A Cost-Effective Framework to Evaluate LLM-Generated Relevance Judgements

Simone Merlo
University of Padua
Padua, Italy
simone.merlo@phd.unipd.it

Stefano Marchesin
University of Padua
Padua, Italy
stefano.marchesin@unipd.it

Guglielmo Faggioli
University of Padua
Padua, Italy
guglielmo.faggioli@unipd.it

Nicola Ferro
University of Padua
Padua, Italy
ferro@dei.unipd.it

Abstract

Large Language Models (LLMs) hugely impacted many research fields, including Information Retrieval (IR), where they are used for many sub-tasks, such as query rewriting and retrieval augmented generation. At the same time, the research community is investigating whether and how to use LLMs to support, or even replace, humans to generate relevance judgments. Indeed, generating relevance judgements automatically – or integrating an LLM in the annotation process – would allow us to improve the number of evaluation collections, also for scenarios where the annotation process is particularly challenging. To validate relevance judgements produced by an LLM they are compared with human-made relevance judgements, measuring the inter-assessor agreement between the human and the LLM.

Our work introduces an innovative framework for estimating the quality of LLM-generated relevance judgments, providing statistical guarantees while minimizing human involvement. The proposed framework allows to: i) estimate the quality of LLM-generated relevance judgments with a defined confidence while minimizing human involvement; and ii) estimate the quality of LLM-generated relevance judgments with a fixed budget while providing bounds on the estimate. Our experimental results on three well-known IR collections using multiple LLMs as assessors show it is sufficient to assess 16% of the LLM-generated relevance judgments to estimate the LLM's performance with a 95% confidence.

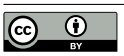
CCS Concepts

• **Information systems → Relevance assessment**.

Keywords

Relevance Assessment; Large Language Models; Quality Estimation

ACM Reference Format:
Simone Merlo, Stefano Marchesin, Guglielmo Faggioli, and Nicola Ferro. 2025. A Cost-Effective Framework to Evaluate LLM-Generated Relevance Judgements. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3746252.3761200



CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761200

1 Introduction

Traditionally, Information Retrieval (IR) evaluation collections were built following the Cranfield paradigm [13]. According to the paradigm, IR evaluation collections are composed of three elements: a set of information needs, represented as queries, a corpus of documents, and a set of relevance judgments that describe the relevance of the documents in response to the information needs. While the information needs can be extracted from a query log and the corpus can be crawled or obtained from existing repositories (e.g., news or research papers), relevance judgements are far more complex to collect as they require manual human work. In this regard, the main evaluation campaigns, such as TREC [46], CLEF [25], FIRE, and NTCIR [41], rely on expert assessors to collect the relevance judgements. While this approach ensures high-quality judgments, it also comes with a huge investment in temporal and economic costs. Furthermore, there are specific scenarios where it is particularly hard to find experts, such as when it comes to low-resource languages, or when the investment might be particularly demanding, like in legal or medical domains. Thus, the advent of Large Language Models (LLMs), which are capable of mimicking the human language, propelled the research community to investigate new, cheaper approaches to produce relevance judgements.

Despite these approaches are gaining popularity in both academia [21, 51] and industry [48], it is not clear yet how to properly validate the automatically-generated relevance judgements. Most of the proposed works [21, 48, 51, 55] evaluate the quality of LLMs as assessors by first using them to generate relevance judgements for historical IR collections, and measuring the agreement with the human-made judgements available in the collections. Although reasonable, this approach assumes that an IR collection is already available. Using historical test collections – often dated and possibly leaked – does not guarantee that the quality of the LLM as assessor generalizes to previously unseen topics or documents. If we wanted to evaluate it on new data, the cost of validating an LLM as assessor on such data would be at least equal to the cost of constructing a new human-made collection.

Here, we propose a cost-effective strategy to validate an LLM assessment process compared to an ideal gold standard, represented

by the human assessor. To exemplify this, we focus on fully automatically generated relevance judgements.

With this perspective in mind, we articulate our work on two research questions:

- **RQ1:** How to estimate the quality of LLM-generated relevance judgments minimizing the human involvement (cost) while providing statistical guarantees?
- **RQ2:** How to estimate the quality of LLM-generated relevance judgments with a limited budget while providing statistical guarantees?

Contributions. To overcome the discussed limitations, we propose a statistical framework to validate an LLM as an assessor by using the least amount of human-made relevance judgments – which remain an integral part of the evaluation process. Our framework relies on a combination of sampling strategies, point estimators, and Confidence Intervals (CIs), which enable the estimation of the LLM effectiveness as an assessor with strong statistical guarantees. Specifically, we focus on estimating the Mean Absolute Error (MAE) and Cohen's κ of the judgments generated by the LLM, adopting Simple Random Sampling (SRS) as the sampling strategy and the Wald interval as CI. The SRS sampling strategy ensures that the sampled judgments are representatives of the entire set of LLM-generated relevance judgments. The Wald interval makes it possible to provide statistical guarantees on the estimate obtained considering only the sampled judgments. Our work advances the current state of the art [24, 47, 48] by: (1) minimizing the human made judgments, hence avoiding overpowered tests; and (2) providing strong statistical guarantees on the outcome of the evaluation.

The proposed framework is very effective. Indeed, human assessors are required to validate fewer than 6% to 16% of the LLM-generated relevance judgments to estimate the human-LLM inter-assessor agreement or the LLM's MAE, respectively. This holds for the considered collections and LLMs and for a confidence of 95% on the estimate.

The rest of this work is organized as follows: in Section 2 we review the current methodologies to evaluate LLM-generated relevance judgments; in Section 3 we describe our proposed quality estimation framework; in Sections 4 and 5 we describe the annotation process and validate its evaluation, respectively; in Section 6 we detail challenges and opportunities related to the framework. Finally, in Section 7 provide the final remarks.

2 Related Work

Recent developments concerning LLMs opened up new research frontiers. In the IR field, the creation of the relevance judgments represents one of the aspects strongly affected by those advancements. Traditionally, relevance judgments were created manually [4, 52] by human expert assessors or by exploiting real user click logs [8, 16–20, 36]. This is the case for the relevance judgments of the best-known collections. To mitigate the costs of the human annotations many works started investigating how to use LLMs to replace humans for the relevance judgments creation [21, 48, 51, 55]. Faggioli et al. [24] first discussed the perspectives related to the use of LLMs in IR for the generation of the relevance judgments. Thomas et al. [48] proposed a first methodology to generate the relevance judgments with LLMs while exploring the effectiveness of several different prompt configurations. Upadhyay et al. [51] provided practical guidelines and a prompt refining of the approach introduced by Thomas et al. [48]. In particular, the approach proposed by Upadhyay et al. [51] is known as UMBRELA and has been used to generate the relevance judgments related to the TREC Retrieval Augmented Generation (RAG) 2024 track [37]. Others, such as Türkmen et al. [49], hypothesise that entire collections, including topics, documents and relevance judgements, can be created by LLMs.

In addition, LLMs started to be exploited not only to generate relevance judgments from scratch but also to fill in the blanks (*i.e.,* non-annotated query-document pairs) present in many test collections [33, 50].

In this view, also challenges and workshops in major top conferences like SIGIR 2024 [1] and TREC 2024[1] have been organized. An example is represented by the "LLMJudge" challenge [40], part of the "LLM4Eval" workshop [38], which took place at SIGIR 2024. This challenge was aimed at studying whether the LLMs can match the accuracy of the human assessors and which are the most effective prompts and models.

The consequences of using LLMs to generate relevance judgments are still debated. Some researchers argue that LLMs can be comparable or even better than humans in the relevance judgments generation task and/or similar tasks [29, 30, 43, 48]. Others claim that the reliability of LLMs is not even close to that of humans [3, 12] and that LLMs should not be used in this context [45]. Many studies have been conducted in this direction [7, 9, 22, 24, 56], but there is still no agreement among the research community.

In this context, the most frequently used comparison metrics are MAE and Cohen κ. Another approach consists of considering as a proxy of the quality of LLM-generated relevance judgements their effectiveness on the downstream task of evaluating IR systems [48, 50, 51]. In particular, it focuses on measuring the consistency of the ranking of the IR systems using either human-made relevance judgements or LLM-generated ones. Ideally, LLM-generated relevance judgements are more useful the more they rank IR systems the same way human-made relevance judgements would. The most frequently used metrics for this purpose are Kendall τ and Spearman ρ. Nonetheless, these approaches do not provide guarantees on the quality of the generated data if the LLM-generated relevance judgements are created for a new set of data, such as new topics, documents, or both. For this reason, in this work, we propose a new methodology to estimate the quality of the relevance judgment produced by an LLM minimizing the human involvement and, therefore, the related costs.

2.1 The Need for a Statistical Framework

Several works on this research line test the performance of the proposed approaches on historical collections for which human-made relevance judgments already exist [21, 35, 39, 48, 51]. This approach allows us to evaluate the quality of the LLM-generated relevance judgements only on the available test collections — which could be dated or could have been ingested by the LLM. For this reason, there is no guarantee that the quality of a given LLM and used prompt will generalise on previously unseen topics, documents,

[1] https://trec.nist.gov/pubs/call2024.html

or even test collections. This approach does not help us evaluate how good the LLM would be if we wanted to use it as a judge to annotate a novel collection.

Other works have also employed sampling to reduce annotation costs when evaluating the performance of automatic or semi-automatic LLM-based relevance judgement creation [24, 47, 48].

Compared to such efforts, our work aims at providing practitioners with a sound and grounded methodology to carry out such a sampling. In particular, our contribution revolves around two key axes. First, our work details the theoretical foundations to build an evaluation procedure that minimises the annotation cost – for humans and LLMs alike– saving time, money, and environmental resources. Secondly, our work promotes good practices, describing how to correctly compute the estimators, the variance, and the CI when evaluating LLMs as assessors. Finally, it is worth noting that the proposed framework can be seamlessly applied to both existing collections – serving as a robust and efficient evaluation mechanism – and new ones – acting as a reliable and cost-effective solution for assessing the practical utility of LLM-generated judgements.

3 LLM Quality Estimation Framework

In this section, we first provide an intuition on the LLM quality estimation framework, then we describe the proposed pipeline.

3.1 Intuition

To overcome the limitations highlighted in Section 2, it would beneficial to have a validation methodology that: i) allows determining the quality of the LLM-generated relevance judgements with fewer human-made assessments than those available in the whole collection; ii) provides statistical guarantees about the extent to which the quality of such relevance judgements generalises to unseen parts of the collection. By relying on sampling to minimize assessments while remaining representative of the entire collection and by computing CIs to gauge the uncertainties inherent in the estimated LLM quality, the evaluation methodology we propose provides the above mentioned, desired properties.

Figure 1 illustrates the quality estimation pipeline. Assume we are interested in evaluating the quality of a set of LLM-generated relevance judgements. This set of relevance judgements contains queries, documents, and corresponding relevance judgements. During the first step, indicated with (A), the researcher decides, based on external aspects – such as the available budget or the desired level of confidence – how the procedure should be designed. The proposed procedures are two, depending on whether it is required to satisfy confidence (purple box) or budget (blue box) constraints.[2] The second step, indicated with (B), consists of sampling a subset of queries, documents, and corresponding relevance judgements. During the third step ((C)), these "triplets" are annotated in terms of correctness. In other words, the LLM-generated judgments are compared with the human-made ones and considered correct if they match, and incorrect otherwise. Note that the assessment can be done by an expert assessor, as in the TREC paradigm, or collected through other procedures – such as via crowdsourcing or click logs. During the fourth step, indicated with (D), the LLM quality estimate and the corresponding CI are computed using the triplets sampled and manually assessed in the previous steps. Depending on the considered quality measure, an appropriate estimator must be selected. In this work we consider MAE and Cohen's κ measures. During the last step ((E)), for the confidence based procedure (purple box), a constraint check is performed to verify whether the obtained CI meets the required confidence, i.e. the Margin of Error (MoE) is lower than a predefined threshold. If the check is satisfied, the procedure outputs the quality estimate and the corresponding CI. Otherwise, the procedure loops back to the sampling step ((B)).

Below, we provide an example illustrating the advantages of the proposed quality estimation pipeline over prior work.

Example 3.1. Let us assume a researcher wants to create a new IR collection. They collected, by crawling the Web, around $9M$ documents. Furthermore, let us also assume that, from a set of related query logs, the researcher obtained 43 queries. At this point, the researcher needs to create relevance judgments, but they have a limited budget. Thus, the researcher decides to employ 3 LLMs to generate the relevance judgments for $9k$ query-document pairs. To make an informed use, they must validate the LLM-generated relevance judgments. However, verifying all the LLM-generated judgments would require an effort equal to simply creating human-made relevance judgments. To overcome this limitation, the researcher decides to adopt the proposed quality estimation framework for every considered LLM. In this regard, they require a quality estimate with a confidence level of 95%, setting a threshold $\epsilon = 0.05$ for the MoE. With this setup, the researcher applies the confidence-based iterative procedure. First, they start by sampling and validating, in each iteration, an LLM-generated relevance judgment. Then, once the sample is validated, the researcher estimates the quality with the corresponding estimator and computes the corresponding CI. Finally, the iterative procedure stops when the MoE (i.e., half the width of the CI) is small enough. At the end of the procedure, the researcher validated fewer than 1500 LLM-generated relevance judgments: not even 17% of the total judgements![3] Based on the performed validation, the researcher can rely on the top performing LLM and its relevance judgments. In this way, the researcher can publish their test collection, together with the LLM estimated quality and the corresponding CI.

3.2 Pipeline

In the following, we first introduce the required notation, and then describe each step of the quality estimation pipeline.

Notation. Let us denote with $\mathcal{R}$ the set of relevance judgments produced by an LLM, where $r \in \mathcal{R}$ represents a single relevance judgment and $t(r)$ refers to its real (human-determined) relevance value. Let us also denote with $\mathcal{S}$ a sampling strategy and with $\mathcal{R}_\mathcal{S} \subset \mathcal{R}$ the set of LLM judgments sampled from $\mathcal{R}$ according to $\mathcal{S}$, where $n = |\mathcal{R}_\mathcal{S}|$ is the sample size. Moreover, let us define with $\text{cost}(\mathcal{R}_\mathcal{S})$ the cost of the human work required to validate the sampled LLM-generated relevance judgments, with θ the value of the considered quality metric measured on $\mathcal{R}$, with $\hat{\theta}$ the estimate

[2] The confidence based procedure is inspired by the work of Gao et al. [27] in the context of data quality management.

[3] The numbers used in this example represent real values obtained while estimating the MAE of an LLM assessor for the TREC Deep Learning (DL) 2019 collection (see Section 5).

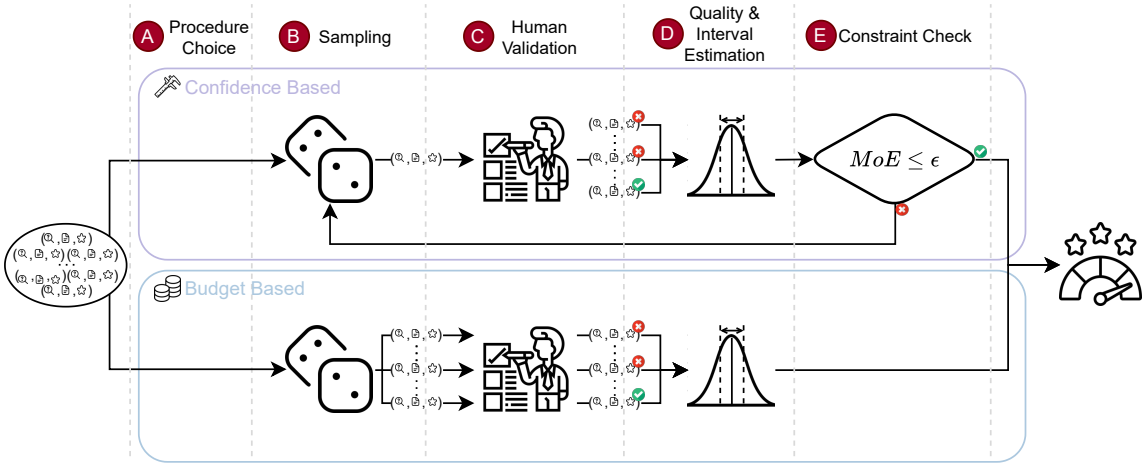


Figure 1: Pipeline of the proposed LLM quality estimation framework.

of the quality metric measured on $\mathcal{R}_\mathcal{S}$, and with $V(\hat{\theta})$ the variance of such estimate. To gauge the uncertainties in the estimate $\hat{\theta}$ due to sampling, a $1-\alpha$ CI is provided, where α denotes the significance level. Finally, let us denote half the width of a CI as MoE.

3.2.1 Procedure Choice. The LLM quality estimation framework is articulated in two procedures, focusing on two different facets:

(1) **Confidence Based Procedure (RQ1):** Estimating the quality of LLM-generated relevance judgments with a desired confidence while minimizing human involvement.
(2) **Budget Based Procedure (RQ2):** Estimating the quality of LLM-generated relevance judgments with a fixed budget.

In the first step of the quality estimation pipeline (block A in Figure 1), the researcher needs to choose among one of the two procedures depending on whether they have to satisfy confidence or budget constraints. The former requires to set a threshold ϵ for the MoE, while the latter requires to set a threshold b for the cost. Note that the validation cost can be computed in multiple ways, such as considering the time or money required to have human involvement in the validation of the sample. In this work, we consider the number of sampled LLM-generated relevance judgments as the validation cost, since it represents a reasonable proxy for temporal or monetary costs. To obtain a reasonable estimate of the costs, it would be sufficient to multiply the number of sampled relevance judgments by the average temporal or monetary cost for a human to validate a single LLM-generated judgment.

Below, we provide the formulation of each procedure.

Confidence Based Procedure Formulation. Given the LLM-generated relevance judgments $\mathcal{R}$ and an upper bound ϵ for the MoE of a $1-\alpha$ CI, the confidence based procedure can be formulated as a minimization problem:

$$\begin{aligned} \text{minimize}_{\mathcal{R}_\mathcal{S}} \quad & \text{cost}(\mathcal{R}_\mathcal{S}) \\ \text{subject to} \quad & \mathbb{E}[\hat{\theta}] = \theta, \text{MoE}(\hat{\theta}, \alpha) \leq \epsilon \end{aligned} \tag{1}$$

This formulation allows to find the smallest sample $\mathcal{R}_\mathcal{S}$ that minimizes the cost for validating LLM-generated relevance judgments while satisfying the confidence constraint $\text{MoE}(\hat{\theta}, \alpha) \leq \epsilon$. Notably, the parameter ϵ controls the confidence of the estimate, inducing higher costs (i.e., a larger sample) when we want high confidence and lower costs (i.e., a smaller sample) otherwise. Furthermore, the constraint $\mathbb{E}[\hat{\theta}] = \theta$ ensures that the obtained estimate is unbiased [31], and therefore representative of the entire population $\mathcal{R}$. This procedure minimises the costs required to estimate the quality of the LLM-generated judgments with a desired confidence level.

Budget Based Procedure Formulation. Given the LLM-generated relevance judgments $\mathcal{R}$ and an upper bound b for the cost of the sample, the budget based procedure requires drawing a sample $\mathcal{R}_\mathcal{S}$ such that $\mathbb{E}[\hat{\theta}] = \theta$ and $\text{cost}(\mathcal{R}_\mathcal{S}) = b$. That is, the procedure allows drawing a sample $\mathcal{R}_\mathcal{S}$ that satisfies the constraint $\text{cost}(\mathcal{R}_\mathcal{S}) = b$, from which an unbiased estimate $\hat{\theta}$ and the corresponding $\text{MoE}(\hat{\theta}, \alpha)$ can be computed. Through this procedure, it is possible to estimate the quality of the LLM-generated relevance judgments with the highest possible level of confidence given the available budget b.

3.2.2 Sampling. The second step of the quality estimation pipeline (block B in Figure 1) involves sampling a batch of LLM-generated relevance judgments from $\mathcal{R}$. Importantly, the sampling strategy $\mathcal{S}$ must be selected with care, as **using a sampling strategy that does not align with the estimator invalidates the statistical guarantees**. In this work, we adopt SRS as sampling strategy [14] which agrees with the chosen MAE and Cohen's κ estimators.

SRS draws a batch of judgments without replacement, each selected with uniform probability from the set $\mathcal{R}$. For the confidence based procedure the batch corresponds to a single judgment, while for the budget based procedure the batch contains the number of judgments required to achieve the budget b. This difference stems from the different nature of the two procedures. While the former is an iterative process that keeps sampling judgments until the confidence constraint is satisfied, the latter represents a one-shot process that samples the highest possible number of judgments given the budget constraint.

3.2.3 Human Validation. In the third step (block C in Figure 1), the newly sampled relevance judgments are manually validated by

humans. For this task, if the LLM judgment is considered incorrect, the annotator must provide their own judgment.

Note that the human assessment follows two different schemes depending on the considered procedure. For the confidence based procedure, which works as an iterative process, the human assessors always validate the newly sampled LLM judgment – which is then added to the pool of assessed judgments. For the budget based procedure, the entire set of sampled LLM judgments is validated together.

3.2.4 Quality and Interval Estimation. In the fourth step (block D in Figure 1) we want to estimate the quality of the LLM judgments. Following the literature on the domain [21, 48, 51, 55] we consider: MAE and Cohen's κ. Since we want to minimize the costs for evaluating such quality while providing statistical guarantees, we need to i) compute an unbiased quality estimate $\hat{\theta}$, and ii) build a corresponding $1-\alpha$ CI to gauge its uncertainties. Thus, in the following, we describe the adopted MAE and Cohen's κ estimators, as well as the considered $1-\alpha$ CI.

MAE Estimator. Let us denote with $f(r) = |r - t(r)|$ the function that takes as value the absolute difference between the LLM-generated relevance judgment $r \in \mathcal{R}$ and the human-determined relevance level $t(r)$. Then, once the sample $\mathcal{R}_\mathcal{S}$ of judgments, drawn via SRS, is validated by humans, we can define an unbiased estimator $\hat{\mu}$ of the real MAE $\theta = \mu(\mathcal{R})$ as the sample proportion [31]:

$$\hat{\theta} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(r_i) \tag{2}$$

where $r_i \in \mathcal{R}_\mathcal{S}$ is the i-th sampled relevance judgment. The variance of this estimator, which corresponds to the standard error, is defined as:

$$V(\hat{\theta}) = V(\hat{\mu}) = \frac{\sum_{i=1}^{n}(r_i - t(r_i))^2}{n(n-1)} \tag{3}$$

Cohen's κ Estimator. Cohen's κ represents a measure of agreement between multiple raters [15]. In this work, we consider two raters: the LLM and the human assessor.[4] Let us define with k the number of considered relevance levels, with p_{ij} the fraction of relevance judgments whose value is set to i by the LLM and to j by the human assessor, with $p_{i.}$ the marginal probability $p_{i.} = \sum_{j=1}^{k} p_{ij}$ and with $p_{.j}$ the marginal probability $p_{.j} = \sum_{i=1}^{k} p_{ij}$. Moreover, we denote as p_o the sum $p_o = \sum_{i=1}^{k} p_{ii}$ and as p_e the sum $p_e = \sum_{i=1}^{k} p_{i.}p_{.i}$. Then, once the sample $\mathcal{R}_\mathcal{S}$ of judgments, collected with SRS, is validated by humans, we can define an unbiased estimator $\hat{\kappa}$ of the real Cohen's κ as the sample κ [44]:

$$\hat{\theta} = \hat{\kappa} = \frac{p_o - p_e}{1 - p_e} \tag{4}$$

Regarding the estimator variance, many studies have been conducted towards finding the best formulation [5, 6, 26, 28, 44]. In this work, we consider the Fleiss estimator variance [26], since it is the most popular and supports multi graded relevance. The Fleiss estimator variance is defined as:

$$V(\hat{\theta}) = V(\hat{\kappa}) = \frac{1}{n(1-p_e)^2}\left\{\sum_{i=1}^{k} p_{i.}p_{.i} \times [1 - (p_{.i} + p_{.i})]^2 + \sum_{i=1}^{k}\sum_{j=1, i\neq j}^{k} p_{i.}p_{.j}(p_{.i} + p_{j.})^2 - p_e^2\right\} \tag{5}$$

Interval Estimation. To quantify the uncertainties in the estimated quality, we adopt the Wald CI [11]. The Wald interval is obtained by inverting the Wald large-sample normal test, leading to the well-known formulation:

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{V(\hat{\theta})} \tag{6}$$

where $\hat{\theta}$ represents the estimated quality, $V(\hat{\theta})$ the estimator variance, α the considered significance level, and $z_{\alpha/2}$ the critical value of the standard normal distribution for α.

Note that, asymptotically, the larger the sample size n, the smaller the CI, and the more confident we can be about the estimate $\hat{\theta}$ of θ.

3.2.5 Constraint Check. The last step of the quality estimation pipeline (block E in Figure 1) only concerns the confidence based procedure. In this step, the MoE of the obtained $1-\alpha$ CI must be checked to determine if it satisfies the constraint or if more iterations are required to converge. The process terminates if the CI is sufficiently small, hence, as soon as the required criterion, $\text{MoE}(\hat{\theta}, \alpha) \leq \epsilon$, is met. This avoids oversampling while providing statistical guarantees.

Note, however, that the considered stopping condition may incur into some issues when the sample size is (very) small, as the variance may shrink too much. To overcome this limitation, a minimum number of assessments over the sampled relevance judgments should be obtained before activating this control step. A typical rule-of-thumb is to gather at least 30 assessments, satisfying the condition to apply the Central Limit Theorem [27].

Once the constraint is satisfied, we can report the final quality estimate and $1-\alpha$ CI. For the budget based procedure, the quality estimate and the $1-\alpha$ CI obtained in the previous step are directly returned as the final quality estimates.

4 Experimental Methodology

In this section we describe the experimental setup and how we generate and process the relevance judgments.

4.1 Experimental Setup

Collections. To evaluate the quality estimation framework, we use three different test collections: the TREC DL 2019 [19, 36] passage collection, the TREC DL 2020 [16, 36] passage collection and the TREC Robust 2004 [52, 53] document collection. The tree collections contain respectively 43, 54 and 249 queries. TREC DL collections are based on the MSMARCO passages corpus [36] which contains 8.8M passages. The TREC Robust 2004 collection relies on the TIPSTER disks 4 and 5 document corpus, minus congressual records, containing 528k documents. For readability, we refer to passages and documents as "documents".

[4] Note that here we refer to a human assessor as either a single assessor or the combination of multiple assessors, whose judgments are aggregated in some way.

Relevance judgements for TREC DL 2019 and TREC DL 2020 correspond to the following four-graded relevance labels: (0) "irrelevant"; (1) "related"; (2) "highly relevant"; and (3) "perfectly relevant". On the other hand, TREC Robust 2004 relevance judgements correspond to one of the following three labels: (0) "not relevant", (1) "relevant", and (2) "highly relevant".

We consider only query-document pairs for which the corresponding relevance judgments are available (i.e., belonging to the original pool). This allows us to simulate the annotation process of the LLM-made relevance judgments.

In terms of relevance judgements, TREC DL 2019 has approximately 9.3k annotated query-document pairs while TREC DL 2020 has around 11k annotated pairs. TREC Robust 2004 has approximately 311k annotated query-document pairs. Therefore, we employ a subset of the available relevance judgments for this collection. Specifically, we use 5% of the query-document pairs ($\sim 15k$), sampled uniformly. This stems from the non-negligible temporal and economic cost of the LLMs, and to adopt a more ethical approach towards IR research and to reduce our carbon footprint [10, 42]. Furthermore, this allows us to use relevance judgment sets with comparable sizes for all the three collections.

Large Language Models. In our experiments, we exploit the following LLMs: **Llama 3.1 405B Instruct** [23], an open LLM developed by Meta with 405 Billion parameters; **Llama-3.1-8B-Instruct** [23], an open LLM developed by Meta with 8 Billion parameters, representing a smaller version of Llama 3.1 405B Instruct; **Mistral 7B-Instruct v0.3** [32], an open LLM developed by Mistral AI with 7 Billion parameters; **Phi 3.5 mini Instruct** [2], an open LLM developed by Microsoft with 3.8 Billion parameters. We use the "Instruct" version of these LLMs as it is fine-tuned to improve their ability to follow instructions and respond to structured questions.

Other Parameters. For the confidence based procedure, we experimented both with a 95% and a 99% CI, *i.e.,* α=0.05 and α=0.01, respectively. ϵ is also set to 0.05. To have a meaningful initial estimate, as discussed in Section 3.2.5, we consider an initial sample $\mathcal{R}_S$ of 30 human-annotated relevance judgments.

For the budget based procedure, we evaluate the cases where the researcher can collect up to b human-made relevance judgements, with $b \in \{50, 100, 200, 500, 1000, 2000, 5000, 10000\}$. To provide a time estimate, we assume 1 and 4 minutes to annotate an LLM-generated query-passage and a query-document judgment, respectively. Therefore, converting the number of annotations in hours, we assume the researcher can allocate between 1 and 167 hours to annotate passages (i.e., the values of b multiplied by the time to annotate a passage), and between 3 and 667 hours for the documents.

4.2 Relevance Judgements Generation

We employ the UMBRELA [51] prompt to generate the relevance judgements. This prompt has been used to generate the relevance judgments for the TREC [46] 2024 RAG Track[5]. The prompt combines a query and a document and causes the LLM to answer with a label among: "irrelevant", "related", "highly relevant" and "perfectly relevant" (i.e., those used for TREC DL 2019 and TREC DL 2020). Thus, to generate the relevance judgements for a given LLM and a given collection, we (i) select a (previously unseen) query-document pair; (ii) customize the UMBRELA prompt for the chosen query-document pair; (iii) provide the prompt in input to the LLM; (iv) obtain the relevance judgment for the query-document pair from the LLM; and (v) halt the procedure if all the query-document pairs in the collection have been judged or repeat from (i) otherwise.

Differently from the TREC DL collections, the relevance judgments of the TREC Robust 2004 collection have three grades. Thus, in this case, we map "perfectly relevant" of UMBRELA to "highly relevant" in the collection, the "highly relevant" to "relevant" and "related" and "irrelevant" to "not relevant".

[5] https://trec-rag.github.io/

5 Experimental Evaluation

In this section, we discuss the empirical evaluation of the proposed framework. In Section 5.1 we present the MAE and Cohen's κ estimation results computed for a 95% CI and in Section 5.2 we report the MAE and the Cohen's κ estimation results for a 99% CI.

5.1 Standard Confidence Relevance Results

Our first two experiments focus on obtaining an estimate of the MAE and the Cohen's κ with a confidence of 95% (α=0.05).

MAE Estimate. The results obtained applying the proposed framework to estimate the MAE for a confidence of 95% are reported in Table 1. The row "Real MAE" contains the MAE computed by comparing all the human-made relevance judgements with the LLM-generated ones: this is the value for the entire population that we wish to estimate through sampling with the proposed procedures. For the confidence based procedure, the table shows the value of the estimate along with the related CI and the number of human-made relevance judgments (cost) used to estimate the value. For the budget based procedure, the table contains a row for each budget value with the corresponding estimate and CI.

The results for the confidence based procedure show that it is very cost-effective. Indeed, for the collections and LLMs used in our experimental setup, less than 1700 LLM generated relevance judgments must be validated by a human to obtain an MAE estimate and a $1 - \alpha$ CI for an α and ϵ values of 0.05. Thus, given the number of relevance judgments generated by the LLM for each of the collections, by exploiting the proposed confidence based procedure, the MAE estimate can be computed by considering less than 16% of the created judgments.

Concerning TREC DL 2019, we notice that the number of human-made relevance judgements required to estimate the error of the relevance judgements generated by Llama 8B with a confidence level of 5% is 1470, which decreases for the other LLMs – with LLama 405B requiring only 787 relevance judgements. These values are correlated with the MAE (both real and estimated). This pattern is in line with what was observed by Marchesin and Silvello [34]: the larger the error, the more the annotations that are needed to correctly estimate it. This can be framed as an information theory problem: the more uniform a set is, the fewer samples we need to describe it, with the perfectly uniform sets on the extremes (i.e., when we have MAE equal to 0 or equal to 3 since we consider 4 relevance levels). Nonetheless, in this case, Phi 3.5 mini represents an exception since it requires less annotations than LLama 8B even

Table 1: MAE estimation results for the confidence and budget based procedures when α=0.05

		TREC DL 2019				TREC DL 2020				TREC robust 2004			
	Model	**Llama 405B**	**Llama 8B**	**Mistral 7B**	**Phi 3.5 mini**	**Llama 405B**	**Llama 8B**	**Mistral 7B**	**Phi 3.5 mini**	**Llama 405B**	**Llama 8B**	**Mistral 7B**	**Phi 3.5 mini**
	Real MAE	0.632	1.001	0.908	1.170	0.604	0.983	0.913	1.221	0.565	0.882	0.856	1.494
conf. based	Estimate (C.I.)	0.626 ±0.050	1.022 ±0.050	0.958 ±0.050	1.197 ±0.050	0.604 ±0.050	0.962 ±0.050	0.893 ±0.050	1.207 ±0.050	0.529 ±0.050	0.853 ±0.050	0.820 ±0.050	1.453 ±0.050
	Judgments # (hours)	787 (13.1h)	1470 (24.5h)	1050 (17.5h)	1120 (18.7h)	810 (13.5h)	1446 (24.1h)	1160 (19.3h)	1205 (20.1h)	867 (57.8h)	1594 (106.2h)	724 (48.2h)	1644 (109.6h)
cost based	Judgments #:												
	50	0.580 ±0.178	1.060 ±0.293	1.100 ±0.258	1.340 ±0.221	0.660 ±0.214	1.140 ±0.269	0.900 ±0.239	1.220 ±0.258	0.420 ±0.169	0.760 ±0.283	0.840 ±0.180	1.620 ±0.285
	100	0.590 ±0.134	1.100 ±0.196	1.070 ±0.165	1.270 ±0.150	0.670 ±0.145	1.120 ±0.190	0.930 ±0.168	1.210 ±0.190	0.510 ±0.141	0.840 ±0.196	0.800 ±0.131	1.600 ±0.201
	200	0.625 ±0.096	1.070 ±0.136	1.020 ±0.115	1.260 ±0.112	0.665 ±0.107	1.025 ±0.131	0.940 ±0.122	1.215 ±0.131	0.520 ±0.105	0.855 ±0.139	0.835 ±0.097	1.510 ±0.147
	500	0.614 ±0.061	1.020 ±0.086	0.964 ±0.075	1.206 ±0.075	0.580 ±0.065	0.974 ±0.084	0.902 ±0.076	1.200 ±0.078	0.544 ±0.067	0.882 ±0.090	0.808 ±0.061	1.462 ±0.093
	1000	0.645 ±0.045	1.033 ±0.061	0.962 ±0.051	1.215 ±0.053	0.596 ±0.045	0.940 ±0.059	0.898 ±0.053	1.213 ±0.055	0.542 ±0.047	0.862 ±0.063	0.835 ±0.043	1.461 ±0.065
	2000	0.636 ±0.032	1.028 ±0.043	0.952 ±0.036	1.202 ±0.038	0.596 ±0.032	0.962 ±0.043	0.910 ±0.038	1.222 ±0.039	0.548 ±0.032	0.855 ±0.045	0.842 ±0.031	1.465 ±0.045
	5000	0.624 ±0.020	1.005 ±0.027	0.907 ±0.022	1.162 ±0.024	0.593 ±0.020	0.971 ±0.028	0.912 ±0.024	1.216 ±0.025	0.556 ±0.021	0.854 ±0.028	0.861 ±0.019	1.469 ±0.029
	10000	-	-	-	-	0.607 ±0.014	0.986 ±0.020	0.916 ±0.017	1.228 ±0.018	0.562 ±0.015	0.866 ±0.020	0.855 ±0.014	1.492 ±0.020

if its MAE is larger. This is due to the randomness of the sampling procedure that, for the case of Phi 3.5 mini, allows to sample a more uniform set earlier. The pattern repeats almost identically also for the TREC DL 2020. In such a case, to estimate the MAE of Llama 8B, we need 1446 human-made relevance judgements (~12.7% of those available in the TREC DL 2020). On the contrary, to estimate the performance of LLama 405B we need 810 human-made relevance judgements (~7.1%). The general pattern slightly changes if we consider TREC Robust 2004. In fact, in this case, Phi 3.5 mini requires the largest amount of human-made relevance judgements: 1644 (10.3% of the 15k relevance judgements that we considered for the TREC Robust 2004). Moreover, for the TREC Robust 2004 Mistral 7B is the model requiring the lowest amount of human-made judgments even if it does not have the smallest MAE. Similarly for the case of Phi 3.5 mini on the TREC DL 2019, this is due to the randomness of the sampling procedure that allows to sample a uniform set.

As expected, the results for the budget based procedure (lower part of Table 1) show that, if we increase the budget, the CIs shrink. If we consider 50 relevance judgements in Table 1, we notice that the CIs sizes are between 0.169 and 0.293 MAE points. While most of them consistently contain real MAE, confirming the validity of the framework, they are likely considered too large to provide a sufficiently precise estimate of the MAE. In line with the analysis of the confidence based procedure, when we consider around 1000 human-made relevance judgements, we obtain CIs that are between 0.045 and 0.065 MAE points. In several settings, this level of precision of the estimate might be considered sufficient and was achieved with, approximately 16 to 66 hours, depending on whether the human annotator is annotating paragraphs or documents. If we move to 5000 human-made relevance judgements, the CI sizes are between 0.019 and 0.029 MAE points. Doubling the number of human-made relevance judgements allows us to have intervals as small as 0.014 to 0.020 MAE points. That is, doubling the budget did not impact substantially the precision of our estimates. Importantly we want to stress that the interval can be made arbitrarily small with a sufficiently high budget. That is, using more and more data could induce overly narrow intervals and overpowered statistical test which would not generalise to previously unseen data. Figure 2 reports the MoE computed for each budget using different LLMs as annotators and explicitly illustrates how increasing the budget allows us to increase the confidence and reduce the variability of the estimate and the size of the CIs.[6]

Cohen's κ Estimate. The results obtained applying the proposed framework to estimate the Cohen's κ for α=0.05 are reported in Table 2. The table is structured in the same way as for the MAE (Table 1) and the row "Real Cohen's κ" contains the Cohen's κ computed by comparing all the human-made relevance judgements with the LLM-generated ones.

The effectiveness of the framework is confirmed by the confidence based procedure result. For the collections and LLMs used in our experimental setup, less than 550 LLM generated relevance judgments must be evaluated by a human to obtain a Cohen's κ estimate and a $1-\alpha$ CI for an α and ϵ values of 0.05. Thus, less than 6% of the LLM-generated judgments must be manually validated to estimate the Cohen's κ – considering the number of judgments generated by the LLMs for each of the collections.

Moreover, for the confidence based procedure, the closer the real Cohen's κ value is to 0.5, the more human annotations are required for the estimate. Indeed, the Cohen's κ takes values in the range [-1, 1], where a value of 1 indicates that there is complete agreement between the raters, a value of 0 indicates that there is no agreement.[7] Thus, a Cohen' s κ value of 0.5 represents a uniform distribution which requires more annotations to be handled [34].

The results of the confidence based procedure confirm such a behaviour. In TREC DL 2019, Llama 405B, whose Cohen's κ is the closest to 0.5 (0.288), requires the annotation of the highest number of LLM-generated judgments (512), while Phi 3.5 mini, whose Cohen's κ is the furthest from 0.5 (0.053), requires to annotate the least amount of LLM-generated judgments (268). The same occurs in the TREC DL 2020 and TREC Robust 2004.

As expected, the results for the budget based procedure (lower part of Table 2) show that, if we increase the budget, the CIs shrink. Indeed, it is possible to notice that, for a budget value b of 50, the CI sizes are between 0.069 and 0.173 Cohen's κ points, while for a value b of 500, we obtain CIs that are between 0.018 and 0.052 Cohen's κ points. Such confidence levels on the estimate might often be considered satisfying and can be achieved with approximately 8

[6]Note that, occasionally the interval might not contain the real value (e.g., TREC DL 19 with Mistral 7B and $b = 1000$). Being our procedure based on $\alpha = 0.05$ we can expect this to occur in 5% of the cases: reducing α decreases this probability at the cost of more annotations.

[7]Negative values indicate specular ratings

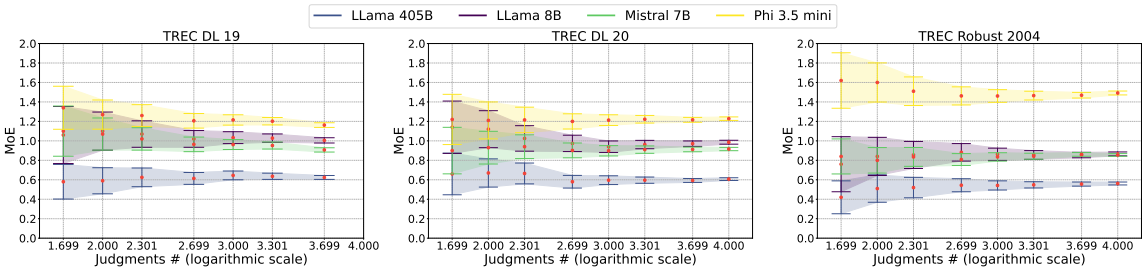


Figure 2: MAE estimation MoE computed for each budget for all the collections and LLMs considered.

Table 2: Cohen's κ estimation results for the confidence and budget based procedures when α=0.05

		TREC DL 2019				TREC DL 2020				TREC robust 2004			
	Model	**Llama 405B**	**Llama 8B**	**Mistral 7B**	**Phi 3.5 mini**	**Llama 405B**	**Llama 8B**	**Mistral 7B**	**Phi 3.5 mini**	**Llama 405B**	**Llama 8B**	**Mistral 7B**	**Phi 3.5 mini**
	Real Cohen κ	0.288	0.164	0.136	0.053	0.269	0.154	0.131	0.064	0.066	0.042	0.010	0.009
conf. based	Estimate (C.I.)	0.298 ±0.050	0.179 ±0.050	0.136 ±0.050	0.023 ±0.050	0.293 ±0.050	0.137 ±0.050	0.168 ±0.050	0.103 ±0.050	0.077 ±0.050	0.042 ±0.049	0.025 ±0.050	0.002 ±0.049
	Judgments # (hours)	521 (8.7h)	405 (6.8h)	362 (6.0h)	268 (4.5h)	541 (9.0h)	414 (6.9h)	354 (5.9h)	232 (3.9h)	330 (22h)	184 (12.3h)	164 (10.9h)	84 (5.6h)
budget based	Judgments #:												
	50	0.287 ±0.164	0.200 ±0.134	0.042 ±0.132	-0.014 ±0.100	0.225 ±0.160	0.081 ±0.135	0.170 ±0.130	0.166 ±0.092	0.160 ±0.173	0.148 ±0.128	0.015 ±0.105	0.003 ±0.069
	100	0.309 ±0.117	0.162 ±0.094	0.032 ±0.096	-0.049 ±0.080	0.221 ±0.113	0.057 ±0.100	0.140 ±0.094	0.146 ±0.075	0.134 ±0.112	0.082 ±0.083	0.038 ±0.077	0.004 ±0.046
	200	0.276 ±0.081	0.154 ±0.069	0.066 ±0.066	-0.019 ±0.058	0.238 ±0.081	0.104 ±0.072	0.159 ±0.066	0.112 ±0.055	0.092 ±0.067	0.044 ±0.050	0.015 ±0.043	0.007 ±0.029
	500	0.287 ±0.051	0.167 ±0.045	0.127 ±0.042	0.048 ±0.037	0.296 ±0.052	0.134 ±0.046	0.156 ±0.043	0.067 ±0.035	0.069 ±0.039	0.050 ±0.029	0.019 ±0.026	0.016 ±0.018
	1000	0.275 ±0.036	0.151 ±0.033	0.118 ±0.030	0.040 ±0.027	0.266 ±0.037	0.144 ±0.033	0.149 ±0.031	0.053 ±0.025	0.070 ±0.027	0.047 ±0.020	0.016 ±0.017	0.016 ±0.012
	2000	0.290 ±0.025	0.152 ±0.023	0.116 ±0.021	0.044 ±0.019	0.274 ±0.026	0.153 ±0.023	0.143 ±0.022	0.065 ±0.017	0.062 ±0.019	0.045 ±0.015	0.018 ±0.013	0.010 ±0.009
	5000	0.290 ±0.016	0.158 ±0.015	0.135 ±0.014	0.054 ±0.012	0.281 ±0.016	0.157 ±0.015	0.130 ±0.014	0.063 ±0.011	0.070 ±0.012	0.050 ±0.010	0.011 ±0.008	0.010 ±0.005
	10000	-	-	-	-	0.266 ±0.012	0.151 ±0.010	0.130 ±0.010	0.062 ±0.008	0.070 ±0.009	0.045 ±0.007	0.011 ±0.006	0.009 ±0.004

to 33 hours, depending on whether the human annotator validates passages or documents. When we focus on budget values of 5000 and 10000, the CI sizes are between 0.005 and 0.016 and between 0.004 to 0.012 Cohen's κ points, respectively. Thus, for high budget levels, doubling the number of human-made relevance judgements did not substantially impact the precision of our estimates. As before, being our procedure statistical with an $\alpha = 0.05$, the CI can occasionally miss the real Cohen's κ value – as with Phi 3.5 mini in the TREC DL 20 collection ($b = 50$). However, it is also important to notice that Cohen's κ variance estimators (especially lower one-sided ones) are known in literature [44] to be less reliable when Cohen's κ has values closer to or lower than 0.

By comparing the results of the confidence and budget based procedures, some singular behaviours can be observed. For instance, when considering the LLama 8B model and the TREC Robust 2004 collection, the confidence based procedure has a MoE of 0.049 and requires to evaluate 184 judgments, whereas the budget based estimate results in a MoE of 0.050 when the budget b is set to 200. Although in most of the cases evaluating more samples corresponds to a gain in confidence, occasionally, it may happen that larger samples contain several outliers – thus making the procedure less stable and forcing the MoE to increase. Nevertheless, this behaviour is very limited and has a negligible impact in our results.

5.2 High Confidence Relevance Results

Our third and fourth experiments focus on obtaining an estimate of the MAE and the Cohen's κ with a confidence of 99% (α=0.01).

MAE Estimate. The results obtained applying the proposed framework to estimate the MAE for α=0.01 are reported in Table 3 (structured as the one reported for α=0.05, Table 1).

The general behaviour of the framework is completely consistent with what discussed in Section 5.1, when considering α=0.05. Nonetheless, for the confidence based procedure, to reach a higher confidence on the estimate it is necessary to sample more judgments. Indeed, if we consider the TREC DL 2019 and LLama 405B, 1416 judgments must be sampled to estimate the MAE for α=0.01, while 787 (~55%) are sufficient for α=0.05. However, given the number of relevance judgments generated by the LLM for each of the collections, by exploiting the proposed confidence based procedure, the MAE estimate can be computed by considering less than 27% of the created judgments.

For the cost based procedure,instead, a comparison of the MAE estimation results for α=0.05 and α=0.01, shows that given a model, collection and cost threshold the CIs are larger when the required confidence is higher. This behaviour is consistent since, fixed a certain judgments number, to be more confident on the estimate it is necessary to increase the size of the CI.

Finally, increasing the confidence increases the reliability of the CIs. Indeed, all of the estimated CIs contain the real MAE value, both for the confidence and cost based estimates. This highlights the robustness and reliability of the procedures.

Cohen's κ Estimate. When α=0.01 the same considerations made for MAE apply for the estimation results of Cohen's κ. This highlights the reliability of the proposed framework in different estimation environments and its independence from the used quality metric. For the confidence based procedure, compared to when the

Table 3: MAE estimation results for the confidence and budget based procedures when α=0.01

		TREC DL 2019				TREC DL 2020				TREC robust 2004			
	Model	**Llama 405B**	**Llama 8B**	**Mistral 7B**	**Phi 3.5 mini**	**Llama 405B**	**Llama 8B**	**Mistral 7B**	**Phi 3.5 mini**	**Llama 405B**	**Llama 8B**	**Mistral 7B**	**Phi 3.5 mini**
	Real MAE	0.632	1.001	0.908	1.170	0.604	0.983	0.913	1.221	0.565	0.882	0.856	1.494
conf. based	Estimate (C.I.)	0.643 ±0.050	1.012 ±0.050	0.947 ±0.050	1.206 ±0.050	0.591 ±0.050	0.957 ±0.050	0.910 ±0.050	1.219 ±0.050	0.555 ±0.050	0.858 ±0.050	0.827 ±0.050	1.478 ±0.050
	Judgments # (hours)	1416 (23.6h)	2497 (41.6h)	1809 (30.2h)	1971 (32.8h)	1364 (22.7h)	2546 (42.4h)	2008 (33.4h)	2106 (35.1h)	1484 (98.9h)	2765 (184.3h)	1298	2822 (192.1h)
budget based	Judgments #:												
	50	0.580 ±0.234	1.060 ±0.385	1.100 ±0.339	1.340 ±0.291	0.660 ±0.281	1.140 ±0.353	0.900 ±0.314	1.220 ±0.340	0.420 ±0.222	0.760 ±0.372	0.840 ±0.237	1.620 ±0.374
	100	0.590 ±0.176	1.100 ±0.258	1.070 ±0.217	1.270 ±0.197	0.670 ±0.190	1.120 ±0.249	0.930 ±0.220	1.210 ±0.249	0.510 ±0.185	0.840 ±0.258	0.800 ±0.172	1.600 ±0.264
	200	0.625 ±0.126	1.070 ±0.178	1.020 ±0.152	1.260 ±0.147	0.665 ±0.141	1.025 ±0.173	0.940 ±0.160	1.215 ±0.172	0.520 ±0.138	0.855 ±0.182	0.835 ±0.128	1.510 ±0.193
	500	0.614 ±0.080	1.020 ±0.112	0.964 ±0.098	1.206 ±0.099	0.580 ±0.085	0.974 ±0.110	0.902 ±0.100	1.200 ±0.102	0.544 ±0.088	0.882 ±0.118	0.808 ±0.080	1.462 ±0.122
	1000	0.645 ±0.059	1.033 ±0.080	0.962 ±0.068	1.215 ±0.070	0.596 ±0.059	0.940 ±0.077	0.898 ±0.070	1.213 ±0.072	0.542 ±0.061	0.862 ±0.083	0.835 ±0.057	1.461 ±0.085
	2000	0.636 ±0.042	1.028 ±0.056	0.952 ±0.047	1.202 ±0.050	0.596 ±0.042	0.962 ±0.056	0.910 ±0.050	1.222 ±0.051	0.548 ±0.043	0.855 ±0.059	0.842 ±0.040	1.465 ±0.059
	5000	0.624 ±0.026	1.005 ±0.035	0.907 ±0.029	1.162 ±0.031	0.593 ±0.027	0.971 ±0.036	0.912 ±0.032	1.216 ±0.033	0.556 ±0.028	0.854 ±0.037	0.861 ±0.026	1.469 ±0.038
	10000	-	-	-	-	0.607 ±0.019	0.986 ±0.026	0.916 ±0.022	1.228 ±0.023	0.562 ±0.020	0.866 ±0.026	0.855 ±0.018	1.492 ±0.027

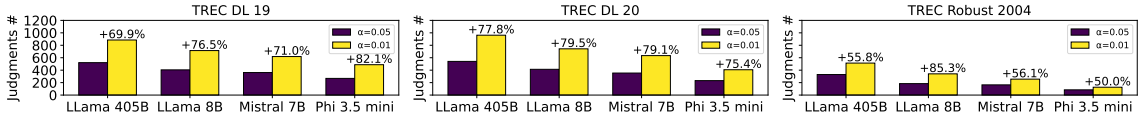


Figure 3: Number of judgments required for the confidence-based Cohen's κ estimate with $\alpha = 0.05$ and $\alpha = 0.01$.

confidence is set to 95%, the number of judgments to be validated grows. Figure 3 shows the increment in the number of validations needed for each collection and model. However, this number is still very limited and corresponds to less than 10% of the LLM-Generated relevance judgments. For the cost based procedure, given a certain model, collection and cost threshold, the CI grows. Finally, increasing the confidence results in the computed CI always containing real Cohen's κ value. This underlines the reliability of the framework. The table reporting the complete results is available online.[8]

6 Challenges and Opportunities

The validity of LLM-generated relevance judgements is a topic under debate, as such judgments may be affected by biases or circularity issues [12, 24]. In light of this, we stress that demonstrating whether and how LLMs should be used as assessors is well beyond the scope of this paper. Nevertheless, assuming that LLMs will eventually be integrated systematically in the creation of relevance judgments for test collections, we further stress the importance of cost-effective and reliable solutions to evaluate their quality. Furthermore, as a guideline for future collections involving LLM-generated relevance judgments,we recommend releasing also the prompt used to generate them, its effectiveness on historical collections, and the estimate of the quality of the judgement computed following the procedure described in this work. This documentation could act as a "reliability badge" in a similar spirit to the one introduced by Webber et al. [54] to make systems' performance comparable within and between test collections. If we establish proper releasing and sharing practices from the outset, we can ensure consistency and make the process more future-proof.

On a different note, the proposed framework employs SRS as sampling strategy, which is a relatively simple solution. However, using more sophisticated approaches, like stratified sampling, would require the definition of appropriate strata, a challenging step in this context that, if done inappropriately, might increase the variance instead of reducing it [31]. Moreover, it is essential to use appropriate estimators to provide statistical guarantees on the quality estimates. As described above, the estimators depend on the sampling strategy. For several measures, especially those that can be computed on a downstream task of the relevance judgements, such estimators have not been discovered yet. Hence, the study of more sophisticated sampling strategies and estimators for IR measures represents a critical direction for future research. Finally, the proposed framework can be adapted to operate in many other contexts, such as hybrid scenarios where human assessors are supported by LLMs – e.g., as summarizers or fact checkers – or where LLMs work as pre-assessors [24],thus emphasizing the generality of our proposal.

7 Conclusions

In this paper we introduced a new framework to estimate the quality of the relevance judgments produced by an LLM while providing statistical guarantees. We analysed two different facets: (1) computing an estimate of the quality with a fixed confidence while minimizing the cost, and (2) computing an estimate of the quality, along with the confidence of the estimate, when the budget is fixed. For each of these facets we introduced a new procedure to estimate the MAE of the LLM and the inter-assessor agreement between the LLM and the humans. We showed that the proposed procedures allow to estimate the quality of the relevance judgments generated by LLMs independently from the collection considered. Furthermore, the results highlighted that it is possible to strongly limit the human involvement required for the estimation while maintaining strong statistical guarantees.

Acknowledgments

The work was supported by the HEREDITARY project, as part of the EU Horizon Europe program under Grant Agreement 101137074. We acknowledge support from CAMEO, PRIN 2022 n. 2022ZLL7MW.

[8] https://github.com/MerloSimone/LLMQualityEstimation

GenAI Usage Disclosure

Following ACM's guidelines on the use of generative AI tools, we disclose that this work generative AI was employed only for grammar checking purposes. The authors carried out and reviewed every research idea, experiment and analysis. The scientific content and the creative o have not been significantly rewritten or generated using generative AI tools.

References

[1] 2024. *SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA). Association for Computing Machinery, New York, NY, USA.

[2] Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *CoRR* abs/2404.14219 (2024). arXiv:2404.14219 doi:10.48550/ARXIV.2404.14219

[3] 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 *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2024, Tokyo, Japan, December 9-12, 2024*, Tetsuya Sakai, Emi Ishita, Hiroaki Ohshima, Faegheh Hasibi, Jiaxin Mao, and Joemon M. Jose (Eds.). ACM, 32–41. doi:10.1145/3673791.3698431

[4] James Allan, Donna Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen M. Voorhees. 2017. TREC 2017 Common Core Track Overview. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017 (NIST Special Publication, Vol. 500-324)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec26/papers/Overview-CC.pdf

[5] Nicole J.-M. Blackman and John J. Koval. 2000. Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine* 19, 5 (2000), 723–741. doi:10.1002/(SICI)1097-0258(20000315)19:5<723::AID-SIM379>3.0.CO;2-A

[6] Daniel A. Bloch and Helena Chmura Kraemer. 1989. 2 x 2 Kappa Coefficients: Measures of Agreement or Association. *Biometrics* 45, 1 (1989), 269–287. http://www.jstor.org/stable/2532052

[7] Nuo Chen, Jiqun Liu, Xiaoyu Dong, Qijiong Liu, Tetsuya Sakai, and Xiao-Ming Wu. 2024. AI Can Be Cognitively Biased: An Exploratory Study on Threshold Priming in LLM-Based Batch Relevance Assessment. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2024, Tokyo, Japan, December 9-12, 2024*, Tetsuya Sakai, Emi Ishita, Hiroaki Ohshima, Faegheh Hasibi, Jiaxin Mao, and Joemon M. Jose (Eds.). ACM, 54–63. doi:10.1145/3673791.3698420

[8] Qi Chen, Xiubo Geng, Corby Rosset, Carolyn Buractaon, Jingwen Lu, Tao Shen, Kun Zhou, Chenyan Xiong, Yeyun Gong, Paul N. Bennett, Nick Craswell, Xing Xie, Fan Yang, Bryan Tower, Nikhil Rao, Anlei Dong, Wenqi Jiang, Zheng Liu, Mingqin Li, Chuanjie Liu, Zengzhong Li, Rangan Majumder, Jennifer Neville, Andy Oakley, Knut Magne Risvik, Harsha Vardhan Simhadri, Manik Varma, Yujing Wang, Linjun Yang, Mao Yang, and Ce Zhang. 2024. MS MARCO Web Search: A Large-scale Information-rich Web Dataset with Millions of Real Click Labels. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw (Eds.). ACM, 292–301. doi:10.1145/3589335.3648327

[9] David Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 15607–15631. doi:10.18653/V1/2023.ACL-LONG.870

[10] Gobinda Chowdhury. 2012. An agenda for green information retrieval research. *Inf. Process. Manag.* 48, 6 (2012), 1067–1077. doi:10.1016/J.IPM.2012.02.003

[11] Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data Cleaning: Overview and Emerging Challenges. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, Fatma Özcan, Georgia Koutrika, and Sam Madden (Eds.). ACM, 2201–2206. doi:10.1145/2882903.2912574

[12] 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

[13] Cyril Cleverdon. 1997. *The Cranfield tests on index language devices.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 47–59.

[14] William G. Cochran. 1977. *Sampling Techniques, 3rd Edition.* John Wiley.

[15] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. arXiv:https://doi.org/10.1177/001316446002000104 doi:10.1177/001316446002000104

[16] 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). https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.DL.pdf

[17] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. Overview of the TREC 2021 Deep Learning Track. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021 (NIST Special Publication, Vol. 500-335)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec30/papers/Overview-DL.pdf

[18] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2022. Overview of the TREC 2022 Deep Learning Track. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022 (NIST Special Publication, Vol. 500-338)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec31/papers/Overview_deep.pdf

[19] 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 https://arxiv.org/abs/2003.07820

[20] 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 *The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, November 14-17, 2023 (NIST Special Publication, Vol. 500-xxx)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec32/papers/Overview_deep.pdf

[21] 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. https://ceur-ws.org/Vol-3752/paper2.pdf

[22] Laura Dietz, Oleg Zendel, Peter Bailey, Charles Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. LLM-Evaluation Tropes: Perspectives on the Validity of LLM-Evaluations. arXiv:2504.19076 [cs.IR] https://arxiv.org/abs/2504.19076

[23] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024). arXiv:2407.21783

doi:10.48550/ARXIV.2407.21783

[24] 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

[25] Nicola Ferro and Carol Peters (Eds.). 2019. *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. The Information Retrieval Series, Vol. 41. Springer. doi:10.1007/978-3-030-22948-1

[26] Joseph L. Fleiss, Jacob Cohen, and B. S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 72 (1969), 323–327. https://api.semanticscholar.org/CorpusID:85541244

[27] Junyang Gao, Xian Li, Yifan Ethan Xu, Bunyamin Sisman, Xin Luna Dong, and Jun Yang. 2019. Efficient Knowledge Graph Accuracy Evaluation. *Proc. VLDB Endow.* 12, 11 (2019), 1679–1691. doi:10.14778/3342263.3342642

[28] J. Barry Garner. 1991. The standard error of Cohen's Kappa. *Statistics in Medicine* 10, 5 (1991), 767–775. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780100512 doi:10.1002/sim.4780100512

[29] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *CoRR* abs/2303.15056 (2023). arXiv:2303.15056 doi:10.48550/ARXIV.2303.15056

[30] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar (Eds.). Association for Computational Linguistics, 165–190. doi:10.18653/V1/2024.NAACL-INDUSTRY.15

[31] Robert V. Hogg, Dale L. Zimmerman, and Elliot A. Tanis. 2015. *Probability and statistical inference* (ninth edition. global edition ed.). Pearson, Boston. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1419274

[32] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR* abs/2310.06825 (2023). arXiv:2310.06825 doi:10.48550/ARXIV.2310.06825

[33] 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

[34] Stefano Marchesin and Gianmaria Silvello. 2024. Efficient and Reliable Estimation of Knowledge Graph Accuracy. *Proc. VLDB Endow.* 17, 9 (2024), 2392–2404. doi:10.14778/3665844.3665865

[35] Jack McKechnie, Graham McDonald, and Craig Macdonald. 2025. Context Example Selection for LLM Generated Relevance Assessments. 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, 293–309. doi:10.1007/978-3-031-88708-6_19

[36] 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

[37] Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Ragnarök: A Reusable RAG Framework and Baselines for TREC 2024 Retrieval-Augmented Generation Track. 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, 132–148. doi:10.1007/978-3-031-88708-6_9

[38] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. 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 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 3040–3043. doi:10.1145/3626772.3657992

[39] Hossein A. Rahmani, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. 2024. JudgeBlender: Ensembling Judgments for Automatic Relevance Assessment. *CoRR* abs/2412.13268 (2024). arXiv:2412.13268 doi:10.48550/ARXIV.2412.13268

[40] 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. https://ceur-ws.org/Vol-3752/paper8.pdf

[41] Tetsuya Sakai, Tetsuya. Sakai, Douglas W. Oard, and Noriko. Kando. 2021 - 2021. *Evaluating information retrieval and access tasks: NTCIR's legacy of research impact.* Springer Nature, Singapore.

[42] Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. 2022. Reduce, Reuse, Recycle: Green Information Retrieval Research. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2825–2837. doi:10.1145/3477495.3531766

[43] 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. In *Proceedings of The Web Conference (WebConf '25)*.

[44] Guogen Shan and Weizhen Wang. 2017. Exact one-sided confidence limits for Cohen's kappa as a measurement of agreement. *Statistical Methods in Medical Research* 26, 2 (2017), 615–632. arXiv:https://doi.org/10.1177/0962280214552881 doi:10.1177/0962280214552881 PMID: 25288510.

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

[46] Nicola Stokes. 2006. *TREC: Experiment and Evaluation in Information Retrieval* Ellen M. Voorhees and Donna K. Harman (editors) (National Institute of Standards and Technology), Cambridge, MA: The MIT Press (Digital libraries and electronic publishing series, edited by William Y. Arms), 2005, x+462 pp; hardbound, ISBN 0-262-22073-3. *Comput. Linguistics* 32, 4 (2006), 563–567. doi:10.1162/COLI.2006.32.4.563

[47] Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Support Evaluation for the TREC 2024 RAG Track: Comparing Human versus LLM Judges. arXiv:2504.15205 [cs.CL] https://arxiv.org/abs/2504.15205

[48] 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

[49] Mehmet Deniz Türkmen, Mucahid Kutlu, Bahadir Altun, and Gokalp Cosgun. 2025. GenTREC: The First Test Collection Generated by Large Language Models for Evaluating Information Retrieval Systems. arXiv:2501.02408 [cs.IR] https://arxiv.org/abs/2501.02408

[50] Shivani Upadhyay, Ehsan Kamalloo, and Jimmy Lin. 2024. LLMs Can Patch Up Missing Relevance Judgments in Evaluation. *CoRR* abs/2405.04727 (2024). arXiv:2405.04727 doi:10.48550/ARXIV.2405.04727

[51] 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). arXiv:2406.06519 doi:10.48550/ARXIV.2406.06519

[52] Ellen Voorhees. 2004. Overview of the TREC 2004 Robust Retrieval Track. In *TREC*.

[53] Ellen M. Voorhees. 1996. NIST TREC Disks 4 and 5: Retrieval Test Collections Document Set. doi:10.18434/t47g6m

[54] William Webber, Alistair Moffat, and Justin Zobel. 2008. Score standardization for inter-collection comparison of retrieval systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong (Eds.). ACM, 51–58. doi:10.1145/1390334.1390346

[55] 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. https://ceur-ws.org/Vol-3752/paper7.pdf

[56] Yiming Zhu, Peixian Zhang, Ehsan ul Haq, Pan Hui, and Gareth Tyson. 2023. Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks. *CoRR* abs/2304.10145 (2023). arXiv:2304.10145 doi:10.48550/ARXIV.2304.10145