Reducing Human Effort to Validate LLM Relevance Judgements via Stratified Sampling

Simone Merlo[0009−0003−8003−4795], Stefano Marchesin[0000−0003−0362−5893], Guglielmo Faggioli[0000−0002−5070−2049], and Nicola Ferro[0000−0001−9219−6239]

University of Padua

Abstract. Information Retrieval (IR) evaluation deeply relies on human-made relevance judgments. To overcome the high costs of the judgment collection process, a potential solution is to utilize LLMs as judges to replace human annotators. However, the validation of LLM-generated judgments is fundamental for informed use. Standard validation approaches typically rely on simple sampling techniques to collect a sample of the LLM-generated judgments and estimate the LLM agreement with the human. In this work, we propose using stratified sampling, a more sophisticated sampling strategy that, by leveraging appropriate stratification features, reduces human involvement in the validation process while still providing statistical guarantees on the human-LLM agreement estimate. Through the analysis of various candidate features, we identify the LLM-generated judgments themselves as the most promising one. Our approach achieves up to an 85% reduction in the required human involvement in the validation process.

Keywords: Relevance Judgments · Large Language Models · Agreement Estimation

1 Introduction

Information Retrieval (IR) evaluation is deeply rooted in empirical experimentation that employs evaluation collections to gauge the effectiveness of the systems. In detail, a standard IR evaluation collection, built according to the Cranfield paradigm [7], includes three elements: a set of topics to represent possible information needs, a corpus of documents, and the relevance judgements, describing which documents are relevant to each topic. Constructing high-quality, realistically sized test collections is both time-consuming and expensive, primarily because of the need for manually created relevance labels. Adding to this challenge is the dynamic nature of topics, documents, and their relevance; continuously updated annotated data is essential for reliably evaluating an IR system. The recent advances in the development and widespread adoption of Large Language Models (LLMs) offer the potential to reduce the cost of building IR experimental collections by assisting humans in the annotation process—or even replacing them in certain scenarios [16, 18, 23, 29, 30]. While an active debate is ongoing within the research community about whether LLMs should [1, 15, 30] or should not [6, 20, 25] be used for relevance assessment, we argue that LLMs are likely to be increasingly integrated in the construction of experimental collections—at least for specific tasks or to replace less reliable and imprecise annotators, such as crowd-workers [11].

Determining whether or not LLMs are suitable and reliable relevance annotators is far beyond the scope of this paper. Here, in the same spirit of Merlo et al. [19], we are interested in determining how to measure the agreement between human-made and LLM-generated relevance judgements and how to compute strong statistical bounds for such a measurement, while minimizing the human effort required. In detail, Merlo et al. [19] demonstrated that by employing an iterative procedure and an estimation pipeline based on the statistical properties of sampling, it is possible to use a few hundred human-made query-document relevance judgments to estimate the alignment between LLM labels and human annotations, while also constructing meaningful confidence intervals around this estimate. However, their approach relied exclusively on Simple Random Sampling (SRS), a reliable and stable sampling method whose downside is that it often requires a larger number of samples to achieve convergence. Advanced sampling methods, such as Stratified Sampling (SS), have the potential to reduce the number of samples required to achieve the same confidence level in the obtained estimate [8, 17]. To fully realize this potential, however, it is necessary to partition the population—relevance judgements in our case—into meaningful strata. In this work, we define and compare possible stratification features that allow us to organize LLM-generated relevance judgments into strata. This enables us to employ the SS technique, reducing the number of LLM-generated relevance judgments that need to be validated by humans to estimate the level of agreement, compared to existing approaches. From this perspective, we articulate our work into two research questions:

- **RQ1 - Stratification Features**: What are the most reliable stratification features for LLM-generated relevance judgments?
- **RQ2 - Impact of Stratified Sampling on the Cost**: Does SS allow for reducing the number of human-made relevance judgments required to validate the LLM-generated ones?

Our empirical results show that adopting SS and appropriate estimators for Mean Absolute Error (MAE) and Cohen's κ allows to reduce the amount of LLM-generated relevance judgments that must be validated by humans up to 85%, compared to the standard SRS technique [19], to achieve a 95% confidence on the estimated measure of interest.

The rest of this work is organized as follows: Section 2 reviews current approaches for LLM-generated judgment validation; Section 3 and 4 describes the SS-based estimation pipeline and the experimental methodology and setup; Section 5 discusses the obtained results; Finally, Section 6 provides the final remarks.

2 Related Work

The recent improvements of LLMs reasoning capabilities favored their adoption in many research fields. In IR, the creation of relevance judgments represents one of the most impacted processes. Indeed, to reduce the costs derived from human work, LLMs started to be employed to support or even replace humans in the assessment of the query-document pairs. Thomas et al. [29] first proposed a methodology and some prompts to allow for the generation of judgments using LLMs. Upadhyay et al. [30] refined the work of Thomas et al. [29], introducing UMBRELA, a prompt that has been extensively used [14, 22] and studied [13]. Moreover, workshops and challenges started to be organized, among which the "LLM4Eval" workshop [21, 24].

Nonetheless, IR researchers hold mixed opinions on whether [1, 15] or not [6, 20, 25] LLMs should be employed in the judgements generation process. In this perspective, several studies are being performed trying to identify potential LLM biases or strengths [3, 4, 11, 14, 28, 33]

Besides the reliability of LLMs, the generated judgments must be properly validated to make an informed use. Existing work estimate the quality of LLM-generated relevance judgments based on a pre-defined sample of query-document pairs [11, 27, 29]. However, most of them focus on evaluating the LLM-human agreement on test collections for which human judgments are available and, thus, an appropriate size of the sample can be defined in advance. The only work that focuses on minimizing the human involvement in the validation process is represented by [19]. This approach exploits an iterative sampling procedure that, relying on SRS, first draws a sample of LLM-generated judgments to be validated by humans and then estimates the human-LLM agreement. Moreover, through the use of appropriate estimators, it also allows to provide statistical guarantees on the computed agreement estimates. Nonetheless, statistical literature has long established that more advanced sampling techniques can further reduce the sample size required to achieve a specified confidence level in the estimate [8]. For this reason, in this work we propose a novel agreement estimation pipeline that leverages the SS technique to reduce the human effort required to validate LLM-generated judgments.

3 The Stratified Sampling Pipeline

Our objective is to evaluate the extent the relevance judgments produced by an LLM agree with those made by humans, while minimizing the requirement for human annotations. Following Merlo et al. [19], we use an estimation pipeline that employs an iterative sampling procedure to achieve this result. Unlike prior work [19], we use SS instead of SRS for sampling.

In detail, we assume to have a population of relevance judgments produced by an LLM, called $\mathcal{R}$, whose size is N. At each step of the iterative procedure, we sample a relevance judgment r and add it incrementally to the sample R_{SS}. For each sampled LLM-generated relevance judgment r, we assume to have the human-made ground truth label $t(r)$. By taking θ as the real quality value and $\hat{\theta}$ as the quality estimate, it is possible to define the human-LLM agreement evaluation as a minimization problem:

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

The goal of this minimization problem is to estimate the human-LLM agreement on a sample $\mathcal{R}_{SS}$ of $\mathcal{R}$, obtained by applying the SS technique, while minimizing the cost (cost($\mathcal{R}_{SS}$)) and guaranteeing that the constraints set on the confidence of the estimate are satisfied – i.e., the Margin of Error (MoE) is below the threshold ϵ. In this work, we set the annotation cost cost($\mathcal{R}_{SS}$) to the number of LLM-generated judgments that are sampled and need to be validated by humans. For the estimation to be valid, the employed estimators $\hat{\theta}$ of θ must be unbiased ($\mathbb{E}[\hat{\theta}] = \theta$). In practice, we address this task through the following iterative pipeline:

Step 0 - Stratify the population: we use one or more stratification variables to partition the population into non-overlapping strata. This step is done beforehand, and once fixed, the stratification cannot be modified.

Step 1 - Sampling: we sample a stratum proportionally to its weight. From that stratum, we uniformly sample an LLM-generated relevance judgment and add it to the global sample R_{SS}.

Step 2 - Estimation of the statistics of interest: given the sampled LLM-generated relevance judgments and the corresponding human-made counterparts, we employ estimators tailored for SS to compute the target statistics, in our case the MAE and Cohen's κ, together with their variance.

Step 3 - Constraint check: using the estimation obtained from the previous step, we compute the Confidence Interval (CI) and evaluate if we have reached the required level of confidence. If this is the case, the procedure is interrupted; otherwise, we repeat the procedure iteratively from step 1.

Below, we describe each of these steps in detail.

3.1 Step 0 - Stratify the Population

SS relies on stratifying the population $\mathcal{R}$ into a set of H non-overlapping strata $\mathcal{P} = \{p_1, p_2, ..., p_H\}$, such that $\mathcal{P}$ represents a partition of the sample space [8]. If the stratification is effective, SS usually requires fewer samples than SRS to achieve the desired confidence level on the estimate. However, the effectiveness of a stratification depends on two factors: (i) how individuals are allocated across strata, and (ii) the number of strata H. Regarding (i), a stratification is more effective the more "uniform" the individuals are in each stratum. As for (ii), while increasing the number of strata likely increases the internal homogeneity, it also increases the number of samples required to converge.

Stratification Features Our first task is to find the feature, or the combination of features, that maximizes the per-stratum homogeneity. In this work, we investigate the following stratification features:

- Assigned Labels (A): the LLM-generated relevance judgments;
- Probability (P): the probability computed by the LLM in the output layer that the token corresponding to the relevance label is the next one;
- Probability Delta (ΔP): the difference between the probabilities in the output layer of the LLM for tokens corresponding to the most likely and least likely labels;
- Probability Delta Second (ΔP_2): similar to ΔP, but considering the most and second-most likely labels;
- Perplexity (PPL): the perplexity of the output generated by the LLM. Perplexity can be computed as: $PPL(X) = \exp\left(-\frac{1}{|X|}\sum_{i=1}^{|X|} \log P(x_i \mid x_1^{i-1})\right)$, where X represents the LLM output and x_i denotes its i-th token.

To limit the influence of the stratification strategy, we either partition the data according to their assigned labels, as in the case of feature A, or employ k-means to derive H strata. Exploring the impact of other stratification strategies is left for future work.

3.2 Step 1 - Sampling

The second step of the pipeline is articulated in two sampling phases. The first phase requires sampling one of the strata. Each h-th stratum has a weight defined as $W_h = \frac{N_h}{N}$, where N_h represents the size of the stratum. According to SS, the strata are randomly sampled with a probability that is proportional to their weight W_h. This guarantees that, at the end of the process, the relevance judgments in $\mathcal{R}_{SS}$ are distributed across strata proportionally to their weights.

The second phase, instead, requires uniformly sampling the relevance judgments from the stratum selected during the previous phase. Within each stratum, we employ SRS as sampling technique; that is, a single LLM-generated relevance judgment is selected among those in the stratum sampled during the first phase.

3.3 Step 2 - Estimation of the Statistics of Interest

After sampling, it is necessary to compute both the point estimate of the target statistics and its variance using an appropriate estimator. Applying an incorrect estimator—e.g., an estimator assuming SRS instead of SS as the underlying sampling strategy—can yield misleading results, such as biased estimates or invalid CIs. Since we rely on SS to draw data, it is essential to adopt estimators suitable for this sampling strategy.

In this work, we assess the human-LLM agreement using two measures: Mean Absolute Error (MAE), defined as the mean absolute difference between LLM-generated and human-made relevance labels, and Cohen's κ. Below, we describe the corresponding SS-based estimators for each statistics.

MAE Estimator under Stratified Sampling. When the statistics we want to compute can be represented as a mean, as in the MAE case, an unbiased point estimate under SS is given by:

$$\hat{\theta} = \sum_{h=1}^{H} W_h \hat{\theta}_h, \tag{2}$$

where $\hat{\theta}_h$ represents the MAE estimate computed on the samples $r \in \mathcal{R}_{SS} \cap p_h$ from the h-th stratum. Specifically, we define the per-stratum MAE as $\hat{\theta}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} f(r_i)$, where $f(r) = |r - t(r)|$ and n_h is the number of judgements drawn from the h-th stratum.

Accordingly, the estimation variance is computed as:

$$V(\hat{\theta}) = \sum_{h=1}^{H} W_h^2 V(\hat{\theta}_h), \tag{3}$$

where $V(\hat{\theta}_h) = \frac{\sum_{i=1}^{n_h} (r_i - t(r_i))^2}{n_h(n_h - 1)}$.

Cohen's κ Estimator under Stratified Sampling. Cohen's κ cannot be represented as a mean, thus an ad hoc estimator has been defined by Stehman [26]. Let us define M as the confusion matrix where the i-th, j-th cell contains the number $n_{i,j}$ of query-document pairs in the sample for which the LLM and the human assigned respectively the labels i and j. Then, according to [26], the Cohen's κ point estimator can be defined as:

$$\hat{\theta} = \frac{N \cdot \hat{D} - \hat{C}}{N^2 - \widehat{C}}, \tag{4}$$

where $\hat{C} = \sum_{j=1}^{H} N_j \cdot \hat{M}_j$ with $\hat{M}_j = \sum_{h=1}^{H} \frac{N_h}{n_h} n_{hj}$, and $\hat{D} = \sum_{h=1}^{H} \frac{N_h}{n_h} n_{hh}$.

The corresponding estimation variance can be defined as:

$$V(\hat{\theta}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{V}_h}{n_h}, \tag{5}$$

where $\hat{V}_h = \frac{\sum_{i=1}^{n_h} \hat{u}_{hi}^2 - n_h \bar{u}_h^2}{n_h - 1}$, $\bar{u}_h = \sum_{i=1}^{n_h} \frac{\hat{u}_{hi}^2}{n_h}$, $\sum_{i=1}^{n_h} \hat{u}_{hi} = n_{hh}B$, $\sum_{i=1}^{n_h} \hat{u}_{hi} = n_{hh}B^2$ and $B = [\frac{N}{N^2 - \hat{C}} + N_h \frac{N\,(\hat{D} - N)}{(N^2 - \hat{C})^2}]$.

It is important to note that the estimator proposed by Stehman works only when the stratification is based on the assigned label (i.e., the LLM-generated relevance judgment). To the best of our knowledge, no SS estimator for Cohens κ currently exists that accommodates arbitrary stratification schemes. The development of such an estimator remains an open direction for future work.

3.4 Step 3 - Constraint check

Once the sample estimate and its variance have been produced, we can use them to compute the CI. In this work, we use the Wald CI [5], which is defined as:

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

where $z_{\alpha/2}$ is the critical value of the standard normal distribution for the significance level α. The Wald CI allows for quantifying the uncertainties on the estimated agreement, thus a larger CI corresponds to a lower confidence level on the estimate.

The final step of the pipeline evaluates whether the target level of confidence in the estimate has been achieved. Specifically, the MoE—i.e., half the width of the CI—is compared against the predefined threshold ϵ. If the MoE falls below ϵ, the procedure halts; otherwise, it loops back to the sampling step (Step 1).

4 Experimental Methodology and Setup

4.1 Experimental Methodology to answer RQ1

Our first research question revolves around determining what are the optimal stratification features to measure the effectiveness of LLMs as assessors. As mentioned in Section 3.1, the more a stratification variable produces uniform strata, the better it is. Furthermore, in our case, the ideal stratification corresponds to perfectly separating correct and wrong LLM-generated relevance judgments—or partitioning them by the size of the error if we consider a multi-graded scenario. In other terms, we are interested in finding partitioning features, the covariates X, that allow us to predict whether the label assigned by the LLM is the same as the one assigned by the human, our response Y. In practical terms, to measure the quality of the features, we fit a logistic regression using each of the features identified above individually. For example, $Y = A$ is the model where we use the label generated by the LLM to predict if the LLM label was correct; similarly, $Y = PPL$ is the model where we try to predict the correctness of the label using the perplexity score. Notice that, while the label assigned by the LLM is a categorical value (our variable A), all the other variables are continuous.

Besides assessing individual features, we also test more complex models, combining multiple variables at once. For instance, $Y = A + P$ is the logistic regression model where the label assigned by the LLM and the probability assigned to the label are used to predict the error.

To identify the best set of features, we compare these models in terms of the R^2 coefficient.[1] The R^2 coefficient is computed as:

$$R^2 \;=\; 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}, \tag{7}$$

[1] Being a logistic regression, we employ McFadden's pseudo R^2.

where y_i represents a binary indicator of the correctness of $r_i \in \mathcal{R}$, $\hat{y}_i$ denotes the prediction of the logistic regression model for r_i and $\bar{y}_i = \frac{1}{N}\sum_{i=1}^{N} y_i$. This coefficient measures how much of the variability in the dependent variable Y is explained by the independent variable(s) X and takes values in the interval $[0, 1]$, with 1 indicating that the model fits the data perfectly.

4.2 Experimental Methodology to answer RQ2

While the methodology described in the previous section allows us to quantify the quality of different stratification features, there are two important limitations to consider. First, it requires access to ground truth labels to compute Y, and therefore it cannot be directly applied in a production setting. Second, even a strong stratification variable is only a part of the solution. The effectiveness of the stratification also depends on several other factors: the number of strata, the threshold values for continuous features, the grouping of categories for categorical features, how to combine multiple stratification features, and the sizes of the resulting strata. All these elements significantly influence the overall cost of the annotation process.

For this reason, we move to a more practical scenario. Inspired by Merlo et al. [19], we directly evaluate the full SS pipeline by measuring the annotation cost required to achieve a specified confidence level in estimating the LLM effectiveness as an annotator. Specifically, we simulate the annotation process using available experimental IR collections that include human-made relevance judgments. We begin by annotating all query-document pairs for relevance using an LLM. Then, we stratify the judgments using the unsupervised features described in Section 3.1. Next, we apply the sampling procedure: we sample an LLM-generated relevance judgment, fetch the corresponding query-document pair, and simulate the annotation step by using its human-provided judgment from the collection. After each sample, we compute the CI for either the MAE or Cohen's κ. This process is repeated until the desired confidence level is achieved. Thus, a stratification strategy is considered effective if it minimizes the number of iterations needed—which directly translates into the number of required human annotations.

4.3 Experimental Setup

We evaluate our pipeline on three widely-used collections: TREC Deep Learning (DL) 2019 [10] and 2020 [9], based on the MSMARCO corpus (8.8M passages), with 43 and 54 queries and 10k relevance judgments on a four graded scale; and the TREC Robust 2004 collection [31, 32] (TIPSTER disks 4 & 5, 528k documents), with 249 queries and 311k judgments on a three-graded scale. Following Merlo et al. [19], we sample 5% of TREC Robust 2004 ($\sim$15k pairs) to mitigate the costs of the LLM annotation and truncate documents to match the average TREC DL passage lengths.

Table 1: R^2 of linear regression models fit using different stratification features. The symbol "$\star$" indicates statistical significance at confidence level $\alpha = 0.05$ with respect to the null model, according to the likelihood ratio test.

	TREC robust 2004			TREC DL 2019			TREC DL 2020		
	Mistral 7B	Llama 8B	Qwen 30B	Mistral 7B	Llama 8B	Qwen 30B	Mistral 7B	Llama 8B	Qwen 30B
A	$0.211^\star$	$0.622^\star$	$0.538^\star$	$0.249^\star$	$0.313^\star$	$0.310^\star$	$0.389^\star$	$0.458^\star$	$0.440^\star$
P	$0.145^\star$	$0.173^\star$	$0.256^\star$	$0.084^\star$	$0.027^\star$	$0.027^\star$	$0.147^\star$	$0.032^\star$	$0.104^\star$
ΔP	$0.097^\star$	$0.031^\star$	0.000	$0.023^\star$	$0.013^\star$	$0.003^\star$	$0.063^\star$	$0.010^\star$	$0.026^\star$
ΔP_2	$0.092^\star$	$0.113^\star$	$0.008^\star$	$0.022^\star$	$0.047^\star$	$0.017^\star$	$0.065^\star$	$0.065^\star$	$0.071^\star$
PPL	$0.040^\star$	$0.026^\star$	$0.015^\star$	$0.096^\star$	$0.017^\star$	$0.028^\star$	$0.213^\star$	$0.007^\star$	$0.043^\star$
$A + P$	$0.238^\star$	$0.634^\star$	$0.569^\star$	$0.253^\star$	$0.331^\star$	$0.311^\star$	$0.390^\star$	$0.474^\star$	$0.450^\star$
$A + PPL$	$0.225^\star$	$0.623^\star$	$0.541^\star$	$0.250^\star$	$0.313^\star$	$0.311^\star$	$0.390^\star$	$0.458^\star$	$0.440^\star$

We test three LLMs of varying sizes: Llama 3.1 8B, Mistral v0.3 7B, and Qwen3 30B, using their Instruct versions when available to improve prompt adherence. To generate relevance judgments, we adopt the UMBRELA prompt [30], modified to return only relevance levels, ensuring consistent token probability and perplexity scores across query-documents pairs.

For the estimation pipeline, we set $\alpha = 0.05$ and $\epsilon = 0.05$, and we vary the number of strata $H \in [2, ..., 8]$. Note that, since A is categorical, when it is used alone as a stratification feature, the only meaningful values of H are: 2, emulating binarized relevance labels, and 3 or 4, depending on the number of relevance levels of the considered collection. In the case H=2, to emulate the binarization, following common approaches [12, 30], we assign the query-document pairs with labels "related" and "irrelevant", for the TREC DL collections, and label "not relevant", for the TREC Robust 2004, to the first stratum, while all the other pairs are assigned to the second stratum.

To account for the variability introduced by the sampling step, we repeat the procedure 10 times and average the outcomes.

5 Results and Discussion

In this section, we first present the results of evaluating the considered stratification features, and then we analyze the results obtained by applying the SS pipeline to estimate MAE and Cohen's κ.

5.1 Answering RQ1: Quality of the Stratification Features

Table 1 reports the R^2 of the models fit using different stratification features to predict the correctness of the LLM labels.[2] All the models, except the one based

[2] To favour readability, we report two configurations that include the most promising features according to the single-features models assessment. Other approaches tend to be on a par or worse. The full table is available here: `https://github.com/MerloSimone/StratifiedSamplingLLMEstimation/blob/main/complete_feature_table.pdf`.

on ΔP for Qwen 30B when applied to the TREC Robust 2004, are statistically significantly better than the null model, indicating some relation between the chosen stratification features and the response variable. If we consider single-feature models, we observe that, in general, the most expressive feature is the relevance judgement assigned by the LLM (A). In other terms, this indicates that all the LLMs tend to make mistakes more often when assigning specific classes. This kind of result aligns with past findings in the effectiveness of LLMs as evaluators [2, 19, 29]. All the other features, taken individually, appear suboptimal, with the probability assigned to the label (P) and the perplexity (PPL) being the second and third best, depending on the setting. If we move towards more complex models that include multiple features, we notice that, in most cases, there is an improvement in the fitness of the model to the data. In particular, the model that relies on the label and on the probability ($A + P$) always appears the most effective. The magnitude of the improvement over the model with only A depends on the experimental setup—despite this, it is always significant according to the likelihood ratio test that compares the two models.

The most relevant finding is that the patterns, i.e., which models tend to perform best, remain stable both across different collections and LLMs: this suggests that, if we manage to find good stratification features on historical collections, such as the ones we employed in this work, it is likely that the same (combination of) features will be useful also to stratify the relevance judgements in future collections and for future LLMs. Consequently, we do not need to find dataset-specific stratification features that would be impossible to evaluate in the absence of ground truth signals.

5.2 Answering RQ2: Impact of Stratified Sampling on the Cost

MAE Estimation. Based on the results from the evaluation of the stratification features, we assess the proposed estimation pipeline only on the best feature configurations: A , $A + P$ and $A + PPL$.

In Table 2 we report the obtained results. The first row of the table reports the real MAE value, computed by validating all the LLM-generated relevance judgments. The second row reports the estimation results and the related cost (between brackets) obtained adopting the SRS framework introduced by Merlo et al. [19]. The remaining portion of the table is splitted in 3 sections, one for each of the considered stratification features configurations. Here, each row reports the estimation results and costs of the proposed pipeline when a different number of strata H is used. Both for SRS and SS we omit the CI since the MoE is always equal to (or lower than) $\epsilon = 0.05$.

The results obtained for $Y = A$ reveal that, while providing a correct estimate of the MAE, the proposed estimation pipeline also allows to consistently reduce the number of relevance judgments that need to be validated by humans. In particular, for the TREC Robust 2004 collection the reduction in cost spans from 25% (for Mistral 7B) up to 76% (for Llama 8B). For the TREC DL collections, instead, the cost reduction is less pronounced and more consistent, spanning from 12% (for Mistral 7B) to 29% (for Llama 8B) for TREC DL 2019, and from

Table 2: MAE estimation results for different stratification features. The value indicates the estimated MAE, while the number between parentheses is the number of annotations required to achieve it.

	TREC robust 2004			**TREC DL 2019**			**TREC DL 2020**		
Model	**Mistral 7B**	**Llama 8B**	**Qwen 30B**	**Mistral 7B**	**Llama 8B**	**Qwen 30B**	**Mistral 7B**	**Llama 8B**	**Qwen 30B**
Real MAE	0.358	0.505	0.391	0.908	1.002	0.792	0.913	0.985	0.729
SRS	0.363 (451.4)	0.507 (1044.8)	0.391 (823.9)	0.903 (1015.9)	1.003 (1419.6)	0.785 (1151.8)	0.915 (1170.7)	0.978 (1533.8)	0.727 (1175.0)
Strata #:					$Y = A$				
2	0.353 (335)	0.518 (272)	0.391 (302)	0.910 (895)	0.996 (1078)	0.777 (920)	0.913 (809)	0.986 (878)	0.731 (744)
3	0.360 (337)	0.521 (246)	0.388 (257)	-	-	-	-	-	-
4	-	-	-	0.920 (794)	1.006 (1008)	0.793 (855)	0.916 (639)	0.983 (748)	0.730 (642)
					$Y = A + P$				
2	0.353 (335)	0.518 (272)	0.391 (302)	0.904 (817)	0.996 (1078)	0.802 (915)	0.908 (798)	0.985 (870)	0.732 (750)
3	0.367 (312)	0.521 (246)	0.409 (185)	0.902 (796)	1.014 (1067)	0.802 (869)	0.910 (654)	0.996 (830)	0.727 (632)
4	0.348 (306)	0.500 (208)	0.400 (162)	0.905 (808)	1.005 (1019)	0.796 (837)	0.917 (631)	0.992 (722)	0.728 (596)
5	0.344 (297)	0.496 (187)	0.406 (133)	0.911 (796)	1.015 (1021)	0.777 (818)	0.915 (645)	1.002 (713)	0.723 (576)
6	0.350 (298)	0.493 (202)	0.403 (115)	0.901 (787)	1.001 (958)	0.782 (818)	0.911 (642)	0.990 (672)	0.729 (580)
7	0.348 (273)	0.502 (217)	0.408 (147)	0.920 (793)	0.998 (964)	0.800 (879)	0.920 (643)	0.989 (681)	0.723 (588)
8	0.359 (278)	0.507 (201)	0.401 (127)	0.917 (796)	1.000 (940)	0.789 (839)	0.923 (641)	0.988 (713)	0.724 (581)
					$Y = A + PPL$				
2	0.331 (242)	0.504 (1042)	0.392 (294)	0.911 (1047)	0.996 (1078)	0.802 (915)	0.912 (1102)	0.985 (870)	0.732 (750)
3	0.354 (294)	0.511 (971)	0.386 (283)	0.904 (968)	0.998 (1033)	0.794 (874)	0.904 (982)	0.983 (1006)	0.712 (696)
4	0.354 (262)	0.492 (350)	0.378 (261)	0.903 (922)	1.008 (1026)	0.796 (913)	0.896 (909)	0.991 (882)	0.730 (630)
5	0.347 (248)	0.508 (345)	0.378 (240)	0.902 (894)	1.006 (1032)	0.793 (885)	0.890 (886)	0.973 (842)	0.720 (596)
6	0.359 (238)	0.518 (335)	0.381 (239)	0.909 (897)	1.000 (1026)	0.805 (889)	0.914 (879)	0.987 (785)	0.732 (626)
7	0.346 (215)	0.495 (221)	0.382 (241)	0.902 (879)	1.001 (1007)	0.797 (886)	0.909 (882)	0.990 (736)	0.721 (607)
8	0.358 (264)	0.509 (234)	0.392 (233)	0.905 (905)	0.997 (969)	0.795 (856)	0.916 (930)	0.978 (735)	0.723 (620)

31 % to 51% for TREC DL 2020. Moreover, it is possible to notice that defining separate strata for each relevance level ($H \in 3, 4$) allows to reduce the costs.

Considering the results obtained for $Y = A + P$, we notice that the cost reductions span from a minimum of 19% (with Mistral 7B on the TREC DL 2019) to a maximum of 86% (with Qwen 30B on the TREC Robust 2004). Nonetheless, the pipeline performance appears similar to the one obtained with $Y = A$. An interesting behavior can be observed when considering the performance for the different number of strata of a single model on a specific collection. Indeed, if we focus on Qwen 30B and the TREC DL 2019 collection we can notice how the estimation cost when using 5 strata (818) is lower than the costs obtained when using 7 or 8 strata (879 and 839, respectively). This highlights that defining a large number of strata may not always be beneficial. Indeed, SS sampling requires at least one judgment from each of the strata. Thus, defining more strata may lead to sampling more judgments. For example, if we set $H = N$ (a stratum for each judgment in the population) we are forced to sample all the judgments.

When considering $Y = A + PPL$, instead, both the best and worst cases in terms of cost reduction are achieved with Llama 8B on the TREC Robust 2004, spanning from 0.25% to 79% saved annotations. In addition, on the TREC DL 2019, when Mistral 7B is employed and 2 is chosen as the strata number, the proposed estimatiom pipeline requires to annotate more judgments than SRS. A closer look at the results reveals that, in general, when PPL and A are jointly used and the number of strata H is set to 2, the cost is higher than for the other analyzed stratification features configurations. This suggest that using A

Table 3: Cohen's κ estimation results when assigned relevance A is used as stratification feature. The value indicates the estimated Cohen's κ, while the number between parentheses is the number of annotations required to achieve it.

	TREC robust 2004			**TREC DL 2019**			**TREC DL 2020**		
Model	**Mistral 7B**	**Llama 8B**	**Qwen 30B**	**Mistral 7B**	**Llama 8B**	**Qwen 30B**	**Mistral 7B**	**Llama 8B**	**Qwen 30B**
Real Cohen's κ	0.173	0.148	0.211	0.136	0.163	0.233	0.131	0.152	0.225
SRS	0.167 (1298)	0.157 (956)	0.217 (950)	0.129 (368)	0.141 (420)	0.236 (463)	0.141 (396)	0.162 (429)	0.237 (518)
Strata #:									
2	0.193 (3025)	0.119 (1279)	0.148 (1655)	0.149 (382)	0.150 (368)	0.235 (425)	0.132 (353)	0.151 (352)	0.223 (468)
3	0.166 (1782)	0.115 (1282)	0.130 (1512)	-	-	-	-	-	-
4	-	-	-	0.138 (362)	0.143 (335)	0.233 (428)	0.131 (336)	0.151 (363)	0.223 (438)

and PPL to stratify the data in a low amount of strata may not be enough to achieve a high per-stratum homogeneity.

A comparison among the results obtained for A, $A+P$ and $A+PPL$, reveals that using A as the only stratification feature, even if it limits the amount of strata that can be created, allows to achieve the most stable performance while drastically reducing the cost with respect to standard sampling techniques. At the same time, using $A+P$ allows to reduce even more the costs with respect to SRS but appears to lead to less stable results. Finally, the $A+PPL$ appears to be the worst solution in terms of stability but still allows to reduce the costs with respect to both SRS and A.

Cohen's κ Estimation. Due to the definition of the Cohen's κ estimator employed, when estimating Cohen's κ the only stratification feature that can be considered is the assigned label A. In Table 3 we report the results when applying the proposed pipeline to estimate Cohen's κ. The table is structured in the same way as for the MAE but it has only one section: $Y = A$.

Considering the TREC DL collections, the proposed pipeline confirms its validity. The only case in which SRS requires less annotations than SS is when Mistral 7B is employed on the TREC DL 2019 collection with H=2. Nonetheless, when the number of strata is set to be the same as the relevance levels, the proposed approach allows to reduce the costs up to 20%. On the TREC Robust 2004, instead, the proposed solution performs worse than SRS. We ascribe this unexpected behavior to the fact that, for this collection, we sample the relevance judgments and we truncate the documents, inducing the LLM to declare many query-document pairs as irrelevant. Indeed, this leads to a highly unbalanced stratification, since all the pairs marked as irrelevant (over 85%) are assigned to a single stratum, while the remaining are distributed across the other strata. Thus, given that the Cohen's κ estimator that we employ requires drawing a consistent number of samples from each stratum, this results in increased costs.

6 Conclusions and Future Work

In this paper, we introduced an iterative pipeline that leverages Stratified Sampling (SS) to reduce human effort in validating LLM-generated relevance judgments. Our results demonstrate that, compared to standard sampling techniques, SS can reduce the number of human annotations required to estimate the agreement between LLMs and human assessors by up to 85%. However, current estimators of Cohen's κ are limited in that they only support stratification based on the assigned label A and are susceptible to unbalanced stratifications, where the majority of data falls within a single stratum.

As future work, we plan to develop Cohen's κ estimators that are better suited for SS settings, and to investigate how to quantify uncertainties in IR system evaluations when relying on LLM-generated judgments.

Acknowledgments. This work has received support from CAMEO, PRIN 2022 n. 2022ZLL7MW.

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

Bibliography

[1] Abbasiantaeb, Z., Meng, C., Azzopardi, L., Aliannejadi, M.: Can we use large language models to fill relevance judgment holes? In: Acharya, P., Clarke, C.L.A., Crestani, F., Fu, X., Jones, G.J.F., Kando, N., Kato, M.P., Lipani, A., Liu, Y. (eds.) Joint Proceedings of the 1st Workshop on Evaluation Methodologies,Testbeds and Community for Information Access Research (EMTCIR 2024) and the 1st Workshop on User Modelling in Conversational Information Retrieval (UM-CIR 2024) co-located with the 2nd International ACM SIGIR Conference on Information Retrieval in the Asia Pacific (SIGIR-AP 2024), Tokyo, Japan, December 12, 2024, CEUR Workshop Proceedings, vol. 3854, CEUR-WS.org (2024), URL https://ceur-ws.org/Vol-3854/emtcir-2.pdf

[2] Alaofi, M., Thomas, P., Scholer, F., Sanderson, M.: Llms can be fooled into labelling a document as relevant: best café near me; this paper is perfectly relevant. In: Sakai, T., Ishita, E., Ohshima, H., Hasibi, F., Mao, J., Jose, J.M. (eds.) 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, pp. 32–41, ACM (2024), https://doi.org/10.1145/3673791.3698431, URL https://doi.org/10.1145/3673791.3698431

[3] Chen, N., Liu, J., Dong, X., Liu, Q., Sakai, T., Wu, X.: AI can be cognitively biased: An exploratory study on threshold priming in llm-based batch relevance assessment. In: Sakai, T., Ishita, E., Ohshima, H., Hasibi, F., Mao, J., Jose, J.M. (eds.) 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, pp. 54–63, ACM (2024), https://doi.org/10.1145/3673791.3698420, URL https://doi.org/10.1145/3673791.3698420

[4] Chiang, D.C., Lee, H.: Can large language models be an alternative to human evaluations? In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 15607–15631, Association for Computational Linguistics (2023), https://doi.org/10.18653/V1/2023.ACL-LONG.870, URL https://doi.org/10.18653/v1/2023.acl-long.870

[5] Chu, X., Ilyas, I.F., Krishnan, S., Wang, J.: Data cleaning: Overview and emerging challenges. In: Özcan, F., Koutrika, G., Madden, S. (eds.) Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016, pp. 2201–2206, ACM (2016), https://doi.org/10.1145/2882903.2912574, URL https://doi.org/10.1145/2882903.2912574

[6] Clarke, C., Dietz, L.: Llm-based relevance assessment still can't replace humanrelevance assessment. In: Proceedings of the Eleventh International

Workshop on Evaluating Information Access, EVIA 2025, a Satellite Workshop of the NTCIR-18 Conference, Tokyo, Japan, June 10, 2025, National Institute of Informatics (NII) (2025), https://doi.org/10.20736/0002002105, URL `https://doi.org/10.20736/0002002105`

[7] Cleverdon, C.W.: The aslib cranfield research project on the comparative efficiency of indexing systems. ASLIB Proceedings **12**(12), 421–431 (1960), ISSN 0001-253X, https://doi.org/10.1108/eb049778

[8] Cochran, W.G.: Sampling Techniques. John Wiley (1963)

[9] Craswell, N., Mitra, B., Yilmaz, E., Campos, D.: Overview of the TREC 2020 deep learning track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020, NIST Special Publication, vol. 1266, National Institute of Standards and Technology (NIST) (2020), URL `https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.DL.pdf`

[10] Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M.: Overview of the TREC 2019 deep learning track. CoRR **abs/2003.07820** (2020), URL `https://arxiv.org/abs/2003.07820`

[11] Faggioli, G., Dietz, L., Clarke, C.L.A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., Wachsmuth, H.: Perspectives on large language models for relevance judgment. In: Yoshioka, M., Kiseleva, J., Aliannejadi, M. (eds.) Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023, pp. 39–50, ACM (2023), https://doi.org/10.1145/3578337.3605136, URL `https://doi.org/10.1145/3578337.3605136`

[12] Faggioli, G., Ferro, N., Fuhr, N.: Detecting significant differences between information retrieval systems via generalized linear models. In: Hasan, M.A., Xiong, L. (eds.) Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, pp. 446–456, ACM (2022), https://doi.org/10.1145/3511808.3557286, URL `https://doi.org/10.1145/3511808.3557286`

[13] Farzi, N., Dietz, L.: Does UMBRELA work on other llms? In: Ferro, N., Maistro, M., Pasi, G., Alonso, O., Trotman, A., Verberne, S. (eds.) Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025, pp. 3214–3222, ACM (2025), https://doi.org/10.1145/3726302.3730317, URL `https://doi.org/10.1145/3726302.3730317`

[14] Fröbe, M., Parry, A., Schlatt, F., MacAvaney, S., Stein, B., Potthast, M., Hagen, M.: Large language model relevance assessors agree with one another more than with human assessors. In: Ferro, N., Maistro, M., Pasi, G., Alonso, O., Trotman, A., Verberne, S. (eds.) Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025, pp.

2858–2863, ACM (2025), https://doi.org/10.1145/3726302.3730218, URL https://doi.org/10.1145/3726302.3730218

[15] Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd-workers for text-annotation tasks. CoRR **abs/2303.15056** (2023), https://doi.org/10.48550/ARXIV.2303.15056, URL https://doi.org/10.48550/arXiv.2303.15056

[16] de Jesus, G., Nunes, S.S.: Exploring large language models for relevance judgments in tetun. In: Siro, C., Aliannejadi, M., Rahmani, H.A., Craswell, N., Clarke, C.L.A., Faggioli, G., Mitra, B., Thomas, P., Yilmaz, E. (eds.) 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, pp. 19–30, CEUR-WS.org (2024), URL https://ceur-ws.org/Vol-3752/paper2.pdf

[17] Marchesin, S., Silvello, G.: Efficient and reliable estimation of knowledge graph accuracy. Proc. VLDB Endow. **17**(9), 2392–2404 (2024), https://doi.org/10.14778/3665844.3665865, URL https://www.vldb.org/pvldb/vol17/p2392-marchesin.pdf

[18] Meng, C., Arabzadeh, N., Askari, A., Aliannejadi, M., de Rijke, M.: Query performance prediction using relevance judgments generated by large language models. ACM Trans. Inf. Syst. **43**(4), 106:1–106:35 (2025), https://doi.org/10.1145/3736402, URL https://doi.org/10.1145/3736402

[19] Merlo, S., Marchesin, S., Faggioli, G., Ferro, N.: A cost-effective framework to evaluate llm-generated relevance judgements. In: Cha, M., Park, C., Park, N., Yang, C., Roy, S.B., Li, J., Kamps, J., Shin, K., Hooi, B., He, L. (eds.) Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025, pp. 2115–2126, ACM (2025), https://doi.org/10.1145/3746252.3761200, URL https://doi.org/10.1145/3746252.3761200

[20] Otero, D., Parapar, J., Barreiro, Á.: Limitations of automatic relevance assessments with large language models for fair and reliable retrieval evaluation. In: Ferro, N., Maistro, M., Pasi, G., Alonso, O., Trotman, A., Verberne, S. (eds.) Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025, pp. 2545–2549, ACM (2025), https://doi.org/10.1145/3726302.3730221, URL https://doi.org/10.1145/3726302.3730221

[21] Rahmani, H.A., Siro, C., Aliannejadi, M., Craswell, N., Clarke, C.L.A., Faggioli, G., Mitra, B., Thomas, P., Yilmaz, E.: Llm4eval: Large language model for evaluation in IR. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, pp. 3040–3043, ACM

(2024), https://doi.org/10.1145/3626772.3657992, URL `https://doi.org/10.1145/3626772.3657992`

[22] Rahmani, H.A., Siro, C., Aliannejadi, M., Craswell, N., Clarke, C.L.A., Faggioli, G., Mitra, B., Thomas, P., Yilmaz, E.: Judging the judges: A collection of llm-generated relevance judgements. CoRR **abs/2502.13908** (2025), https://doi.org/10.48550/ARXIV.2502.13908, URL `https://doi.org/10.48550/arXiv.2502.13908`

[23] Rahmani, H.A., Yilmaz, E., Craswell, N., Mitra, B., Thomas, P., Clarke, C.L.A., Aliannejadi, M., Siro, C., Faggioli, G.: Llmjudge: Llms for relevance judgments. In: Siro, C., Aliannejadi, M., Rahmani, H.A., Craswell, N., Clarke, C.L.A., Faggioli, G., Mitra, B., Thomas, P., Yilmaz, E. (eds.) 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, pp. 1–3, CEUR-WS.org (2024), URL `https://ceur-ws.org/Vol-3752/paper8.pdf`

[24] Siro, C., Rahmani, H.A., Aliannejadi, M., Craswell, N., Clarke, C.L.A., Faggioli, G., Mitra, B., Thomas, P., Yilmaz, E.: Llm4eval: Large language model for evaluation in IR. In: Ferro, N., Maistro, M., Pasi, G., Alonso, O., Trotman, A., Verberne, S. (eds.) Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025, pp. 4188–4191, ACM (2025), https://doi.org/10.1145/3726302.3730367, URL `https://doi.org/10.1145/3726302.3730367`

[25] Soboroff, I.: Don't use llms to make relevance judgments. Inf. Retr. Res. J. **1**(1), 29–46 (2025), https://doi.org/10.54195/IRRJ.19625, URL `https://doi.org/10.54195/irrj.19625`

[26] Stehman, S.V.: Estimating standard errors of accuracy assessment statistics under cluster sampling. Remote Sensing of Environment **60**(3), 258–269 (1997), ISSN 0034-4257, https://doi.org/https://doi.org/10.1016/S0034-4257(96)00176-9, URL `https://www.sciencedirect.com/science/article/pii/S0034425796001769`

[27] Thakur, N., Pradeep, R., Upadhyay, S., Campos, D., Craswell, N., Lin, J.: Support evaluation for the trec 2024 rag track: Comparing human versus llm judges (2025), URL `https://arxiv.org/abs/2504.15205`

[28] Thomas, P., Oard, D.W., Yang, E., Lawrie, D.J., Mayfield, J.: System comparison using automated generation of relevance judgements in multiple languages. In: Ferro, N., Maistro, M., Pasi, G., Alonso, O., Trotman, A., Verberne, S. (eds.) Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025, pp. 2812–2816, ACM (2025), https://doi.org/10.1145/3726302.3730252, URL `https://doi.org/10.1145/3726302.3730252`

[29] Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th Inter-

national ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, pp. 1930–1940, ACM (2024), https://doi.org/10.1145/3626772.3657707, URL `https://doi.org/10.1145/3626772.3657707`

[30] Upadhyay, S., Pradeep, R., Thakur, N., Craswell, N., Lin, J.: UMBRELA: umbrela is the (open-source reproduction of the) bing relevance assessor. CoRR **abs/2406.06519** (2024), https://doi.org/10.48550/ARXIV.2406.06519, URL `https://doi.org/10.48550/arXiv.2406.06519`

[31] Voorhees, E.: Overview of the trec 2004 robust retrieval track. In: TREC (2004)

[32] Voorhees, E.M.: Nist trec disks 4 and 5: Retrieval test collections document set (1996), https://doi.org/10.18434/t47g6m

[33] Zhu, Y., Zhang, P., ul Haq, E., Hui, P., Tyson, G.: Can chatgpt reproduce human-generated labels? A study of social computing tasks. CoRR **abs/2304.10145** (2023), https://doi.org/10.48550/ARXIV.2304.10145, URL `https://doi.org/10.48550/arXiv.2304.10145`