Variations in Relevance Judgments and the Shelf Life of Test Collections

Andrew Parry
University of Glasgow
United Kingdom, Glasgow

Maik Fröbe
Friedrich-Schiller-
Universität Jena
Germany, Jena

Harrisen Scells
University of Tübingen
Germany, Tübingen

Ferdinand Schlatt
Friedrich-Schiller-
Universität Jena
Germany, Jena

Guglielmo Faggioli
University of Padua
Italy, Padua

Saber Zerhoudi
Universität Passau
Germany, Passau

Sean MacAvaney
University of Glasgow
United Kingdom, Glasgow

Eugene Yang
Johns Hopkins University
USA, Baltimore

Abstract

The fundamental property of Cranfield-style evaluations, that system rankings are stable even when assessors disagree on individual relevance decisions, was validated on traditional test collections. However, the paradigm shift towards neural retrieval models affected the characteristics of modern test collections, e.g., documents are short, judged with four grades of relevance, and information needs have no descriptions or narratives. Under these changes, it is unclear whether assessor disagreement remains negligible for system comparisons. We investigate this aspect under the additional condition that the few modern test collections are heavily re-used. Given more possible query interpretations due to less formalized information needs, an "expiration date" for test collections might be needed if top-effectiveness requires overfitting to a single interpretation of relevance. We run a reproducibility study and re-annotate the relevance judgments of the 2019 TREC Deep Learning track. We can reproduce prior work in the neural retrieval setting, showing that assessor disagreement does not affect system rankings. However, we observe that some models substantially degrade with our new relevance judgments, and some have already reached the effectiveness of humans as rankers, providing evidence that test collections can expire.

CCS Concepts

• **Information systems** → **Evaluation of retrieval results**.

Keywords

Evaluation; Test collections; Collection Reliability

Table 1: Overview of annotator agreement studies that we reproduce in the TREC Deep Learning scenario. We contrast information needs (length of title, description, narrative), documents (domain and length), evaluations (measure and relevance grades) and observed results of each study (we show the agreement as overlap and the system correlations as τ).

Study	Information Need			Docs.		Evaluation		Results	
	Title	Desc.	Narr.	Dom.	Len.	Meas.	Rel.	Agr.	Corr.
Lesk, Salton [23]	×	100.0	×	Medical	103.7	Prec.	0–1	0.30	×
Burgin [4]	×	30.4	×	IR	15.7	MAP	0–1	0.49	×
Voorhees [46]	2.6	20.4	64.7	News	550.5	MAP	0–1	0.36	0.890
Our Reproduction	8.3	×	×	QA	30.4	nDCG	0–3	0.15	0.897

1 Introduction

TREC-style test collections enable the evaluation and, thereby, the development of retrieval systems. Making these collections robust and reusable is not trivial, partly because of the inherent subjectivity of relevance [45]. Assessing the reliability of relevance judgments involves assessing inter-annotator agreement and is comparably expensive [4, 23, 46]. Consequently, only a few studies have investigated how the variability of relevance judgments affects the evaluation of retrieval systems. However, those findings were highly important for the design of retrieval test collections as they showed that even when relevance is subjective and therefore annotators disagree [38, 42], subsequent aggregated effectiveness evaluations [4, 23] and system rankings [46] are stable, ensuring the reusability of test collections.

Contrasting older collections, topics in modern test collections often do not have descriptions or narratives [8], which is known to affect agreement [33, 50], documents are often shorter, and relevance is assessed on a graded scale, which has not been investigated in re-annotation studies. Table 1 contrasts previous studies and the properties of the corresponding test collections with the TREC Deep Learning 2019 [8] (DL'19) test collection we use in our reproducibility work. The changes in modern collections potentially increase assessment variability because an assessor can freely interpret the query intent and context that may be missing from the document to assess its relevance on a fine-grained scale. In this reproducibility study, we investigate if the robustness of classical test collections still holds for modern test collections under these new conditions.

These shifts in test collection design coincide with advances in retrieval methodologies; retrieval pipelines increasingly comprise components where neural language models estimate relevance. These systems have continuously improved over multiple iterations through improved training processes [31], through distillation from large language models [32, 40], or by directly using large language models for ranking [43]. We hypothesize that iterative model improvements on static benchmarks may overfit to the specific opinions expressed in static relevance judgments, introducing systemic bias, which can be characterized as "bias by research design" in which the influence of previous works can bias new findings implicitly towards a common benchmark [18]. The TREC DL'19 track remains a commonly used standard for benchmarking systems, being cited and most likely evaluated hundreds of times (Cited 32 times at ECIR'24 and SIGIR'24 alone). Craswell et al. [9] alluded to this point, warning against the iterative design processes employed in ad-hoc neural ranking.

Thus, we look to replicate the classic setting for evaluating the robustness of collections proposed by Voorhees [46] for modern collections, with the additional hypothesis that bias in the system orderings may be introduced through model iteration in addition to human disagreement. Since a single expert human assessor's relevance grades on a given topic are subjective and "equally plausible" as another assessor's, we can compare the effectiveness of systems using multiple "equally correct" sets of relevance judgments. If a system's relative effectiveness degrades when evaluated using alternative judgments, it may indicate that models overfit the preferences expressed by the collection's official judgments.

In this paper, we conduct a re-annotation of TREC DL'19 using two "secondary" annotators for every topic. Based on our new relevance judgments, we first provide a qualitative analysis of this process and validate the findings of previous work in a modern setting. That is, we observe low agreement among assessors but system rankings remain stable. Secondly, treating human annotators as manual TREC runs, we find that one can no longer meaningfully discriminate between the state-of-the-art and human judges on this collection. A reasonable "upper bound" effectiveness is around nDCG@10 of 0.81 (due to ambiguities in relevance assessment). We suggest that reported results that are higher than this bound on DL'19 may already overfit to the relevance labels. Note that we are not aware of any existing system reporting results exceeding this bound.[1] However, the leading systems today are very close, with RankZephyr [32] and RankGPT [43] reaching between 0.78–0.80 nDCG@10 with a sufficiently strong first stage. We additionally find that closed-sourced large language models and systems distilled from them can fluctuate greatly in relative effectiveness, often degrading system rank on our new relevance judgments.

Our reproducibility effort spawned from the ECIR 2024 Collaba-thon on Neural IR [24]. A subset of the IR community discussed that the current generation of retrieval models may have reached a practical upper bound on existing test collections, but we were unsure when to consider a test collection to be saturated or expired. Our study reproduced prior findings on the inherent subjectivity and variability of relevance judgments (especially in the absence of a well-defined narrative), and it established an empirical upper bound on TREC DL'19. Therefore, we encourage caution when attempting to maximize a given metric as we approach and exceed human bounds on relevance estimation. Exceeding human effectiveness, given the subjectivity of relevance, may not lead to an optimal system but instead stagnation of progress in the field. We publish our code and relevance judgments for reproducibility and to encourage future studies.[2]

2 Related Work

This section outlines evaluation frameworks in retrieval, previous efforts to assess collection reliability, and judgment stability.

Offline Evaluation. The Cranfield experiments introduced the modern offline test collection in ad-hoc ranking; experts curated a set of queries, and texts were retrieved from a static collection and exhaustively annotated, allowing for offline relevance judgments [6]. Completeness of relevance judgments, i.e., the exhaustive annotation of corpora, is a key factor in the reliability of an offline test collection [48]; however, the scale of modern corpora makes such a task infeasible. As such, TREC-style collections are collated by pooling the runs of multiple diverse systems to create a set of query–document pairs to be annotated under the assumption that unpooled documents are not relevant [47]. Assessors then label these pairs and assign a grade of relevance. A single annotator is often assigned to each topic; this decision generally reduces costs and has been validated by several studies measuring agreement between annotators as an effect on the reliability of a collection.

Relevance judgments are subjective and, therefore, can vary among annotators [25], with low inter-annotator agreement [37], which can indicate low collection reliability. As such, narratives and training are usually provided to annotators describing the information need underlying a query [15]. We consider that these homogenized judgments may be a weak point for ongoing collection reliability as we learn from them in a semi-supervised fashion through effective systems acting as teachers.

Notions of Relevance. The reliability of offline test collections is contested due to the intrinsic ambiguity and biases present in relevance assessment under different human annotators [34, 45], annotation settings [33], relevance grading [10], and, more recently, the process of collecting candidate documents via pooling [54]. As these concerns have been raised, one method to validate offline test collections is to re-annotate judgments. Lesk and Salton [23] and Burgin [4] empirically validated the collection reliability under annotator disagreement, observing that system order is maintained under different annotators by measuring variance in precision and recall. When web-scale corpora rendered exhaustive annotation impossible, depth pooling became popularized to provide a minimal set of documents that should be annotated to discriminate system effectiveness. Concerns were two-fold: that systems that contributed to the pool would have an advantage over new systems and that the reduced annotations may disrupt the stability of system orderings. Voorhees [46] found that system order stability could be maintained under the re-annotation of test collection pools from TREC-4 [14] and -6 [50].

---

[1]We do not consider large ensemble systems, given their risks [2].

[2]https://github.com/Parry-Parry/sigir25-annotation

Collection Retirement. Previous work has discussed the retirement of a test collection in recent years, though not retired before dissemination the use of TREC 2021 Deep Learning test collection [9] is discouraged by NIST due to the large increase in corpora size leading to pooling bias; both NIST [49] and participants in the track [20] had concerns regarding the ability to measure recall due to incomplete labeling. However, we are not aware of any work that considers the effect of collection popularity and subsequent bias induced by fitting to relevance judgments over time; our focus on precision is in line with the philosophy of neural systems and the Deep Learning track.

Craswell et al. [9] warns against using TREC Deep Learning test collections for experimental iteration to reduce the chance of overfitting to judgments. Though this guidance is generally followed with the MSMARCO-dev set acting as a suitable alternative, the broad use of existing strong systems validated on TREC Deep Learning 2019 [17, 32, 52] may lead to overfitting via semi-supervision. We draw a parallel with the work of Armstrong et al. [3], which warned that progress might stagnate due to poor practice in evaluation. Similarly, we note that the continued comparison to a dataset that guides our architecture and training pipeline may lead to stagnation in system development.

Within a modern ranker training pipeline, several components depend on existing models for effectiveness. Hard negative mining is a process in which, instead of random negative examples in a contrastive loss, samples are taken from the top most relevant documents to a query with the assumption that the majority will be non-relevant [21]. It is common to employ one or several models' pooled rankings to sample negatives [41, 53], with these models often chosen based on their in-domain effectiveness. More explicitly, pipelines are frequently composed of multiple stages of distillation [39, 51, 53], which is a semi-supervised process in which a student model is optimized to approximate the output of a teacher model [16]. Again, these models will be chosen based on in-domain effectiveness and coupled with hard negatives.

3 Evaluating Reliability through Re-Annotation

The reliability of a test collection, given the subjectivity of relevance, is often validated by re-annotation [4, 23, 46]. Modern IR models are often optimized to mimic previously effective systems through data-driven processes. We examine how these models, designed after a collection's release, perform on new judgments following the process of Voorhees [46].

3.1 Reproduction Source

We describe the original annotation process applied by Voorhees [46]. This work was carried out as a component of the curation of the TREC-4 [14] and TREC-6 [50] corpora. As is standard in the curation of pooled TREC corpora, one "primary" annotator was assigned to each topic, subsequently judging each document in the pool. Two additional annotators were assigned to each topic to validate the effect of re-annotation on depth-pooled corpora. A sub-sample of judgments was taken from the pool and re-annotated. In TREC-4, a maximum of 200 relevant documents per topic were taken as a secondary pool for re-annotation. Additionally, 200 non-relevant documents were sampled per topic. A binary judgment was

assigned to each query–document pair, relevant or non-relevant. The texts constituting the primary annotation pool, left unjudged in the secondary pool, are included in all system evaluations, maintaining identical pooling biases. The comparison of "secondary" annotations was carried out to assess hypotheses that resurface frequently in IR literature. This work focuses on the subjectivity of relevance and how pooling affects this subjectivity. Thus, experiments were conducted over a depth-pooled corpus under construction at the time. This process produces a subset of judgments where each query–document pair was re-annotated twice.

Where possible, we follow the evaluation process of Voorhees [46], including analysis of the probability of two systems swapping in effectiveness under different annotators. We define, in an annotation set A, an annotation $a \in A$ as a tuple $\{q, d, r\}$ composed of query q, document d, and judgment $r \in [0, 1, 2, 3]$. For a given annotator pair and a primary annotator, assigned n query-document pairs, we have three annotation sets A_1, A_2, A_3 where $A_i = \{a_j\}_{j=0}^{n}$. For m queries, as noted by Voorhees [46], there are 3^m possible combinations. Thus, these combinations allow for the greater exploration of judgments that could have been produced from a given annotation process. From the 3 annotation sets, a combination A^* is sampled s times, and systems are evaluated against each combination. The probability of a swap between systems i and j can be evaluated under a given measure f given judgments A^*. For systems S, we produce a matrix $B \in \mathbb{Z}^{|S| \times |S|}$ where $B[i, j]$ is the number of annotation samples which caused systems S_i to have higher effectiveness than S_j. The swap probability between these systems can then be computed as $\frac{\min(B[i,j], B[j,i])}{s}$. Due to the symmetric treatment of system comparisons, this computation is bounded between 0 and 0.5.

3.2 Re-Annotation of a Modern Collection

While the original study focused on binary judgments, our work adapts this methodology to graded relevance and neural systems under a modern evaluation setting. Our reproduction is conducted on the TREC Deep Learning 2019 track (DL'19) [8]. Specifically, we use the passage collection, which is generally more popular. The proposal of more granular measures that explicitly consider the ordering of a ranking, such as discounted cumulative gain, as opposed to set-based measures, such as Precision@k, generally requires graded relevance as opposed to the binary relevance grades of older collections. DL'19 was annotated under this setting with four grades of relevance: 'perfectly relevant,' 'highly relevant,' 'related', and 'non-relevant.' This distinction leads to some key differences in our process and analysis.

We treat the original judgments of DL'19 as those of the "primary" annotator. In our study, two annotators from a pool of 8 retrieval academics[3] are assigned to each topic and act as "secondary" annotators (corresponding to 4 topic groups), aligning with the original study in that each topic is annotated by two new annotators; however, the original study does not specify the number of total annotators. The size of the annotator pool used in this work

---

[3]These academics contributed to subsequent analysis and write-up in this work. This decision has minimal effect on our findings; nevertheless, we acknowledge this overlap.

is similar to that of TREC collections [42], albeit with a reduced total annotation budget due to cost and following in the sampling setting of Voorhees [46]. We create a secondary pool composed of all documents in the pool with a grade of 1 or more, and sample 100 documents per annotator from the non-relevant documents of the primary pool. We reduce the number of non-relevant judgments from the pool, motivated by the hypothesis that agreement is far greater over what is non-relevant; for set-based measures, this choice leads to a balance between relevant and non-relevant texts and has minimal effect on nDCG.

Our process for pool creation uses all judged documents at least related to a topic and assigns topics to annotators in a round-robin fashion to balance annotation load while maintaining two annotators per topic. A priority queue of annotator pairs ordered by their current annotation load is produced and assigned the next topic with the greatest number of judgments. The load on each annotator is then balanced by swapping large topics to annotators with minimal load in a second round-robin. We then assign non-relevant texts to each annotator pair with a total mean annotation load of 1125 pairs per annotator.

Annotator Instruction. Each annotator was provided with annotation guidelines, including the original relevance grade descriptions from DL'19. Annotators were made aware that instances had been previously annotated for DL'19; however, they were not provided with relevance grades or a grade distribution. Due to the specialist knowledge required to assess relevance in, for example, medical topics, annotators were allowed to familiarise themselves with topics at their discretion. To facilitate reproducibility and prevent exposure to discussion and instances of DL'19 relevance judgments, which may be indexed on common search engines, we instead enforce the use of the ChatNoir search endpoint [30], which indexes ClueWeb22 [28] and retrieves using BM25.

3.3 Further Investigation

Beyond our reproduction setting, we utilise re-annotated query-document pairs to investigate other aspects of collection reliability.

Judgment Aggregation. Voorhees [46] apply several aggregation operations including combination as described in Section 3.1 and aggregate operators. We perform maximum, minimum, and mean pooling over judgments. In the case of mean pooling, we do not perform further quantization; for example, three texts of grades 1, 1.5, and 2, where the second text was annotated 1 and 2 by different annotators, would be compared in that order.

Absent Narratives. The absence of narratives or larger descriptions for the TREC Deep Learning collections allows an additional focus for our investigation: Annotators were prompted to note cases of ambiguity and, more generally, their difficulties in annotation. Furthermore, annotators were asked to create their narratives for each query so that agreement among annotators could be investigated with more granularity in the presence of a natural language description of their thought process during annotation. In doing so, we measure query intent as a factor in agreement.

To ablate this factor, we fix narratives after the curation of judgments to assess how narrative interpretation can affect agreement and downstream measures of effectiveness. We take three topics of low, median, and high agreement measured via Fleiss κ between the primary and secondary annotators; half of the annotators reference one narrative, and half reference the other. We can then further isolate sources of disagreement to query intent and intrinsic ambiguity of a query title, both qualitatively and quantitatively.

Annotators as Oracles. User models underlie many common IR measures such as discounted cumulative gain [19] and mean reciprocal rank ($@k$) [5]. In performing offline evaluations of search systems using these measures, we aim to model effectiveness via user behaviour approximated by relevance judgments. In principle, an optimal system would approach a score of 1 for a normalized metric (assuming k relevant documents exist). However, given the subjective and variable nature of relevance under different query intents, such effectiveness, though possible, may not be indicative of a truly effective system. Human annotators outside the original annotation setting may represent strong search systems as they have exhaustively annotated available documents to some cutoff k. As such, we measure the performance of each annotator under original relevance judgments as an indicator of how non-pooled modern systems are performing relative to a more realistic oracle. Similarly to idealized discounted cumulative gain, we sort by relevance grade post-aggregation. Though graded relevance lacks granularity in contrast to scalar relevance scores, through in-grade permutations (shuffling), we validate that variations in downstream measures of effectiveness are minimal (SD < 0.0001).

4 Evaluation

In this section, we outline our investigation and analyze findings when evaluating using secondary annotations.

4.1 Experiment Design

Following the re-annotation process of Voorhees [46], we observe the impact of re-annotation on the DL'19 test collection, using combinations to assess modern system bias.

Dataset. The MSMARCO passage corpus is a collection of around 8.8 million segmented documents covering a broad range of ad-hoc search topics [26]; this collection and accompanying training topics were mined and annotated from Bing query logs. We re-annotate the TREC Deep Learning 2019 test collection [8], composed of 43 topics. Unlike the development split released originally with MSMARCO, the DL'19 test set uses densely annotated depth-pooled topics taken from the associated TREC track that year.

Pool Description. The DL'19 track primarily focused on the comparison of neural systems in ad-hoc ranking. For equivalence to the re-ranking task, the top 100 texts of all submitted systems from both re-ranking and retrieval tasks were included in our annotation pool. The top 10 texts from each of the 75 systems (44% Neural LM, 27% Neural, 29% lexical/traditional) were added to our pool, and HiCal [1] is used to collect an additional 100 texts per topic. Our approach does not annotate all texts and thus uses a subset of this pool.

Models. We compare several neural architectures and training settings. In addition to models that contributed to the original annotation pool, we include systems that use other models for either data

Table 2: Modern systems included in our analysis. Rightmost columns are the source of labels for semi-supervised learning; neural is any encoder-type model, and LLM is a Large Language Model. Absence of both indicates binary labels.

		Distillation	
Model	Type	Neural	LLM
ColBERT [22]	Late interaction bi-encoder	×	×
SPLADE ED++ [11, 12]	Learned sparse encoder	✓	×
MonoT5 [27]	Seq2Seq cross-encoder	×	×
MonoELECTRA [31, 40]	BERT-based cross-encoder	✓	✓
SetEncoder [39]	BERT-based list-wise cross-encoder	✓	✓
RankZephyr [32]	Decoder-only list-Wise cross-encoder	×	✓
RankGPT [43]	Decoder-only list-wise cross-encoder	?	?

augmentation or semi-supervised training via distillation. Table 2 illustrates the degree of teacher distillation used in each architecture; ColBERT and SPLADE act as end-to-end retrieval or first-stage retrievers, whereas all others re-rank BM25 [35] or ColBERT as noted. For monoT5, monoELECTRA, and SetEncoder, we include multiple model sizes, base and large for BERT/ELECTRA variants and base and 3B for T5. For monoELECTRA, we do not use the original checkpoints released by Pradeep et al. [31]; we instead use distilled checkpoints released by Schlatt et al. [40] as we are interested in systems that use modern processes. When using RankGPT, we employ GPT-4, GPT-4 Turbo, and GPT-4o.

Measures. Voorhees [46] measure inter-annotator agreement in a pair-wise setting using the overlap of judgments defined as the number of common judgments divided by the total number of judgments. We additionally use Cohen's κ over both 4-grade and binarized relevance and Fleiss' κ, allowing for comparisons between 3 annotators (1 primary, 2 secondary) simultaneously. In the TREC-4 and 6 test collections, mean-averaged precision (MAP) was the primary measure of effectiveness, and therefore, the previous study measured systems by MAP. The TREC DL tracks with four grades of relevance instead primarily measure nDCG@10, and thus, our study is conducted over this metric. Aligned with Voorhees [46], we measure downstream system order correlation between judgment sets using measure Kendall's τ; additionally, we measure Spearman's ρ and RBO.

System Order Robustness. As we have multiple annotators per query, we can observe pooled judgments across several hypothetical sets; we take all possible combinations of annotators and evaluate both pooled and new models to observe the stability of system order under new judgments. For a given judgment set, we produce a ranking of systems by a target metric (i.e., nDCG@10) and compare this system order to the system ordering under the original judgments. We aggregate rank Δ for all possible natural combinations over each system to observe how frequently a system degrades over other assessed systems. We measure significant differences in retrieval effectiveness via a paired t-test with Bonferroni correction for different annotator combinations. We measure significant changes in system rank by a Wilcoxon signed test comparing annotator combinations with the system order by official TREC judgments.

Table 3: Comparing the ration of annotation grades between the primary annotators and secondary annotators in a pilot study of 10 topics.

	Grade Ratio				
Annotators	0	1	2	3	$|A|$
Primary	0.487	0.190	0.308	0.015	1143
Secondary	0.560	0.213	0.108	0.118	600

Table 4: Inter-Annotator agreement over combinations of annotators. Agreement is shown under binarised and 4-grade relevance.

	Fleiss' κ		Overlap		Cohen's κ		
Group	4	2	4	2	4	2	$|Q|$
Primary+1	0.22	0.44	0.12	0.46	0.12	0.31	14
1	—	—	0.42	0.72	0.19	0.37	
Primary+2	0.28	0.40	0.17	0.48	0.17	0.31	12
2	—	—	0.47	0.74	0.22	0.38	
Primary+3	0.17	0.31	0.11	0.45	0.11	0.28	8
3	—	—	0.63	0.89	0.27	0.47	
Primary+4	0.28	0.37	0.19	0.48	0.19	0.30	9
4	—	—	0.43	0.71	0.19	0.34	

4.2 Pilot Study Findings

We conducted a pilot study over the top 10 pooled texts for 10 randomly sampled topics. Table 3 summarizes the results. We observe that the prevalence of grades 1 and 3 is inflated in these annotations compared to the original annotations. In discussions with annotators, there was ambiguity about what separates a relevant text from a highly relevant text. Additionally, cultural context leads to ambiguity in interpreting a query; for example, a query "dog day afternoon meaning" was difficult to interpret without broader context, as a film exists with that name. The query could be interpreted as the meaning of the phrase or the film. From these ambiguities, we allowed annotators to use a controlled search engine to familiarise themselves with the topic. Annotators not from the USA generally noted the difficulty in USA-centric topics; however, we leave bias and query difficulty by nationality to future work, as this line of study is not central to our hypotheses.

4.3 Core Assessment of Agreement

Quantification of relevance is confounded by many factors with several works reporting what is generally considered low agreement among annotators [4, 23, 46]. Voorhees [46] report overlap among annotators as a measure of agreement, observing values around 0.4. In Table 4, we present agreement measurements over both 4-grade and binarized relevance. We observe that under 4-grade relevance, all agreement values degrade relative to those stated in previous studies. Fleiss' κ measured over primary and secondary annotators indicates near-random agreement but generally improves by around 10 points in all cases over binary judgments. Overlap values
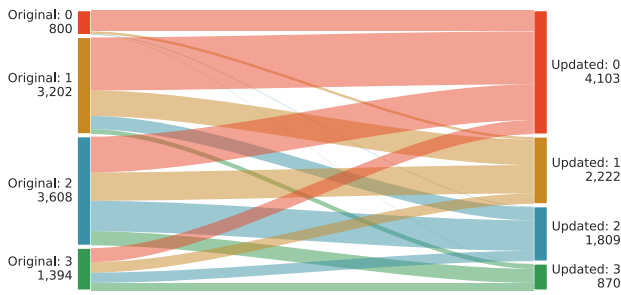
Original: 0 — 800
Original: 1 — 3,202
Original: 2 — 3,608
Original: 3 — 1,394

Updated: 0 — 4,103
Updated: 1 — 2,222
Updated: 2 — 1,809
Updated: 3 — 870

Figure 1: Transitions in relevance grades from primary to secondary annotations.

over binary judgments generally correspond to those observed by Voorhees [46], and over solely secondary annotators, we observe high overlap values, suggesting that while disagreement stems from the subjectivity of relevance and the annotation setting, the process of annotating over 4 grades is generally more variable regardless of topic granularity. In the third group, we find the lowest values in both 4 grade and binary relevance, which is again reflected in low overlap and Cohen's κ values, potentially suggesting diverging query intents or that this allocation of queries was particularly ambiguous. However, we see increases in overlap values when measured solely over secondary annotators, again suggesting that the annotation setting under which judgments are collected contributes to the improved agreement, as noted by Cuadra and Katter [10] and Regazzi [34] in previous annotation studies. Values of κ are low in all cases, suggesting that generally quantifying relevance over 4 grades is more challenging for annotators than the re-annotation setting of Voorhees [46]. The lack of pre-defined narratives clarifying intent means annotators could interpret a query as they please, potentially resulting in three diverging topic interpretations.

In Figure 1, observe a Sankey flow diagram between the original relevance grade assigned by primary annotators and the new grades by secondary annotators. Like previous work, a secondary annotation setting often leads to the reduction of relevance grades, e.g., 1 to 0 or 2 to 1. Validating our decision to sample a subset of non-relevant judgments, 13% of 800 non-relevant texts change grade, with 77% being a change to a grade of 1; we analyse outlier cases (providing a grade of 2 or more) in Section 4.4. Measuring frequencies agnostic of previous judgments, grades 1 and 2 are annotated similarly across annotators, while grades 0 and 3 were annotated with greater variance. This variance in the case of grade 0 is generally consistent across annotators, and we provide a deeper analysis of grade 3 due to varying notions of relevance below.

4.4 Narratives, Intent, and Relevance

The observations of grade transitions in our work and generally low agreement noted by prior work, even in a binary setting, may be attributed to several factors noted in Section 2. Though we cannot fully isolate factors in the subjectivity of relevance, we choose to investigate the effect of narratives on inter-annotator agreement, given that recent corpora often do not have fixed narratives [7, 8],

unlike those used in prior re-annotation studies [14]. We select topics by median, minimum, and max Fleiss' κ values from the annotation pool and consider where disagreement occurs and how narratives differ. Chosen examples illustrate different aspects of agreement, differing query intents, and a narrative's specificity.

The first, taken from Group 1, is a technical query in the medical domain with a Fleiss κ of 0.06, indicating disagreement near random guessing. The query "types of dysarthria from cerebral palsy" was described by one annotator as a query from a sibling of someone with cerebral palsy looking to understand how their sibling would be affected by their condition in the future. The other annotator described a medical context in which the definition of all conditions is known, and therefore, solely the connection between dysarthria and cerebral palsy was required. The second annotator stated a requirement that relevance was determined by the types and specificity of provided types of dysarthria connected to cerebral palsy, with all other documents being non-relevant. Though both are valid information needs, they represent vastly different query intents, with one annotator's grade frequency matching the overall annotation pool and the other stricter narrative leading to near-total non-relevance.

The second, taken from Group 4, is a more general query, "is cdg airport in main paris" with a Fleiss κ of 0.08, again indicating low agreement. However, both annotators described a need for exact distances from the airport to either the center of Paris or, more generally, the city of Paris. Both annotators required a succinct description of the distance or travel times to the airport from Paris to be considered perfectly relevant. To be highly relevant, both considered any mention of either travel time or distance measurements. Intents then diverged for partially relevant documents, with one annotator allowing for the mention of costs and other aspects of travel to and from the airport. In contrast, the other annotator required the airport's position, at minimum, mentioning that it is situated with respect to Paris. Non-relevance was denoted by any other mention of the airport or mentions of distances to destinations near Paris. We find larger differences in lower relevance grades, with one annotator considering the majority of pooled documents to be non-relevant when there is no mention of travel details. In contrast, the other considers them to be relevant because they mention that the airport is situated near Paris in a generic manner. The specificity of a query generally leads to lower agreement in our observations. While queries in domains such as law or medicine have a lower overall agreement, we still find cases where generic queries have low agreement due to the interpretations of information needs.

We analyze outlier cases of texts initially labeled non-relevant that, when evaluated under an alternative query intent, were deemed perfectly relevant. In the 9 cases of this, there are cases where a reference is made to the entity of focus, for example "goldfish" in the query, "do goldfish grow" but only implicitly, for example "it" or "they". The query was satisfied by a given document if the context is inferred by the surrounding spans of text or from knowing that the document has been retrieved by some system (this is inevitable under pooling), but the information is not self-contained. In some cases, annotators explicitly stated that they considered a document relevant, but when the information is not self-contained, it cannot be annotated as relevant; others took this approach without note.

Table 5: System ranking correlation over different annotator groups measuring Kendall's τ, Spearmans's ρ and rank-biased overlap. 'Combination' is the mean of natural permutations of the 8 annotators, 'In-Sample' is the mean of judgment permutations as described by Voorhees [46]. All τ and ρ values represent significant correlation.

Type	τ	ρ	RBO
Combination	0.879	0.972	0.888
In-Sample	0.897	0.977	0.902

Overall, many annotation disagreements can occur not only in the high-level query intent of the narrative but also in descriptions of relevance. A diagnostic tool and query-specific grade specifications could be helpful in future work.

4.5 Narrative and Intent Ambiguity

Given the lack of public narratives provided in the release of TREC Deep Learning 2019, it is infeasible to isolate query intent as a factor in reliability without first collecting narratives. In fixing narratives, we produce an additional set of relevance judgments for queries of low, median, and high agreement by Fleiss κ, including the primary annotator. In Figure 2, We classify these queries as difficult, median and easy with respect to query intent determination and measure Cohen's κ between the judgments of both narrative groupings. It is clear that a fixed narrative reduces ambiguity with Cohen's κ, improving by 24% and 46% for each annotation group between the difficult and median query; however, between the median and easy query, a decrease of 8% and an increase of 19% is observed.

In several cases, high agreement across different narrative groups occurs; in each case, narratives are of similar detail, defining what intent the user had in searching as well as what should be labelled relevant; we consider that one factor in relevance annotation beyond random noise may therefore be interpretation of a narrative which in the case of a primary annotator is of minimal concern but in human-in-the-loop annotation with generative systems may lead to implications for control as ultimately we can only express our need or a demonstration to the system in natural language. When computing global improvements in agreement, we measure a 79% increase in agreement between the difficult and median query and an additional 1% increase comparing the median and easy query. From these observations, we propose that the greatest factor in disagreement is simply the ambiguity of the natural language used to express a query regardless of the narrative applied. Whether or not this is only inherent in pooled corpora would require further investigation as in a setting such as that of TREC DL'19, the ad-hoc realisation of a query intent will be biased by the top-k texts chosen by pooled systems.

4.6 Core Assessment of System Order

The core objective in re-annotation studies is to validate that downstream evaluation is not compromised by human bias in subjective relevance judgments. By measuring the stability of system order and the correlation of system order by target metrics, we can assess collection reliability. Following Voorhees [46] and as described in

Section 3.1, we sample combinations of relevance judgments and observe the stability of pair-wise system comparisons in Figure 3. Comparisons closest to the top right of the figure represent those where nDCG differences are large; however, they frequently swap in system order. A clear pareto-frontier is composed of comparisons between LLM and neural systems under judgment combinations. Lexical-neural comparisons generally lie in the main body of neural-neural and llm-llm comparisons, suggesting similar stability as was noted by Voorhees [46] over TREC-6. However, more generally, we replicate the findings of previous works that despite low annotator agreement, under re-annotation, system ordering is highly correlated with the original DL'19 annotations even under 4-grade relevance. In Table 5, observe aggregate correlation values using both natural combinations of annotators and the in-sample approach described by Voorhees [46].

As the In-Sample approach includes primary judgments, a marginally higher correlation compared to combinations of solely secondary annotators is found. To investigate where system order diverges, we analyse ranking changes at a system level. In Figure 4 see that generally, new annotations led to a reduction in absolute nDCG@10 values relative to the original annotations, effectiveness measures on original annotations order systems, and that system order is not fully stable but LLM-based or distilled systems appear to be outliers, as such we explicitly measure the changes in system order concentrated around these systems. In Figure 5, we combine all possible combinations of annotators across annotation pairs. We measure changes in system order under these combinations and aggregate by system. Many lexical systems remain stable in ordering ($\Delta \approx 0$). Older neural approaches in the original system pool display higher variance but can improve or degrade in rank without a significant difference. When considering only modern systems, we observe that smaller models trained with supervised and semi-supervised approaches consistently improve system rank. Within this group, however, larger semi-supervised models are more robust than their smaller counterparts under new judgments. List-wise encoders generally degrade in rank. These models use an interaction token to facilitate inter-document attention, assisting in out-of-domain effectiveness [39]. Changes in effectiveness are more pronounced and are significant in smaller distilled language models. This is notable as though we see continuing improvement in smaller models out-of-domain through distillation. One should be cautious that in-domain effectiveness may be less robust than benchmarks indicate, which for many applications may be more important than robust effectiveness across heterogeneous corpora.

Runs that are either zero-shot or distilled from LLM re-rankers show the highest variance in relative effectiveness even within a sub-group such as those based on GPT 4 endpoints. GPT-4o-based runs significantly degrade in system rank when re-ranking BM25 but are more stable under a neural first-stage ColBERT. This is most likely caused by positional bias found in list-wise ranking towards already precise first-stage rankings [29, 44]. This hypothesis is validated by the frequent rank improvement of RankZephyr (the left-most model in Figure 4), which was trained explicitly to reduce this bias [32]. Overall, we consider that though high-quality data (i.e., granular training data from strong models) is important in improving the robustness of neural models, these findings would suggest that smaller models in our current training strategies may not have the capacity to robustly generalize even in-domain.
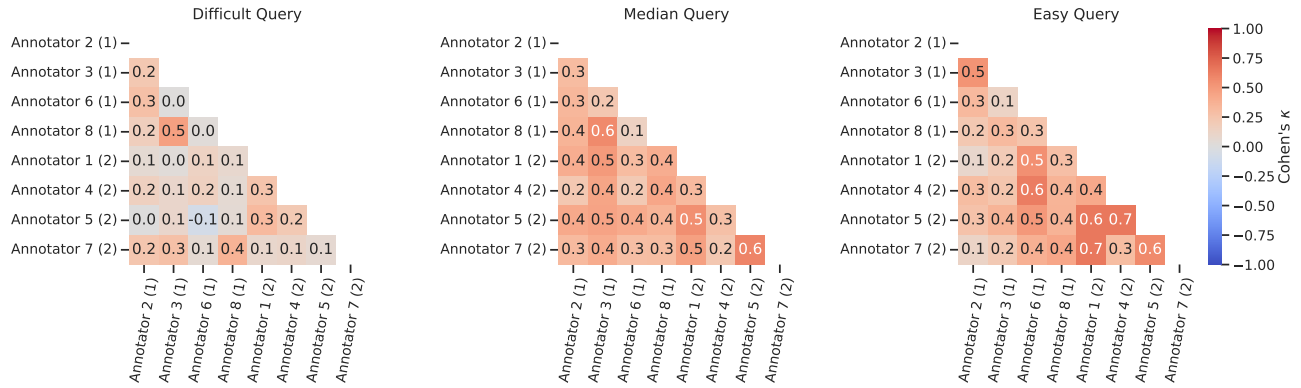
Figure 2: Comparing 3 queries of varying difficulty by inter-annotator agreement under fixed narratives (2 narratives (1) and (2)). Annotators were divided into two groups and were assigned identical document pools. Color indicates strength of agreement; redder indicates stronger agreement, bluer indicates weaker agreement.
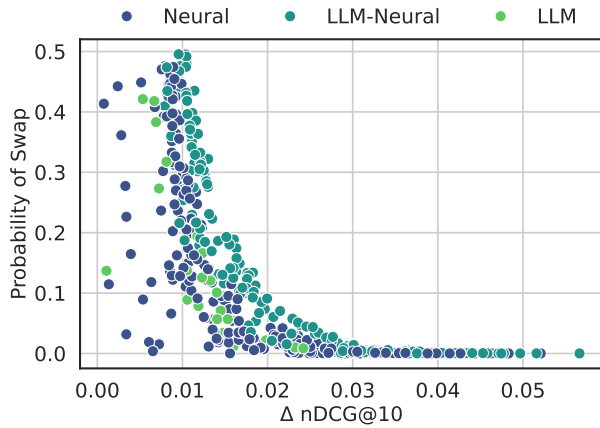

Figure 3: Comparing the stability of judgments through nDCG@10 and system swap probabilities over 10000 judgment combinations. We filter lexical comparisons to improve visibility in the central mass. 'LLM-Neural' denotes that a pair contains one LLM-based and one Neural-based ranker.
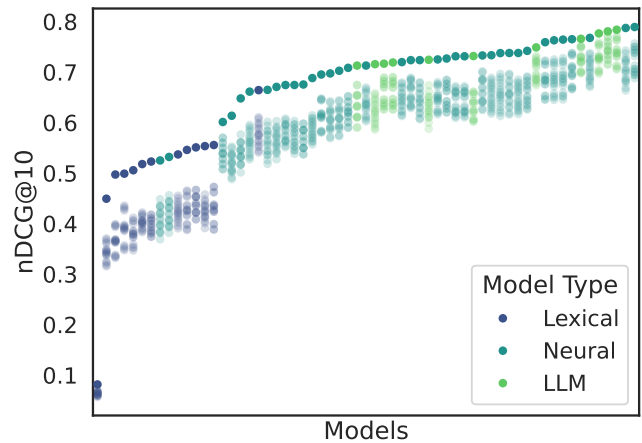

Figure 4: Variance in nDCG@10 over annotator combinations. Effectiveness values on new annotations are translucent. Color of points indicates model type.

4.7 Systems versus Humans

Recall our argument that a "perfect" score on a normalized metric may not be a feasible or desirable goal in ranking evaluation; as such, we measure the effectiveness of human annotators aggregated in multiple forms outlined in Section 3.3. In Table 6, see that systems can exceed or reach parity with secondary annotators, particularly with respect to nDCG@10 and MRR@10, with multiple systems, both LLM and encoder-based, exceeding annotators. When excluding unjudged documents (which is known to overestimate the retrieval effectiveness [13, 36]), annotator performance measured by precision can be exceeded by neural systems across nDCG, P, and MRR; however, recall remains lower in all settings. Given our findings in Figure 4, an observation that systems are

better than humans should be considered cautiously. However, the robustness of the LLM-based RankZephyr observed both by system order robustness over new judgments and its parity with human annotators suggests that with sufficient capacity, a smaller system can be distilled from larger systems while remaining robust. Observing similar effectiveness to humans and confidence intervals providing no way to discriminate systems, we consider that achieving significant improvements would suggest far exceeding humans, which may not be possible in a way conducive to improving information access and may solely serve to improve precision on specific topics and intents. The ability of distilled systems to reach parity with human effectiveness across multiple annotator combinations while
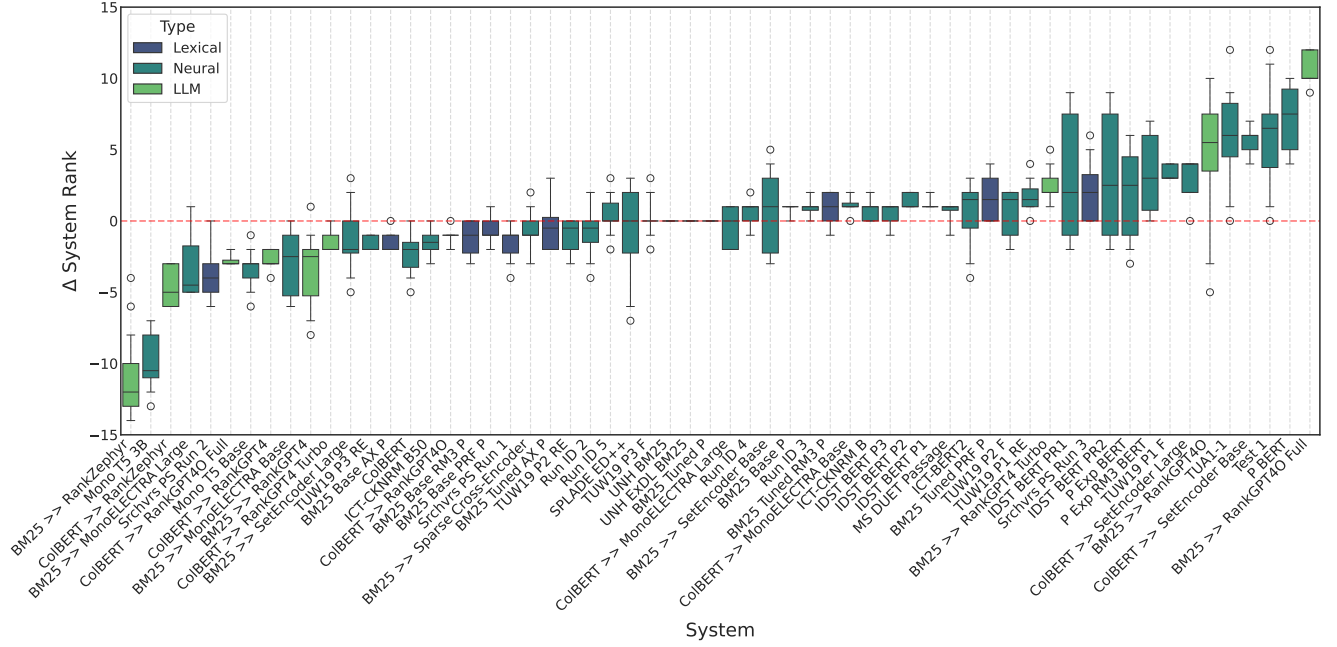
Figure 5: Aggregated system rank changes over all natural combinations of secondary versus primary annotations measured using nDCG@10. Negative Δ indicates a system frequently improves in rank.

Table 6: Comparing aggregate annotations as rankings with several ranking architectures. Judged Only indicates that original runs have been filtered to contain solely annotated documents, allowing for a similar setting to the annotator rankings. Significance is with respect to Minimum (A), Mean (B) or Maximum (C) aggregation via a paired t-test with Bonferroni correction ($p < 0.05$).

Model	nDCG@10	P@10	MRR@10	R@100
Judged & Unjudged				
BM25	0.51 ± 0.08^{ABC}	0.41 ± 0.09^{ABC}	0.70 ± 0.12^{BC}	0.49 ± 0.10^{ABC}
Splade ED++	0.73 ± 0.07^{BC}	0.62 ± 0.10^{BC}	0.91 ± 0.07	0.60 ± 0.09^{ABC}
ColBERT»RankZephyr	0.75 ± 0.07^{B}	0.67 ± 0.10	0.84 ± 0.09	0.67 ± 0.09^{ABC}
ColBERT»RankGPT-4o	0.78 ± 0.06	0.71 ± 0.09	0.87 ± 0.08	0.67 ± 0.09^{ABC}
Judged Only				
BM25	0.51 ± 0.08^{ABC}	0.41 ± 0.09^{ABC}	0.70 ± 0.12^{BC}	0.65 ± 0.09^{ABC}
Splade ED++	0.73 ± 0.07^{B}	0.63 ± 0.10^{BC}	0.91 ± 0.07	0.69 ± 0.09^{BC}
ColBERT»RankZephyr	0.77 ± 0.07	0.70 ± 0.09	0.85 ± 0.08	0.67 ± 0.09^{ABC}
ColBERT»RankGPT-4o	0.80 ± 0.05	0.73 ± 0.09	0.89 ± 0.08	0.67 ± 0.09^{ABC}
Minimum (**A**)	0.76 ± 0.07	0.67 ± 0.10	0.84 ± 0.09	0.75 ± 0.06
Mean (**B**)	0.81 ± 0.05	0.71 ± 0.10	0.90 ± 0.08	0.86 ± 0.06
Maximum (**C**)	0.79 ± 0.05	0.70 ± 0.09	0.86 ± 0.08	0.86 ± 0.06

obtaining largely reduced recall in a judged setting, even when re-ranking a neural first-stage retriever, is undesirable for several downstream tasks. Although precision is approaching or overtaking the effectiveness of an oracle, recall remains a challenge in densely annotated collections.

5 Conclusion

We considered important studies on the impact of the annotator agreement on retrieval evaluations in light of modern neural evaluation scenarios. In reproducing re-annotation processes and evaluations by Voorhees [46] on a modern test collection, we validate the stability of system ordering under re-annotation under 4-grade relevance and ambiguous query intent. We raised concerns that the broad and frequent use of test collections to "tune" learned systems implicitly through influences from published works may reduce the reliability of a collection over time. We identified that the subjectivity of relevance through query intent may be a factor in determining "overfitting" in ranking tasks. As a query may have several interpretations, we re-annotate the popular TREC DL'19 collection. We found that system order varies when evaluating large language models and systems distilled from them. Furthermore, we observed that the current state-of-the-art can outperform combinations of human annotators on original relevance judgments, suggesting we may have reached a realistic bound on precision for this collection. We posit that further improvement on this collection may not indicate that a system is better than another in a meaningful way. Our process has limitations—most notably, re-annotation is incredibly costly. Future work will investigate how to draw similar conclusions about a collection's expiration without the need for extensive human effort.

Acknowledgments

References

[1] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. 2018. A System for Efficient High-Recall Retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 1317–1320.

[2] Xavier Amatriain and Justin Basilico. 2017. Netflix recommendations: Beyond the 5 stars (part 1). https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429

[3] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 601–610. doi:10.1145/1645953.1646031

[4] Robert Burgin. 1992. Variations in relevance judgments and the evaluation of retrieval performance. *Inf. Process. Manage.* 28, 5 (July 1992), 619–627. doi:10.1016/0306-4573(92)90031-T

[5] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*. New York, NY, USA, 621–630. http://doi.acm.org/10.1145/1645953.1646033

[6] C. Cleverdon, J. Mills, and M. Keen. 1966. *Factors Determining the Performance of Indexing Systems. Volume I. Design. Part 2. Appendices.* Technical Report PB169574. Association of Special Libraries and Information Bureau, Cranfield (England). https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB169574.xhtml Num Pages: 261.

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2020 Deep Learning Track. In *Proceedings of the 29th Text REtrieval Conference, TREC 2020, Virtual Event, Gaithersburg, MD, USA, November 16-20, 2020 (NIST Special Publication, Vol. 1266)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[8] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2019. Overview of the TREC 2019 Deep Learning Track. In *28th International Text Retrieval Conference, TREC 2019, Gaithersburg, Maryland, USA (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[9] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Fernando Campos, Ellen M. Voorhees, and Ian Soboroff. 2021. TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021). https://api.semanticscholar.org/CorpusID:233296851

[10] Carlos A. Cuadra and Robert V. Katter. 1967. Opening the Black Box of Relevance. *Journal of Documentation* 23, 4 (April 1967), 291–303. doi:10.1108/eb026436

[11] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2353–2359. doi:10.1145/3477495.3531857

[12] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. *SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking.* Association for Computing Machinery, New York, NY, USA, 2288–2292. https://doi.org/10.1145/3404835.3463098

[13] Maik Fröbe, Lukas Gienapp, Martin Potthast, and Matthias Hagen. 2023. Bootstrapped nDCG Estimation in the Presence of Unjudged Documents. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13980)*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, 313–329. doi:10.1007/978-3-031-28244-7_20

[14] Donna Harman. 1995. Overview of the Fourth Text REtrieval Conference (TREC-4). In *Proceedings of The Fourth Text REtrieval Conference, TREC 1995, Gaithersburg, Maryland, USA, November 1-3, 1995 (NIST Special Publication, Vol. 500-236)*, Donna K. Harman (Ed.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec4/overview.ps.gz

[15] Donna Harman. 2012. TREC-Style Evaluations. In *Information Retrieval Meets Information Visualization - PROMISE Winter School 2012, Zinal, Switzerland, January 23-27, 2012, Revised Tutorial Lectures (Lecture Notes in Computer Science, Vol. 7757)*, Maristella Agosti, Nicola Ferro, Pamela Forner, Henning Müller, and Giuseppe Santucci (Eds.). Springer, 97–115. doi:10.1007/978-3-642-36415-0_7

[16] Geoffrey Hinton. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[17] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation. *CoRR* abs/2010.02666 (2020). arXiv:2010.02666 https://arxiv.org/abs/2010.02666

[18] Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Linguistics and Language Compass* 15, 8 (aug 2021). doi:10.1111/lnc3.12432

[19] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446. doi:10.1145/582415.582418

[20] Jaap Kamps and David Rau. [n. d.]. University of Amsterdam at TREC 2021: Deep Learning Track. https://api.semanticscholar.org/CorpusID:266088115

[21] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 6769–6781. doi:10.18653/V1/2020.EMNLP-MAIN.550

[22] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 39–48. doi:10.1145/3397271.3401075

[23] M.E. Lesk and G. Salton. 1968. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval* 4, 4 (Dec. 1968), 343–359. doi:10.1016/0020-0271(68)90029-6

[24] Sean MacAvaney, Adam Roegiest, Aldo Lipani, Andrew Parry, Björn Engelmann, Christin Katharina Kreutz, Chuan Meng, Erlend Frayling, Eugene Yang, Ferdinand Schlatt, Guglielmo Faggioli, Harrisen Scells, Iana Atanassova, Jana Friese, Janek Bevendorff, Javier Sanz-Cruzado, Johanne Trippas, Kanaad Pathak, Kaustubh D. Dhole, Leif Azzopardi, Maik Fröbe, Marc Bertin, Nishchal Prasad, Saber Zerhoudi, Shuai Wang, Shubham Chatterjee, Thomas Jänich, Udo Kruschwitz, Xi Wang, and Zijun Long. 2024. Report on the Collab-a-Thon at ECIR 2024. *SIGIR Forum* 58, 1 (2024), 1–11. doi:10.1145/3687273.3687287

[25] Stefano Mizzaro. 1997. Relevance: The Whole History. *J. Am. Soc. Inf. Sci.* 48, 9 (1997), 810–832.

[26] 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.

[27] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 708–718. doi:10.18653/v1/2020.findings-emnlp.63

[28] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Rich Information. 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, 3360–3362. doi:10.1145/3477495.3536321

[29] Andrew Parry, Sean MacAvaney, and Debasis Ganguly. 2024. Top-Down Partitioning for Efficient List-Wise Ranking. arXiv:2405.14589 [cs.IR] https://arxiv.org/abs/2405.14589

[30] Martin Potthast, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. 2012. ChatNoir: a search engine for the ClueWeb09 corpus. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson (Eds.). ACM, 1004. doi:10.1145/2348283.2348429

[31] Ronak Pradeep, Yuqi Liu, Xinyu Zhang, Yilin Li, Andrew Yates, and Jimmy Lin. 2022. Squeezing Water from a Stone: A Bag of Tricks for Further Improving Cross-Encoder Effectiveness for Reranking. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I* (Stavanger, Norway). Springer-Verlag, Berlin, Heidelberg, 655–670. doi:10.1007/978-3-030-99736-6_44

[32] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze! *CoRR* abs/2312.02724 (2023). doi:10.48550/ARXIV.2312.02724 arXiv:2312.02724

[33] Alan M. Rees and Douglas G. Schultz. 1967. *A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. Final Report to the National Science Foundation. Volume I.* Technical Report. Clearinghouse for Federal Scientific and Technical Information, Springfield, Va.

[34] John J. Regazzi. 1988. Performance measures for information retrieval systems—an experimental approach. *Journal of the American Society for Information Science* 39, 4 (1988), 235–251. doi:10.1002/(SICI)1097-4571(198807)39:4<235::AID-ASI3>3.0.CO;2-H

[35] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Mike Gatford, and A. Payne. 1995. Okapi at TREC-4. In *Proceedings of The Fourth Text REtrieval Conference, TREC 1995, Gaithersburg, Maryland, USA, November 1-3, 1995 (NIST Special Publication, Vol. 500-236)*, Donna K. Harman (Ed.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec4/papers/city.ps.gz

[36] Tetsuya Sakai. 2008. Comparing metrics across TREC and NTCIR: the robustness to system bias. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury (Eds.). ACM, 581–590. doi:10.1145/1458082.1458159

[37] Tetsuya Sakai. 2019. How to Run an Evaluation Task - With a Primary Focus on Ad Hoc Information Retrieval. In *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, Nicola Ferro and Carol Peters (Eds.). The Information Retrieval Series, Vol. 41. Springer, 71–102. doi:10.1007/978-3-030-22948-1_3

[38] Linda Schamber, Michael B. Eisenberg, and Michael S. Nilan. 1990. A Re-Examination of Relevance: Toward a Dynamic, Situational Definition∗. *Information Processing & Management* 26, 6 (Jan. 1990), 755–776. doi:10.1016/0306-4573(90)90050-C

[39] Ferdinand Schlatt, Maik Fröbe, Harrisen Scells, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Benno Stein, Martin Potthast, and Matthias Hagen. 2024. Set-Encoder: Permutation-Invariant Inter-Passage Attention for Listwise Passage Re-Ranking with Cross-Encoders. *arXiv preprint arXiv:2404.06912* (2024).

[40] Ferdinand Schlatt, Maik Fröbe, Harrisen Scells, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Benno Stein, Martin Potthast, and Matthias Hagen. 2024. A Systematic Investigation of Distilling Large Language Models into Cross-Encoders for Passage Re-ranking. arXiv:2405.07920 [cs.IR] https://arxiv.org/abs/2405.07920

[41] Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Xiaolong Huang, Binxing Jiao, Linjun Yang, and Daxin Jiang. 2022. LexMAE: Lexicon-Bottlenecked Pretraining for Large-Scale Retrieval. *ArXiv* abs/2208.14754 (2022). https://api.semanticscholar.org/CorpusID:251953412

[42] Ian Soboroff. 2024. Don't Use LLMs to Make Relevance Judgments. arXiv:2409.15133 [cs.IR] https://arxiv.org/abs/2409.15133

[43] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14918–14937. doi:10.18653/v1/2023.emnlp-main.923

[44] Yubao Tang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, and Xueqi Cheng. 2024. Listwise Generative Retrieval Models via a Sequential Learning Process. *ACM Trans. Inf. Syst.* 42, 5 (2024), 133:1–133:31. doi:10.1145/3653712

[45] Mortimer Taube. 1965. A Note on the Pseudo-Mathematics of Relevance. *American Documentation* 16, 2 (1965), 69–72. doi:10.1002/asi.5090160204

[46] Ellen Voorhees. 2000. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. 36 No. 5 (2000-01-01 2000).

[47] Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 315–323. doi:10.1145/290941.291017

[48] Ellen M. Voorhees. 2019. The Evolution of Cranfield. In *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, Nicola Ferro and Carol Peters (Eds.). The Information Retrieval Series, Vol. 41. Springer, 45–69. doi:10.1007/978-3-030-22948-1_2

[49] Ellen M. Voorhees, Nick Craswell, and Jimmy Lin. 2022. Too Many Relevants: Whither Cranfield Test Collections?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2970–2980. doi:10.1145/3477495.3531728

[50] Ellen M. Voorhees and Donna K. Harman. 1997. Overview of the Sixth Text REtrieval Conference (TREC-6). *Inf. Process. Manag.* 36 (1997), 3–35. https://api.semanticscholar.org/CorpusID:9854240

[51] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. SimLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:250311114

[52] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2023. SimLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 2244–2258. doi:10.18653/v1/2023.acl-long.125

[53] Xing Wu, Guangyuan Ma, Meng Lin, Zijia Lin, Zhongyuan Wang, and Songlin Hu. 2023. Contextual masked auto-encoder for dense passage retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4738–4746.

[54] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 307–314. doi:10.1145/290941.291014