use the resolved utterances provided by humans as a context-independent query in our prompts.

- We design a zero-shot prompt inspired by the prompt used in Thomas et al. [18] which is shown in Table 2.
- Our next prompt is a one-shot prompt which includes a relevant document with a relevance score of 4. This prompt is inspired by the prompt used by MacAvaney and Soldaini [15]. This prompt is shown in Table 2. We use the canonical response for the corresponding user utterance from the TREC iKAT 2023 collection as the relevant (perfect) example with a relevance score of 4. The canonical responses are provided by the organizers and are supposed to be the best possible answer that can be given to the utterance at every point in the conversation [29].
- The third prompt includes a relevant document and an irrelevant document. We randomly sample these documents from the TREC iKAT 2023 official pool. A document with a relevance score higher than or equal to 2 is selected as relevant and a document with a relevance score lower than 2 is randomly selected as irrelevant. This prompt is shown in Table 2.

2.6. Llama fine-tuning and prompt design

For open-source LLMs, we consider the following setups:

- We use the one-shot prompt shown in Table 2 to prompt Llama-3.
- For fine-tuning Llama-3, we follow Meng et al. [16] to fine-tune Llama-3 using a novel PEFT method, 4-bit QLoRA [23]; the train, test, and validation data used for fine-tuning and inference over the model is explained above.
- Because the Llama model released by Meng et al. [16] is only trained to generate binary relevance judgments given a query and a passage on MS MARCO, we leave out the user personal knowledge when we use the Llama model released by Meng et al. [16].

Table 2
The template of the prompts designed for relevance judgment. The orange rows belong to the one-shot prompt, the blue rows belong to two-shot prompt. The black lines are in all prompts including zero-shot, one-shot and few-shot.

The prompts used for relevance judgment
Instruction: You are a search quality rater evaluating the relevance of web pages. Given the persona of the user, user query, and a web page, you must provide a score on an integer scale of 0 to 4 to indicate to what extent the given document meets the information needs of the user. The scores have the following meanings: 0: fails to meet, 1: slightly meets 2: moderately meets, 3: highly meets, 4: fully meets User persona: { ptkb } Query: { utterance } Document 1: { canonical response } Score: { 4 } Document 1: { passage 1 } Score: { score 1 } Document 2: { passage 2 } Score: { score 2 } Document : { document } Score: Please only generate an int score between 0 to 4 to say to what extent the document is relevant to the user question. Score lower than 2 means the document is irrelevant.

3. Experiments & Results

In this section, we describe in detail the experiment we designed to answer each research question, followed by the results we obtained by doing the experiments.

3.1. LLM relevance label comparison (RQ1)

3.1.1. Experimental design

In this section, we aim to answer our first research question, ***RQ1:*** *How do different LLMs compare in predicting relevance judgments in conversational search?* To do so, as described in Section 2, we randomly sample the human-generated labels into the train, validation, and test sets and use the training data to fine-tune Llama-based models. We then compare the performance of the Llama-based models with the different GPT-3.5-based models.

3.1.2. Results

In Table 3, we report the agreement of our proposed models on the test set. The experiments reveal that we can improve the agreement by fine-tuning the Llama-3-inst model. As can be seen, the fine-tuned Llama-3-inst achieves the agreement of 0.729 on the binary level.

We use fine-tuned and zero-shot Llama to predict the test set. We create a small pool based on

Table 3
Kappa Cohen's agreement between human and LLM labels on the test set at binary and graded levels.

LLM	Prompt	tmp	Binary	Graded
GPT-3.5	zero-shot	1	0.410	0.151
		0	0.489	0.117
	one-shot	1	0.499	0.212
		0	0.543	0.212
	two-shot	1	0.329	0.134
		0	0.454	0.184
Llama-3	one-shot	-	0.015	-0.005
	Fine-tuned	-	0.687	0.527
Llama-3-inst	one-shot	-	0.127	0.092
	Fine-tuned	-	**0.729**	**0.553**
Llama-1 (pre-trained) [16]		-	0.386	-

Table 4
Comparison between the relative ranking of TREC iKAT 2023 runs using (1) subset of human-generated pools on the test set and (2) LLM-generated labels on the same test subset. The relative ranking is compared using Kendall's Tau (τ) metric.

LLM	Prompt	tmp	NDCG@3	NDCG@5	NDCG	P@10	R@10	R@1000	mAP	MRR
GPT-3.5	zero-shot	1	0.788	0.772	0.820	0.857	0.709	0.884	0.746	0.815
		0	**0.852**	**0.799**	**0.862**	0.820	0.767	0.868	0.741	0.810
	one-shot	1	0.836	0.794	**0.862**	0.820	0.804	**0.931**	0.783	0.794
		0	0.783	0.778	0.831	**0.915**	**0.841**	0.921	0.772	0.847
	two-shot	1	0.810	0.762	0.810	0.841	0.794	0.894	**0.794**	0.873
		0	0.767	0.746	0.772	0.884	0.810	0.884	0.735	0.804
Llama-3	Fine-tuned	-	0.788	0.772	0.868	0.905	0.810	0.915	0.772	**0.921**
	one-shot	-	0.614	0.571	0.778	0.630	0.614	0.820	0.640	0.735
Llama-3-inst	Fine-tuned	-	0.762	0.751	0.820	0.894	0.825	0.894	0.741	0.847
	one-shot	-	0.550	0.534	0.788	0.624	0.550	0.868	0.635	0.614

the model's predictions on the test set. We sorted the TREC iKAT 2023 runs based on their retrieval performance two times (1) using the LLM-generated assessments and (2) using the human-generated pool. We compare the ranking of runs by computing the correlation between them. Table 4 reports the result of the relative ranking performance of different LLMs, compared to human ranking. Surprisingly, the Llama-3 model is not performing better than the GPT-3.5 model in this scenario while it has a higher agreement with human labels on the same test set. This could be due to the different labeling biases that the models have where GPT-3.5 labels could be more different from human labels in terms of absolute numbers, but when we compare different documents they are more similar relatively.

We report the binary- and graded-level confusion matrices for prediction of best Llama- and GPT-3.5-based models over the test set in Tables 5 and 6, respectively. Additionally, we report the binary confusion matrix of the Llama-1 which is fine-tuned on the MS MARCO dataset. According to Table 6, the fine-tuned Llama-3-inst has a very lower tendency to assign scores 1 and 4 compared to scores 0 and 2. The behavior of the Llama is natural as the train data has less number of 1 and 4 labels compared to other labels according to Table 1. We do not observe such bias in the one-shot GPT-3.5 as this model is not fine-tuned on the train data. However, the one-shot GPT-3.5 has predicted a large number of 4 labels compared to the Llama. This bias could be the result of putting the canonical answer in the prompt as a positive example with a score of 4.

As can be seen in Table 5, the one-shot GPT-3.5 has more false positives compared to the fine-tuned

Table 5
Binary-level confusion matrix over the test set using different fill holing models.

LLM		0	1	sum
Llama-3	0	384	59*	443
Fine-tuned	1	65*	409	474
GPT-3.5	0	319	79*	398
one-shot (tmp=0)	1	130*	389	519
Llama-1	0	287	119	406
(pre-trained) [16]	1	162	349	511

Table 6
Graded-level confusion matrix over the test set using different fill holing models.

LLM		0	1	2	3	4	sum
Llama-3-inst	0	359	22	42	5	1	429
Fine-tuned	1	2	1	9	2	0	14*
	2	22	30	198	52	4	306
	3	6	7	42	79	11	145
	4	0	0	4	12	7	23*
GPT-3.5	0	221	7	21	9	0	258
one-shot ($tmp = 0$)	1	71	20	37	8	4	142*
	2	42	8	57	22	4	133
	3	32	17	75	37	3	164
	4	23	8	105	74	12	222*

Llama over the binary-level labels. Giving one positive example in the prompt might cause this bias. The distribution of the false positive and false negative are approximately equal for the Llama. This might be because the number of relevant and irrelevant passages in the training set of LLaMA is equal.

3.2. LLM vs. human labels (RQ2)

3.2.1. Experimental design

Here, we aim to answer our second research question, ***RQ2:*** *How do LLM-generated assessments compare to human-generated assessments in both absolute label prediction and relative ranking of retrieval models in CS datasets?* To do so, we regenerate all the relevance labels of the official TREC iKAT 2023 pool using the three prompts described in Section 2. Inspired by Faggioli et al. [19] and MacAvaney and Soldaini [15] we aim to test the hypothetical case of having zero or one assessed passage for each query and rely on LLMs to assess the pool. In this experiment, we evaluate the models based on both the

Table 7
Cohen's Kappa agreement between human and LLM-generated labels from TREC iKAT 2023 dataset.

LLM	Prompt	Temperature	Binary	Graded
GPT-3.5	zero-shot	1	0.170	0.099
		0	0.207	0.041
	one-shot	1	0.235	0.137
		0	**0.269**	**0.155**
	two-shot	1	0.133	0.076
		0	0.218	0.152
Llama-1 (pre-trained) [16]		-	0.186	-

Table 8
Comparison between the relative ranking of TREC iKAT 2023 runs using LLM- and human-generated pools. In this table, GPT-3.5 model is used as LLM, and the pool is re-assessed completely using different variations of prompts. The relative ranking is compared using Kendall's Tau (τ) metric.

Prompt	tmp	NDCG@3	NDCG@5	NDCG	P@10	R@10	R@1000	mAP	MRR
zero-shot	0	**0.862**	**0.862**	0.841	0.836	0.794	0.709	0.730	**0.873**
	1	0.852	**0.862**	0.836	0.847	0.825	0.847	0.794	0.831
one-shot	0	0.836	0.852	**0.926**	0.815	0.820	**0.905**	**0.878**	0.804
	1	**0.862**	**0.862**	0.899	0.836	**0.836**	0.868	0.852	0.804
two-shot	0	0.831	0.825	0.852	**0.868**	0.831	0.889	0.836	0.862
	1	0.804	0.804	0.778	0.772	0.772	0.772	0.735	0.767

quality of individual predicted labels and the relative ranking of the models assessed with each of the generated labels, compared to human labels.

3.2.2. Results

We report the agreement of proposed models with human labels from the complete pool of the TREC iKAT 2023 benchmark in Table 7. As can be seen, the one-shot prompting of the GPT-3.5 has the highest agreement with human labels among zero-shot and two-shot prompting. Additionally, setting the temperature to 0 increases the agreement. A lower value for the temperature parameter means that the model has less randomness in generating the output while higher values mean that the model is more creative and has randomness in generation. Our one-shot prompt has the highest agreement in terms of both binary and graded labels. The better performance of one-shot prompting compared to two-shot prompts indicates that (1) using the canonical response as a positive example is more useful and (2) using two positive and negative examples confuses the GPT-3.5.

We use the pools generated by GPT-3.5 in different settings and the official pool assessed by humans to assess the runs and rank them. More correlation between the rankings obtained by the LLM-generated pool and the human-assessed pool indicates that using LLM-generated assessments is as effective as using human-generated assessments. Table 8 shows the correlation between the relative ranking of runs using different LLM-generated pools with the human-assessed pool. As can be seen, one-shot prompting the ChatGPT model significantly outperforms the other settings over Kendall's Tau correlation metric. Interestingly, we observe that the temperature of 0 is not always better than the temperature of 1 in terms of all retrieval metrics.

In Figure 1, we show the correlation between the relative ranking of LLM-generated and human-generated pools using the K best-performing runs. The best-performing model is selected according to the ranking based on using a human-generated pool. The LLM-generated pool is generated using the best pool generation model from Table 8, i.e., one-shot labeling ChatGPT using temperature of 0. Considering the 4 best-performing runs, the relative ranking using the LLM-generated assessments is the same as using the human-generated assessments over all ranking metrics. Using the LLM-generated assessments, the relative ranking of the 10 best-performing runs based on NDCG@5 is the same as the relative ranking of these runs based on human-generated assessments (Kendall's Tau = 1). According to Figure 1, as the value of K increases (more runs are included in the comparison), the value of the correlation converges. This finding represents the reliability of LLM-generated assessments in terms of relative ranking.

3.3. Filling judgment holes (RQ3)

3.3.1. Experimental design

To answer our ***RQ3**: How are new models with different levels of holes ranked using LLM-generated assessments? Can we rely on LLM-generated labels to compare a new model with existing models?*, we simulate the case where a new model is being tested using TREC iKAT 2023 runs. To do so, we do multiple experiments where in each one we take out all the judgments of one run while keeping those judgments that are in common with other existing runs. This leads to different levels of holes per run, depending on their similarity to other existing models. We then assess the relevance of the unjudged passages using GPT-3.5 and use those labels to compute the performance of the model. To assess performance, we compare the ranking of the model using the original human assessments vs. GPT-3.5-generated assessments and report the absolute difference in the model's ranking in the two cases. This indicates, how reliable LLM-generated labels are in filling the holes for new models. After removing a run and generating labels for it using GPT-3.5, we do the ranking based on the (1) new pool (human pool filled by LLM) and (2) human pool which includes the human judgments for the holes of the current run.

3.3.2. Results

The value of the absolute distance of the run in the two rankings based on the portion of the Unjudged@10 passages for that specific run is shown in Figure 2. We use the one-shot GPT-3.5 model with a temperature of 0 for hole filling. As can be seen, as the value of Unjudged@10 increases, the absolute distance increases which means the missing run also increases. This makes sense because we know that GPT-3.5 is biased to rate the passages with higher scores compared to humans. As a result, we can conclude that given a new ranking model with a lot of missing judgments (a larger value for Unjudged@10), it is advisable to recreate the whole pool using the GPT-3.5, rather than augmenting the existing human-created pool by filling the holes using GPT-3.5.

Interestingly, we see that the results of Llama exhibit a completely different trend where the number of holes does not seem to matter. We see in the plot that Llama can consistently rank the missing run close to its original ranking and even achieves perfect ranking at some points. This is in line with our observation in Table 3, where we observed a higher agreement of Llama-generated labels with human labels, leading to a lower disparity in terms of the absolute value of the labels, which then makes the augmented labels more reliable.

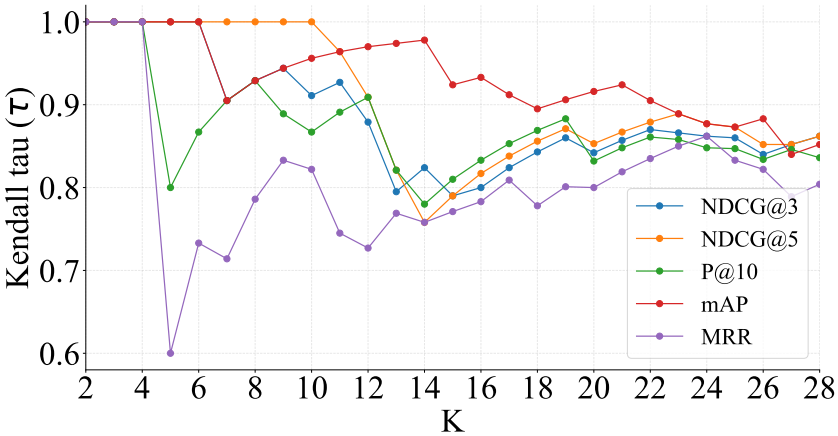

Figure 1: Rank correlation between human- and LLM-generated pools using the K best-performing runs.

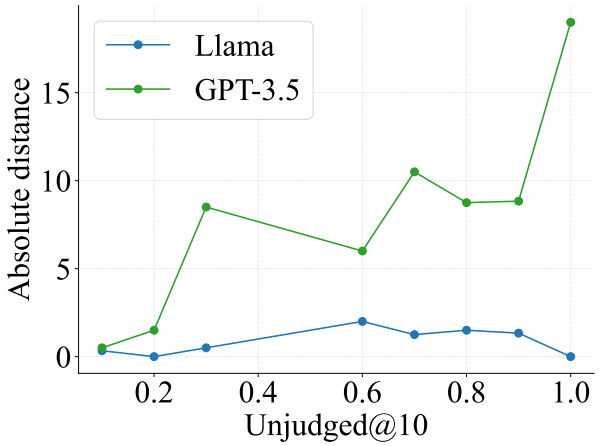

Figure 2: Absolute distance between the location of a new run before and after filling holes using GPT-3.5 and Llama. The X-axis shows the average of unjudged documents among the top 10 documents returned by a new run.

4. Conclusion

In this work, we conducted extensive experiments to study the effect of LLM-generated relevance judgments on incomplete relevance judgments (aka. "holes") of the TREC iKAT 2023 collection. We studied the effectiveness of different open-source and closed-source LLMs on generating relevance assessments on the same set, where we observed that labels by fine-tuned Llama align better with human labels compared to the labels obtained by few-shot prompting the GPT-3.5 model. In line with previous work, we observed that automatic judgments from LLMs result in highly correlated model rankings; however, we found that it substantially correlates lower when human plus automatic judgments were used when a new model was being assessed on the pool. We further found that, depending on the LLM employed, new runs will be highly favored (or penalized), and this effect is magnified proportional to the size of the holes. We conclude that generating automatic labels on the whole pool is more effective, rather than just the missing holes, as it leads to higher correlation and ensures that the same labeling biases are applied to all the models. Further work is needed to refine prompt engineering and fine-tuning of LLMs so they better match and reflect human annotations. This will help align the models more closely with their intended purpose. Moreover, we plan to simulate various labeling strategies to study the effectiveness of fine-tuning in more practical scenarios.

References

[1] C. W. Cleverdon, The cranfield tests on index language devices, Aslib Proceedings 19 (1967) 173–194.

[2] K. Sparck-Jones, C. J. Van Rijsbergen, Report on the Need for and Provision of an 'Ideal' Information Retrieval Test Collection, Technical Report British Library Research and Development Report No. 5266, Computer Laboratory, University of Cambridge, 1975.

[3] E. M. Voorhees, D. K. Harman, TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing), The MIT Press, 2005.

[4] G. V. Cormack, C. R. Palmer, C. L. A. Clarke, Efficient construction of large test collections, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, Association for Computing Machinery, 1998, p. 282–289.

[5] C. Buckley, E. M. Voorhees, Retrieval evaluation with incomplete information, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, 2004, p. 25–32.

[6] M. Sanderson, J. Zobel, Information retrieval system evaluation: effort, sensitivity, and reliabil-

ity, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, 2005, p. 162–169.
[7] X. Lu, A. Moffat, J. S. Culpepper, The effect of pooling and evaluation depth on ir metrics, Information Retrieval Journal 19 (2016) 416–445.
[8] M. Baillie, L. Azzopardi, I. Ruthven, A retrieval evaluation methodology for incomplete relevance assessments, in: Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, volume 4425 of *Lecture Notes in Computer Science*, Springer, 2007, pp. 271–282.
[9] M. Baillie, L. Azzopardi, I. Ruthven, Evaluating epistemic uncertainty under incomplete assessments, Information processing & management 44 (2008) 811–837.
[10] J. A. Aslam, E. Yilmaz, Inferring document relevance from incomplete information, in: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07, Association for Computing Machinery, 2007, p. 633–642.
[11] B. Carterette, J. Allan, R. Sitaraman, Minimal test collections for retrieval evaluation, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 268–275.
[12] A. Lipani, J. Palotti, M. Lupu, F. Piroi, G. Zuccon, A. Hanbury, Fixed-cost pooling strategies based on ir evaluation measures, in: Advances in Information Retrieval, 2017, pp. 357–368.
[13] A. Moffat, W. Webber, J. Zobel, Strategic system comparisons via targeted relevance judgments, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, 2007, p. 375–382.
[14] E. M. Voorhees, N. Craswell, J. Lin, Too many relevants: Whither cranfield test collections?, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, 2022, p. 2970–2980.
[15] S. MacAvaney, L. Soldaini, One-shot labeling for automatic relevance estimation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2230–2235.
[16] C. Meng, N. Arabzadeh, A. Askari, M. Aliannejadi, M. de Rijke, Query performance prediction using relevance judgments generated by large language models, arXiv preprint arXiv:2404.01012 (2024).
[17] E. Khramtsova, S. Zhuang, M. Baktashmotlagh, G. Zuccon, Leveraging llms for unsupervised dense retriever ranking, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, 2024, p. 1307–1317.
[18] P. Thomas, S. Spielman, N. Craswell, B. Mitra, Large language models can accurately predict searcher preferences, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 1930–1940.
[19] 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: ICTIR, 2023, pp. 39–50.
[20] R. Pradeep, S. Sharifymoghaddam, J. Lin, Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze!, arXiv preprint arXiv:2312.02724 (2023).
[21] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, Journal of Machine Learning Research 25 (2024) 1–53.
[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[23] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: efficient finetuning of quantized llms, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., 2024.
[24] A. J. Peterson, Ai and the problem of knowledge collapse, 2024. arXiv:2404.03502.

[25] F. Radlinski, N. Craswell, A theoretical framework for conversational search, in: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR '17, Association for Computing Machinery, 2017, p. 117–126.
[26] L. Azzopardi, M. Dubiel, M. Halvey, J. Dalton, A conceptual framework for conversational search and recommendation: Conceptualizing agent-human interactions during the conversational search process, in: Proceedings of the CAIR'18: Second International Workshop on Conversational Approaches to Information Retrieval at SIGIR 2018, 2018.
[27] C. Meng, N. Arabzadeh, M. Aliannejadi, M. de Rijke, Query performance prediction: From ad-hoc to conversational search, in: SIGIR, 2023, p. 2583–2593.
[28] C. Meng, M. Aliannejadi, M. de Rijke, System initiative prediction for multi-turn conversational information seeking, in: CIKM, 2023, pp. 1807–1817.
[29] M. Aliannejadi, Z. Abbasiantaeb, S. Chatterjee, J. Dalton, L. Azzopardi, Trec ikat 2023: A test collection for evaluating conversational and interactive knowledge assistants, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, 2024, p. 819–829.
[30] Z. Abbasiantaeb, M. Aliannejadi, Generate then retrieve: Conversational response retrieval using llms as answer and query generators, CoRR abs/2403.19302 (2024). arXiv:2403.19302.
[31] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
[32] P. Bajaj, D. Campos, N. Craswell, L. Deng, X. L. Jianfeng Gao, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, Ms marco: A human generated machine reading comprehension dataset, in: NIPS, 2016.