Evaluating D-MERIT of Partial-annotation on Information Retrieval

Anonymous ACL submission

Abstract

Retrieval models are often evaluated on partially-annotated datasets. Each query is mapped to a few relevant texts and the remaining corpus is assumed to be irrelevant. As a result, models that successfully retrieve false negatives are punished in evaluation. Unfortunately, completely annotating all texts for every query is not resource efficient. In this work, we show that using partially-annotated datasets in evaluation can paint a distorted picture. We curate D-MERIT, a passage retrieval evaluation set from Wikipedia, aspiring to contain *all* relevant passages for each query. Queries describe a group (e.g., "journals about linguistics") and relevant passages are evidence that entities belong to the group (e.g., a passage indicating that *Language* is a journal about linguistics). We show that evaluating on a dataset containing annotations for only a subset of the relevant passages might result in misleading ranking of the retrieval systems and that as more relevant texts are included in the evaluation set, the rankings converge. We propose our dataset as a resource for evaluation and our study as a recommendation for balance between resource-efficiency and reliable evaluation when annotating evaluation sets for text retrieval.

1 Introduction

Passage retrieval, the task of retrieving relevant passages for a given query from a large corpus, is a traditional IR task (Kaszkiel and Zobel, 1997; Callan, 1994; Zobel et al., 1995). Within NLP, it has many applications, such as Open-Domain Question-Answering (ODQA) (Karpukhin et al., 2020; Zhu et al., 2021; Mavi et al., 2022; Rogers et al., 2023) and fact verification (Bekoulis et al., 2021; Murayama, 2021; Vallayil et al., 2023).

Recently, the task has experienced a renaissance due to the modern retrieval-augmented-generation setup leveraging LLMs (aka "RAG") (Lewis et al., 2021; Cai et al., 2022; Li et al., 2022). In all of those cases, retrieval makes for a crucial component of the system (Cai et al., 2022; Ram et al., 2023).

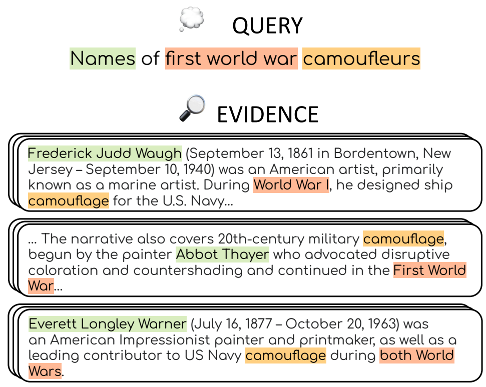

Figure 1: Demonstrating the evidence retrieval task described in Section 2.2. The query is "Names of first world war camoufleurs". Highlighted text corresponds to the query requirements: names (green), "First World War" (red), and "camouflage" (orange). A passage must match all requirements to be considered as evidence.

It is common practice, and often essential to evaluate the retriever component separately from the full system. This is done by using large-scale data resources that map queries to relevant passages.[1] The vast majority of available datasets are only partially-annotated; a query is mapped to a single (or a few) relevant passages and all other passages are assumed to be irrelevant (Bajaj et al., 2018; Kwiatkowski et al., 2019), leading to many false negatives in the dataset. This practice has long been contested (Zobel, 1998; Buckley and Voorhees, 2004; Craswell et al., 2020; Gupta and MacAvaney, 2022), yet due to the massive size of modern corpora, exhaustively annotating all passages for every query is highly impractical. As an example, MS-MARCO (Bajaj et al., 2018) con-

[1] Relevancy is defined according to the task in hand. In this work, we adopt the definition of TREC (Craswell et al., 2020), a popular retrieval research challenge.

sists of ~1M queries and ~8.8M passages, which amounts to ~8.8 *trillion* annotations.

Evaluating retrieval solutions using a partially-annotated dataset is obviously not ideal. A system retrieving a non-annotated relevant passage rather than an annotated one is unjustly penalized. Some work has been done on metrics and methods attempting to deal with this issue (Buckley and Voorhees, 2004; Yilmaz and Aslam, 2006; MacAvaney and Soldaini, 2023). However, the common practice is still using vanilla metrics (e.g. MRR, $Recall$), and the impact of partial annotation during evaluation using these metrics is still unclear. Does the ranking of systems change? Do the inaccurate scores falsely crown the wrong systems as the SOTAs? Moreover, we wonder how many relevant passages are needed in order to sufficiently reduce the error and correctly rank systems.

In this work, we propose **D-MERIT**; ***D**ataset for **M**ulti-**E**vidence **R**etrieval **T**esting*, an evaluation set for retrieval systems, *striving* to pair each query to *all* of its relevant passages. In our setting, relevant passages are evidence that some entity belongs to a group described in the query. While we use it to explore the consequences of having an evaluation dataset with only a few relevant passages annotated, D-MERIT is also highly suitable for use in high-recall settings, where the task is to retrieve as many relevant texts as possible for a given query, as it contains almost all relevant passages available in the corpus for each query.

We first show that evaluation of systems with the common single-relevant setup (for each query, annotate passages until a single relevant passage is found) is sensitive to the way in which passages were selected during annotation. As a result, different selections lead to different rankings of systems. However, we observe that when a system very significantly outperforms another ($p - value < 0.01$), representing a seminal improvement or breakthrough, the single-relevant setup is likely to provide accurate rankings. Then, we mimic partially-annotated setups, gradually adding annotated relevant passages to queries, hence reducing the number of false negatives in the data. Our findings reveal that in order to reliably evaluate retrieval systems that are reasonably close in performance, a significant portion of relevant passages must be found. This is substantial because it implies that when evaluating using partially-annotated datasets, some system might *seem* better-performing than another, while in fact, the opposite is true. To summarize, our contributions are as follows:

- D-MERIT: A publicly available passage retrieval evaluation set, aspiring to contain all relevant passages per query.
- A study on the consequences of leaving too many false negatives in evaluation sets.
- Recommendations for a balance between resource-efficiency and reliable evaluation when annotating retrieval datasets.

2 D-MERIT

2.1 Desiderata

To observe the impact of having false negatives in an evaluation set, we need to have a dataset where the false negatives are marked as such. This calls for a completely-annotated dataset, that will allow us to reliably evaluate systems' performance, as well as examine the effects of partial-annotation. To accentuate the gap between partial and full annotation, queries in the dataset should be mapped to many relevant passages. We are set to try to identify *all* relevant passages for each query, but annotating all passages for each query is unrealistic. Therefore, we desire a framework that offers inherent mappings between queries and high quality candidate passages. To push our method towards exhaustiveness, our automatic approach to candidate collection needs to lean towards recall, followed by an automatic filtering stage.

2.2 Task Definition

Evidence Retrieval. We choose evidence retrieval as our task as it naturally complements our need to collect queries with numerous relevant passages. In this task, passages are considered relevant if they contain text that can be seen as evidence that some answer satisfies the query. Previous work considering this task did not collect more than a single evidence (Malaviya et al., 2023; Amouyal et al., 2023) or did not aspire to be completely-annotated (Zhong et al., 2022). Instead, they map queries to answers, and collect evidence for each answer from a single document. Our goal is to map a query to *all* evidence in the corpus, without the limitation of a single document.

Our setup. In our setup, that can be seen as an extension of the single-evidence setup in (Malaviya et al., 2023) to an all-evidence one, a query describes a group of entities and relevant passages are evidence that an entity is a member of the group. The task is then, given a query representing some group, to retrieve all texts stating that some entity is a part of this group. For instance, Fig. 1 shows evidence for the query "names of first World War camoufleurs". The first passage confirms "Fredrick Judd Waugh" is an entity that belongs to the group of World War 1 camoufleurs. More concretely, each query lists constraints, and an evidence would associate an entity with all of them.[2] In the example above, a query describes the group of all World War 1 camoufleurs, an evidence would then need to indicate an entity (1) took part in World War 1; (2) was a camoufleur. For example, the second passage in Fig. 1 states "Abbot Thayer" advocated for coloration and countershading camouflage during World War 1, which satisfies these requirements.

2.3 Dataset Curation

We adopt the Wikipedia framework [3], which allows us to take advantage of the Wikidata structure (Vrandečić and Krötzsch, 2014) to extract groups and their corresponding members. We use the Wikipedia link network to obtain mappings between an article and all other articles referencing it. Our curation process involves three stages: (1) collecting queries and *candidates* – all passages with high likelihood of containing evidence (Section 2.3.2); (2) automatic annotation of candidate passages (Section 2.3.3); (3) generating natural language queries (Section 2.5).

2.3.1 Corpus

Our corpus is limited to the introduction section of Wikipedia articles. Without limiting our collection process to a specific section, the number of annotations per article would have multiplied by ~5, which would have made the annotation process significantly more expensive. We opted to focus on the introduction section, because it is a section that is consistent across most articles, and it is intuitive that many evidence lie there. In total, our corpus is comprised of $6,477,139$ passages.

2.3.2 Query and Candidate Collection

Extracting list members. The collection process begins by scanning articles prefixed with "list of" for tables using the Wikidata format. We extract columns with "name" in their title, as these are most likely to describe entities. Each such column is extracted separately and makes for a set of members. Columns containing empty values or values without a dedicated Wiki article are discarded.

Collecting candidates We employ the "What Links Here" feature from Wikidata. This tool provides a list of all articles that reference a specific article (and its aliases). The reference count of an article can vary significantly, even for members of the same list. For example, "Shogi" has over 600 references, while "Machi Koro" only has 9. Both appear in the group "Japanese board games". To manage this disparity and keep the candidate count feasible, we discard columns containing an article with more than $10K$ references.

2.3.3 Evidence Identification

To complete the dataset construction, we need to sift through the collected candidates. Human evaluation would have been the most reliable route, however, it does not scale. We thus turn to the current state-of-the-art large language model for automatic filtering, and show it nears human judgement.

Automatic identification. We use `GPT-4`[4] to filter $\sim 250K$ passages across $\sim 2.5K$ queries. Each prompt consists of a passage paired with a query embedded in our definition of relevance, asking the model to judge for relevance. To ensure each query is meaningful in number of evidence, queries with less than five evidence were discarded. For technical details, see Appendix C.

2.4 Evaluation of Construction Process

In order for D-MERIT to contain a significant portion of the positives for each query, some assumptions need to hold. First, Wikipedia list pages need to be exhaustive.[5] This is a common assumption also taken by (Amouyal et al., 2023) and (Malaviya et al., 2023). Our dataset construction method also relies on the accuracy of Wikipedia's linking network. This is a limitation of the method (and is therefore mentioned in the limitations section). Herein, we want to show these assumptions do not meaningfully degrade the quality of the dataset. To this end, we approximate D-MERIT's completeness and soundness by evaluating the candidate

[2] The queries in our setup are somewhat reminiscent to the intersection queries in (Malaviya et al., 2023), where a query makes for a list of requirements.

[3] The Wikidump is from July 1st, 2023.

[4] We used `GPT-4-1106-preview`. Future references to `GPT-4` refer to this version.

[5] Note that we only need the list to be exhaustive with respect to the corpus, i.e. if some set member is not in the list but is also not mentioned in Wikipedia introductions, it will not hinder the exhaustiveness of our collection method.

Query	Member	Candidate	Evidence
names of Indian Marathi romance films	Sairat	Jeur	Jeur is a village in the Karmala taluka of Solapur district in Maharashtra state, India. **Sairat**, the controversial and highest-grossing **Marathi film** of all time based on the theme of **forbidden love** was set and shot in Jeur village.
names of National Wildlife Refuges in West Virginia	Ohio River Islands National Wildlife Refuge	Mill Creek Island	Mill Creek Island is a bar island on the Ohio River in Tyler County, **West Virginia**. The island lies upstream from Grandview Island and the towns of New Matamoras, Ohio and Friendly, West Virginia. It takes its name from Mill Creek, which empties into the Ohio River from the Ohio side in its vicinity. **Mill Creek Island is protected as part of the Ohio River Islands National Wildlife Refuge**.
Names of players on 1992 US Olympic ice hockey team	Dave Tretowicz	Dave Tretowicz	**Dave Tretowicz** (born March 15, 1969) is an **American former professional ice hockey player**. In 1988, he was drafted in the NHL by the Calgary Flames. **He competed in the men's tournament at the 1992 Winter Olympics**.

Table 1: Examples of records in our dataset. **Query** is the generated natural-language query describing a group. **Member** is an entity that belongs to the group described by the query. **Candidate** is the Wikipedia article from which the evidence is taken from. **Evidence** is a passage indicating the member's association with the group.

collection process – if we have missed a meaningful number of evidence during candidate collection. To complete the evaluation of D-MERIT's quality, we also evaluate our automatic identification model, `GPT-4`, to confirm it reliably identifies the vast majority of evidence without adding much false positives.

Evaluation tasks. We turn to Amazon Mechanical Turk (AMT) for sourcing human raters. For the candidate collection evaluation, a human rater is provided with a passage and a prompt containing the query, and is requested to mark whether the passage is evidence or not. In the task designed to gauge the quality of the automatic identification, in addition to the passage and prompt, the annotation of `GPT-4` is also provided. The rater is then requested to judge the correctness of the annotation. Since judging relevance can be subtle[6], we make a decision to judge the correctness of annotations, instead of to annotate and compare results to `GPT-4`. This encourages the rater to consider the annotation's perspective and allows tolerance toward borderline cases. The selection and conditioning process of human raters is detailed in Appendix C.

Exhaustiveness of candidate collection. To ensure our collection process is nearly exhaustive, we need another evidence collection process, independent of ours. We thus adopt the popular TREC approach (Craswell et al., 2020), where a number of systems retrieve the top-k passages given a query, and are then unified to a single set of passages to be judged for relevancy. We use 12 different systems, described in Section 3.1. As for the pool depth, we select $k = 20$ to match our experimental study. Several works researched the relation between pool depth and the completeness of TREC evaluations (Buckley et al., 2007; Keenan et al., 2001; Lu et al., 2016) raising concerns regarding reliability of the shallow pool depth commonly used (the typical TREC setup uses a $k = 10$ depth), hence we also extrapolate the results of this evaluation to a $k = 100$ pool depth.

We select 23 random queries from D-MERIT, and use the TREC approach to retrieve $2,329$ unique passages. Since we are looking for relevant passages that we missed, we discard unique passages that were already annotated by our process (311 such cases, all relevant) and are left with $2,018$ passages. We ask human raters to mark the remaining passages for relevance and find *only* 35 new evidence. In total, the TREC process finds 346 relevant passages, 311 of which were found by our process too. To put this in context, for the same 23 queries, our process finds 990 relevant passages. We note that while our method retrieves many more evidence, it is tailor-made to the Wikidata format, while the method from TREC can be applied to any corpus. To further attest to the exhaustiveness of our approach, we extrapolate the analysis to $k = 100$, and estimate the number of identified evidence to increase to 638, with only 60 new evidence. A more profound discussion of TREC's coverage, including details on the extrapo-

[6]Consider row 2 in Table 1, where the passage does not explicitly say that "Ohio River Islands National Wildlife Refuge" is in "West Virginia". Instead, it says that "Mill Creek Island", which is in "West Virginia", is part of the "Ohio River Islands National Wildlife Refuge".

lation process, can be viewed in Appendix E.

To summarize, the TREC process, with a pool depth of $k = 20$, finds 346 positives and requires 2,329 annotations ($\sim 14.9\%$ positives in the pool). Our method finds 990 positives, requiring 3,206 annotations ($\sim 30\%$ positives in the pool). The TREC process adds only $\sim 3.5\%$ new positives to our method. When TREC is extrapolated to a pool depth of $k = 100$, D-MERIT still has a high (estimated) coverage of 94.5% of identified evidence.

Comparing automatic to manual identification. To verify `GPT-4` is comparable to manual identification, we collect a random sample of 1,300 (query, passage) pairs, consisting of 650 evidence. Out of all the samples, the rater agrees with `GPT-4` 84.7% of the time.[7] Specifically, they disagreed with the model on 141 cases of "relevant" and only 57 cases of "not relevant".

2.5 Natural-language Query Generation

We generate natural sounding queries by providing `GPT-4` the "list of" page title and instructing the model to phrase a natural-language query. For details and examples see Appendix C.

2.6 D-MERIT Overview

The final dataset comprises 1,196 queries, encompassing 60,333 evidence in total. There are 50.44 evidence per query on average, and a median of 22, ranging from a minimum of 5 to a maximum of 682 evidence. On average, each group member contributes about 2 evidence to a query, with 61.8% of the evidence coming from articles other than the members' own articles. The average number of members per query stands at 23.71. We note that it is possible for some members to not contribute any evidence to a query, for example, when the evidence is not in the introduction. In Table 2 we show the members and evidence distributions, and the relation between the number of members and number of evidence mapped to a query.

As accustomed with new datasets, we benchmark D-MERIT on the evidence retrieval task, where all evidence should be retrieved for a given query. Results are reported and discussed in Appendix A.

[7] To further validate this number, we check agreement between two expert annotators. On 400 examples, a 94% agreement is reached. This indicates that the task is less subjective than general relevance tasks which tend to have a lower agreement, explaining the relatively high human-GPT agreement.

# Members	Avg # Evidence	# Queries
1-10	25.5	558
11-20	32.0	282
21-50	69.8	236
51-100	109.7	77
100+	281.2	43

Table 2: Dataset distribution (average number of evidence, number of queries) divided to buckets by number of set members.

3 Experimental Study

With our evaluation set ready, we can address the questions we put forth in the beginning. We experiment to examine the widespread practice of considering only a single evidence per query, and explore whether rankings stabilize as false negatives decrease when adding more labeled evidence.

3.1 Setup

Systems. To ensure our analysis is unbiased towards a specific retrieval paradigm, we utilize the Pyserini information retrieval toolkit (Lin et al., 2021a) to experiment across twelve diverse, out-of-the-box systems: five sparse, four dense, and three hybrid systems. (1) In the sparse category; BM25 (Robertson and Walker, 1994), QLD (Zhai and Lafferty, 2001), UniCoil (Lin and Ma, 2021), SPLADEv2 (Formal et al., 2021) and SPLADE++ (Formal et al., 2022). (2) For the dense methods; DPR (Karpukhin et al., 2020), coCondenser (Gao and Callan, 2022), RetroMAE-distill (Xiao et al., 2022), and TCT-Colbert-V2 (Lin et al., 2021b). (3) In the hybrid category; TCT-Colbert-V2-Hybrid (Lin et al., 2021b), coCondenser-Hybrid, and RetroMAE-Hybrid. Further details regarding the systems can be found in Appendix B.

Evaluation metrics. Needing a metric to quantify the ability of systems to retrieve multiple evidence, we opt to use *recall@k* as this is a simple, common metric for this task. For brevity, we report *recall@20* in the main paper, and show results on *recall@5*, *recall@50*, and *recall@100* in Appendix F. We note that other *k* values show similar trends to *k=20*, and conclusions drawn in this paper generalize to other *k* values reported as well. Other suitable metrics (NDCG, MAP, R-precision) are discussed and reported in Appendix A. After evaluating the performance of each system, we are interested in comparing the recall-based ranking of systems to quantify the gap between the partially- and fully-annotated settings. We utilize Kendall-τ (Kendall, 1938), which can intuitively be understood as a measure of similarity between two ranking orders. This metric evaluates the number of pairwise agreements (concordant pairs) versus disagreements (discordant pairs) in the ranking order of systems between the two settings. A high Kendall-τ score (close to 1) indicates a strong correlation, signifying that the rankings in the partially- and fully-annotated settings are similar, whereas a low score (close to -1) suggests major differences. Specifically, if we have n systems, and C is the number of concordant pairs while D is the number of discordant pairs, then Kendall-τ is given by the formula $\tau = \frac{C-D}{\binom{n}{2}}$, where $\binom{n}{2}$ is the total number of possible pairs. In addition to the vanilla Kendall-τ, we also report the probability of observing a discordant pair, denoted as the *Error-rate*, as it is a more intuitive metric. Formally it is defined as:

$$\textit{Error-rate} = 100 \cdot \frac{D}{\binom{n}{2}} = 100 \cdot \frac{1-\tau}{2}.$$

3.2 Is the single-relevant setup reliable?

To assess the single-relevant setup, we start by randomly sampling an evidence for each query. We evaluate each system on the formed single-relevant evaluation set and compare the resulting system ranking to the ground-truth ranking formed using the fully-annotated dataset. To mitigate the randomness, we run this experiment $1,000$ times, and find that the mean ($\pm$ std) Kendall-τ value is 0.936 ($\pm$0.038), translating to an error-rate of 3.2%. These numbers suggest that sampling a random evidence for each query leads to reliable results. Unfortunately, in order to properly randomly sample an evidence, one would need to annotate a non-feasible amount of passages in most datasets.[8]

In practice, some method is used to select the passages sent for annotation. This method is usually biased[9]. To determine whether selecting an evidence in a biased manner is problematic or not,

[8]For example, in the 2020 TREC challenge (Craswell et al., 2021), operating on the MS-MARCO (Bajaj et al., 2018) dataset, $11,386$ relevant passages were found for 54 queries, an average of 210 per query. In Appendix E we estimate these are only $\sim 50\%$ of the actual relevant passages leading to roughly 500 per query. Given the corpus size, $\sim 8M$ passages, one would need $\sim 16K$ annotations on average to find a single relevant passage randomly for a *single* query.

[9]For example, it has been shown that models tend to suffer from popularity bias (Gupta and MacAvaney, 2022) and that sparse methods tend to prefer longer texts over shorter ones while a human annotator is likely to prefer shorter texts.

Selection	τ-similarity	Error-rate (%)
Random	**0.936**	**3.20**
Most popular	0.696	15.10
Longest	0.545	22.75
Shortest	0.696	15.10
System-based	0.616	19.20

Table 3: Kendall-τ similarities and Error-rate for the different biases in a single-annotation setup.

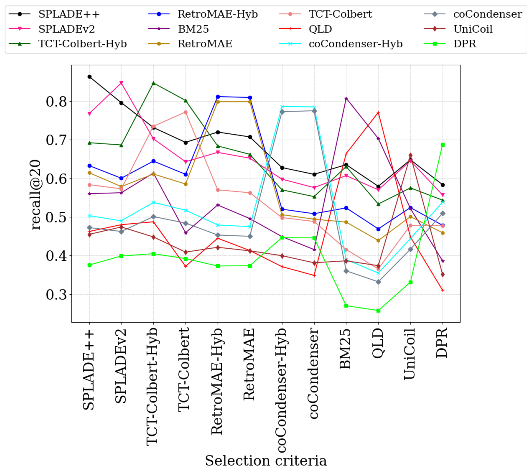

Figure 2: Selection techniques for a single-relevant setting. The x-axis denotes systems used to select passages for annotation. Each tick represents the performance of systems on the same dataset with different annotations. An intersection demonstrates a swap in rankings.

we explore 3 biases: *most popular* selects the most popular[10] evidence for each query. We also consider a length-selection approach, which considers the number of words in a given passage, by selecting the *longest* and *shortest* evidence available for each query. Results are presented in Table 3. It can be seen that as opposed to random selection, in the more likely scenario of a biased selection the error-rate is much higher, suggesting that the single-relevant setting is unreliable. A popular technique for sampling passages for annotation is using an existing retrieval system, and annotating passages in the order they are retrieved until a relevant passage is found. We simulate this by considering each of our 12 considered retrievers as the base system. We then evaluate all of the systems on the 12 formed evaluation sets. Results are plotted in Fig. 2. The graph shows that the selection technique, used to pick which passages are annotated, has a major effect on the systems' measured performance *and*

[10]We define popularity as the number of times an article is referenced, which can be derived using the "What Links Here" feature from Section 2.3.2.

on the ranking of the different systems. For example, when choosing evidence using BM-25, QLD is ranked as the best system (excluding BM-25 itself), while when choosing evidence using either coCondenser, coCondenser-Hybrid, DPR or TCT-Colbert, QLD is the worst performing system. For other systems selecting evidence, it is ranked somewhere in between. When comparing the 12 rankings formed using these evaluation sets to the ranking formed by the completely annotated dataset, the average Kendall-τ score computed is 0.616, translating to an average error-rate of 19.2%.[11] Table 3 indicates that system-based selection is indeed closer to biased selection than it is to random selection. In summary, the experiments presented in this section show that while random selection of evidence can lead to reliable results in the single-relevant scenario, the more realistic case (where the annotated evidence is not randomly selected) is prone to generating misleading results and ranking of systems.

3.3 Is the single-relevant scenario enough when systems are significantly separated?

After establishing that there are cases where the single-relevant scenario is not reliable, we ask in what cases it can be sufficient. To explore this, we first define buckets of pairs of systems as follows. A pair of systems (A, B) is in a $[p_{min}, p_{max})$ bucket if A is better performing than B, and the statistical significance computation for the difference between these two systems leads to a p-value of at least p_{min} and at most p_{max}, using a relative t-test, as computed on the fully annotated evaluation set. We then repeat the final experiment described in Section 3.2, but when calculating Kendall-τ and it's error-rate we only consider pairs of systems that fall in some bucket. We denote this measure as partial-Kendall-τ.[12] We consider 3 buckets: $[0, 0.01)$ represents systems with very low p-values, meaning they are very far apart in performance, hence should be easier to order correctly. $[0.01, 0.05)$ represents systems with a significant, yet not extreme difference. The final bucket, $[0.05, 1)$, contains pairs of systems that do not differentiate in a statistically significant way. Results are shown in Table 4. We observe that, as expected, the error-rate drops when a bucket represents a smaller p-value, indicating higher significance that the systems are ordered correctly.

p_{min}	p_{max}	partial-τ	Error-rate (%)
0.0	0.01	**0.658**	**17.1**
0.01	0.05	0.333	33.3
0.05	1.0	0.0	50.0

Table 4: Partial-Kendall-τ similarity (defined in Section 3.3, denoted partial-τ) and Error-rate computed on pairs of systems that belong to the $[p_{min}, p_{max})$ bucket.

3.4 Do rankings stabilize as false negatives decrease?

Taking the evidence chosen using the different systems as discussed in Section 3.2, we gradually add a fraction of annotated evidence for all queries in the evaluation set. We then evaluate the systems on each partially annotated dataset by comparing the ranking achieved to the fully annotated evaluation set. We divide pairs of systems into buckets based on their p-values, as described in Section 3.3, and for each percentile we average results across the different system pairs falling within each bucket. Results are presented in Fig. 3. Depending on the significance of the difference between systems, results show a different portion of evidence needs to be annotated in order to achieve the correct order. For example, if we are aiming at a ~ 0.8 Kendall-τ score, representing a $\sim 10\%$ error-rate, for very significant pairs of systems acquiring $\sim 20\%$ of the positives should suffice, while for systems with a non-significant difference between them, almost all positives are needed.

4 Related Work

Our work builds on previous efforts in benchmark creations in multi-answer and multi-evidence settings and the complete annotation setting. Below, we detail how our work relates to both.

Multi-answer retrieval. QAMParI (Amouyal et al., 2023) introduce a benchmark of questions with multiple answers extracted from lists in Wikipedia, and Quest (Malaviya et al., 2023) is a dataset with queries containing implicit set operations based on Wikipedia category names. Both limit evidence collection to the Wikipedia article

[11] We eliminate the system used to select the evidence from the computation, as it generates artificial swaps. For example when computing the Kendall-τ for the ranking formed by choosing the first evidence as ranked by BM-25, Kendall-τ is computed on the ranking of all except BM-25.

[12] We opt to use Kendall-τ due to its simplicity, yet it does not accurately capture all the intricacies of ranking system performance. More details on this and an involved metric, taking into account the significance of differences between systems, is presented in Appendix D. Results using this metric validate our choice of Kendall-τ.

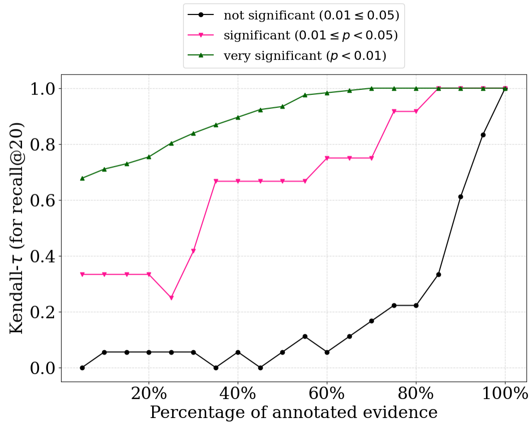

Figure 3: Partial-Kendall-τ between rankings of systems with k percent annotations and ranking with all evidence, using *recall@20*. System pairs are divided into 3 buckets as described in Section 3.3.

of the answer. In contrast, our goal is to identify all relevant evidence for each answer, including other Wikipedia articles. RomQA (Zhong et al., 2022) curates a large multi-evidence and multi-answer benchmark derived from the Wikidata knowledge graph with the goal of challenging the retriever and QA model. Although RomQA provides a large number of evidence, they do not aim for complete annotation nor to understand the negative effect of evaluation with partial annotations. Our paths diverge in that they seek to evaluate QA models and we aim to understand the effects of partial annotations on retriever evaluation, and to collect *all* evidence for each answer.

Exhaustive annotation. TREC Deep Learning (Craswell et al., 2020, 2021, 2022, 2023, 2024) is a yearly effort to completely-annotate queries for passage retrieval from the MS-Marco benchmark (Bajaj et al., 2018). Since annotating the entirety of MS-MARCO is unrealistic (~1M queries and ~8.8M passages), they conduct a competition where participants submit the results of their retrievers. Then, the results are pooled and their relevancy is evaluated. However, manual evaluation is a non-scalable approach, and over a span of five years (2019–2023) only 312 queries were annotated. In addition, exhaustiveness is unlikely as previously observed in (Zobel, 1998) and further corroborated in Appendix E. NERetrieve (Katz et al., 2023) shares our aspiration for a completely-annotated dataset. It proposes a retrieval-based NER task that creates a Wikipedia-based dataset where entity types function as queries and relevant passages contain a span that mentions instances of the entities (e.g., "Dinosaurs" is an entity type and "Velociraptor" is an instance of it). With some similarity to our process, they collect candidates by relaxed matching of mentions of entities in documents that reference them (on DBPedia's link-graph (Lehmann et al., 2015)), and then use a classifier to filter out cases that do not match their query. However, our work annotates evidence and not simply mentions of entities in a passage. Moreover, in addition to creating an exhaustively annotated dataset, we study the effects of partial annotation.

5 Conclusions

In this work we question whether the lack of rigorous annotation in modern retrieval datasets results in false conclusions. To answer this, we create D-MERIT, a dataset aspiring to collect *all* relevant passages in the corpus for each query. We use it to explore the impact of evaluating systems on datasets riddled with false negatives; We demonstrate that evaluation based on queries with a single annotated relevant passage is highly dependent on the passages selected for annotation, unless one system is significantly superior to all others. We also show that the number of annotations required to stabilize the rankings is a factor of the difference in performance between systems. We conclude that there is a clear efficiency-reliability curve when it comes to the amount of annotations invested in a retrieval evaluation set, and that when picking the correct spot on this curve considerations should include the estimated difference between the systems in question and the method used to choose the passages sent to annotation. We show that the commonly used TREC-style evaluation method fails to find a significant portion of the relevant passages in D-MERIT, suggesting that using this annotation approach on D-MERIT would lead to a non-negligible error rate. If it's possible, our recommendation for other datasets would be to estimate the coverage of the TREC method before using it for evaluation. Otherwise, its results should be taken with a grain-of-salt. Finally, our dataset opens a new avenue for research, both as a test-bed for evaluation studies, as well as evaluation in a high-recall setting.

Limitations

Exhaustiveness. Our evidence identification process is automated by `GPT-4`, the current state-of-the-art for text analysis. Despite achieving high agreement with human annotators, it is not perfect. Furthermore, even with a flawless model, computing the relevance of *all* passages in Wikipedia for each member in each query would have resulted in millions of inferences, which would have made the creation of this dataset unfathomably expensive. We thus make the (sensible) assumption that a passage with evidence must contain a link to the article of the entity. It is possible some evidence were never collected, as analyzed in Section 2.4.

Generalization of conclusions. We (and many before us) believe that in order to properly evaluate retrieval systems, the community should *strive* to collect all (or most) relevant passages. We believe this is true for many different datasets and scenarios. Having said that, showing this explicitly requires to completely annotate datasets, which is hard and expensive. Therefore, while we do believe that most of our conclusions can generalize to many other datasets, technically we could show them only on the dataset we used.

Data evaluation compatibility. Our dataset is made of set-queries with multiple members (translating to multiple answers in the QA setting). In such cases, systems are usually evaluated using datasets containing a single relevant *per answer*. In Section 3.2 we evaluate and draw conclusions using a single positive *per query*. We do so in order to draw conclusions regarding cases where single positives per query are used, but in practice these datasets usually contain *single-answer* queries (e.g. MS-MARCO). While we do believe our conclusions generalize to this case, it would have been more accurate to use such a single-answer-per-query dataset. Unfortunately, collecting such a fully annotated dataset is not trivial.

Ethics Statement

Automatic annotation. Since our annotation is automatic, it is model-dependent. This means it is vulnerable to the model's biases. As a result, it may fail to attribute evidence to a query if a candidate is under-represented in the model's training data. This might cause D-MERIT to miss out on evidence that belongs to some under-represented group.

Rater details. To collect annotations on our dataset, we used Amazon Mechanical Turk (AMT). All raters had the following qualifications: (1) over 5,000 completed HITs; (2) 99% approval rate or higher; (3) Native English speakers from England, New Zealand, Canada, Australia, or United States. Raters were paid \$0.07 per HIT, and on average, \$20 an hour. In addition, raters that performed the task well were given bonuses that reached double pay.

Annotation collection and usage policy. Raters were notified that their annotations are intended for research use in the field of Natural Language Processing and Information Retrieval, and will ultimately be shared publicly. The task and collected annotations were objective and excluded personal information. Moreover, all data sources for the study were publicly accessible.

Computing resources. We used only modest computing resources. For both, the dataset creation and the experimentation, we used a single Amazon-EC2-g5.4xlarge instance for 200 hours, which costs \$1.6 per hour. For the annotation of the passages, and creation of the natural-language queries, we utilized `GPT-4-1106-preview`, which at the time of writing, is priced at \$0.01 for 1K input tokens, and \$0.03 for 1K output tokens. In total, we paid ~\$3,000 for our use of the model.

References

Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. QAMPARI: A benchmark for open-domain questions with many answers. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–110, Singapore. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Comput. Surv.*, 55(1).

C Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2007. Bias and the limits of pooling for large collections.

Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 25–32, New York, NY, USA. Association for Computing Machinery.

Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3417–3419, New York, NY, USA. Association for Computing Machinery.

James P. Callan. 1994. Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 302–310, Berlin, Heidelberg. Springer-Verlag.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track. In *Text REtrieval Conference (TREC)*. TREC.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2022. Overview of the trec 2021 deep learning track. In *Text REtrieval Conference (TREC)*. NIST, TREC.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2023. Overview of the trec 2022 deep learning track. In *Text REtrieval Conference (TREC)*. NIST, TREC.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the trec 2019 deep learning track.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Hossein A. Rahmani, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2024. Overview of the trec 2023 deep learning track. In *Text REtrieval Conference (TREC)*. NIST, TREC.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

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*, SIGIR '22, page 2353–2359, New York, NY, USA. Association for Computing Machinery.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Prashansa Gupta and Sean MacAvaney. 2022. On survivorship bias in ms marco. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22. ACM.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick 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)*, pages 6769–6781, Online. Association for Computational Linguistics.

Marcin Kaszkiel and Justin Zobel. 1997. Passage retrieval revisited. In *ACM SIGIR Forum*, volume 31, pages 178–185. ACM New York, NY, USA.

Uri Katz, Matan Vetzler, Amir Cohen, and Yoav Goldberg. 2023. NERetrieve: Dataset for next generation named entity recognition and retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3340–3354, Singapore. Association for Computational Linguistics.

Sabrina Keenan, Alan F. Smeaton, and Gary Keogh. 2001. The effect of pool depth on system evaluation in trec. *J. Am. Soc. Inf. Sci. Technol.*, 52(7):570–574.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, S. Auer, and Christian Bizer. 2015. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6:167–195.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

Jimmy Lin and Xueguang Ma. 2021. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. *arXiv preprint arXiv:2106.14807*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021a. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2356–2362, New York, NY, USA. Association for Computing Machinery.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021b. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173, Online. Association for Computational Linguistics.

Xiaolu Lu, Alistair Moffat, and J. Shane Culpepper. 2016. The effect of pooling and evaluation depth on ir metrics. *Inf. Retr.*, 19(4):416–445.

Sean MacAvaney and Luca Soldaini. 2023. One-shot labeling for automatic relevance estimation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23. ACM.

Chaitanya Malaviya, Peter Shaw, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2023. Quest: A retrieval dataset of entity-seeking queries with implicit set operations.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2022. A survey on multi-hop question answering and generation. *arXiv preprint arXiv:2204.09140*.

Taichi Murayama. 2021. Dataset of fake news detection and fact verification: a survey. *arXiv preprint arXiv:2111.03299*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 232–241, Berlin, Heidelberg. Springer-Verlag.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10).

Alan Stuart. 1953. The estimation and comparison of strengths of association in contingency tables. *Biometrika*, 40(1/2):105–110.

Manju Vallayil, Parma Nand, Wei Qi Yan, and Héctor Allende-Cid. 2023. Explainability of automated fact verification systems: A comprehensive review. *Applied Sciences*, 13(23):12608.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emine Yilmaz and Javed A. Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, page 102–111, New York, NY, USA. Association for Computing Machinery.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 334–342, New York, NY, USA. Association for Computing Machinery.

Victor Zhong, Weijia Shi, Wen tau Yih, and Luke Zettlemoyer. 2022. Romqa: A benchmark for robust, multi-evidence, multi-answer question answering.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments? In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Justin Zobel, Alistair Moffat, Ross Wilkinson, and Ron Sacks-Davis. 1995. Efficient retrieval of partial documents. *Information Processing & Management*, 31(3):361–377. The Second Text Retrieval Conference (TREC-2).

A Benchmarking D-MERIT

While tangential to this paper, the D-MERIT dataset allows us to benchmark the ability of existing retrieval models to perform on the full-recall retrieval setup, as it's coverage is very high as reported in Section 2.4. This section describes this benchmark process.

Benchmark metrics. We select Recall, Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (MAP). In addition, given that we possess complete evidence for every query, we can calculate R-precision– a form of recall where k varies for each query, determined by the specific total evidence count to that query. For instance, if a query corresponds to 40 pieces of evidence, then k is set at 40. Achieving a perfect score means that the top 40 results are all evidence associated with the query.

Results. Performance of all systems is shown in Table 5, with SPLADE++ and SPLADEv2 performing best across all metrics. The scores suggest there is substantial room for improvement on our evidence retrieval task. For example, the *recall@100* score indicates no system successfully retrieves even half of the evidence on average.

B Further Details: Experimental Study

To allow reproduction of our results, we detail the hyper-parameters used in our work. We utilize the Pyserini information retrieval toolkit (Lin et al., 2021a) with the following settings for each system: **BM25** is employed using the standard Lucene index for indexing and retrieving results. Similarly, **QLD** is used but with the QLD reweighing option to refine the process. **UniCoil** embeddings are generated with the *castorini/unicoil-noexp-msmarco-passage* encoder, and retrieval is conducted using Lucene search with the 'impact' option to incorporate unicoil weights. **SPLADEv2** and **SPLADE++** follow a similar approach, where passages and queries are embedded using their respective official code repositories, and retrieval is performed using Lucene with the 'impact' option. **DPR** involves embedding passages and queries with the *facebook/dpr-ctx_encoder-multiset-base* and *facebook/dpr-question_encoder-multiset-base* encoders, respectively, with retrieval via FAISS (Douze et al., 2024). **RetroMAE-distill** adopts a similar strategy, utilizing the *Shitao/RetroMAE_MSMARCO_distill* encoder for both queries and passages. **TCT-Colbert-V2** also mirrors this approach but uses the *castorini/tct_colbert-v2-msmarco* encoder. **coCondenser** involves training document and query encoders on the Natural Questions dataset (Kwiatkowski et al., 2019) using the CoCondenser official code repository. Hybrid models such as **TCT-Colbert-V2-Hybrid**, **coCondenser-Hybrid**, and **RetroMAE-Hybrid** combine the strengths of BM25 with **TCT-Colbert-V2**, **coCondenser**, and **RetroMAE-distill** respectively, using a fusion score with $\alpha = 0.1$.

C Further Details: D-MERIT Creation

License. D-MERIT builds on data from Wikipedia, which carries a Creative Commons Attribution-ShareAlike 4.0 International License. This license requires that any derivative works also carry the same license.

Conditioning human raters. Before the evaluation process begins, we need to assure the raters we use understand the task and can perform it adequately. We thus begin a conditioning process. First, we run a qualification exam, and the raters that get all the questions right, are invited to an iterative training process. The process includes small batches, of up to 100 (passage, prompt) pairs, where the rater submits their response and we provide personal feedback. Moreover, all tasks included an option to mark the example as difficult or provide textual feedback about it, to encourage communication from the raters as they work. After each batch raters are filtered out, until we remain with a single rater with a success rate of over 95% on a single batch. The task is visualized in Fig. 10.

Automatic identification details. To automatically identify evidence, `GPT-4` is provided with a passage and a structured query. In this context, a structured query begins with the article name, followed by its section names arranged hierarchically (separated by "»"), corresponding to the structure of the article, and ultimately culminating in the column value. For instance, a typical structured query could be "Cities and Towns in Cambodia" (article name) » "Cities" (section name) » "Name" (column name). The task for `GPT-4` is to determine whether the passage provides evidence supporting the query. The evaluation involves analyzing the text to ascertain whether the passage directly or indirectly confirms the entity in question is part of

System	Recall@k				NDCG@k				MAP@k				R-precision
	5	20	50	100	5	20	50	100	5	20	50	100	
SPLADE++	**9.43**	**24.11**	**36.02**	**45.16**	**38.17**	**36.54**	**38.05**	**40.56**	**7.11**	**15.0**	**19.35**	**21.72**	**28.16**
SPLADEv2	7.82	21.21	33.29	43.34	32.09	31.43	33.78	37.00	5.74	12.20	16.03	18.27	24.82
TCT-Colbert-Hybrid	7.85	19.62	29.71	37.97	34.86	31.60	32.23	34.33	5.80	11.48	14.78	16.56	22.75
bm25	6.65	17.46	27.54	35.76	28.93	27.20	28.62	31.13	4.76	9.76	12.83	14.61	20.86
RetroMAE-Hybrid	7.30	17.48	25.95	32.85	33.95	29.21	29.19	30.82	5.71	10.63	13.14	14.48	20.12
RetroMAE	7.03	16.62	24.78	31.61	32.71	27.98	27.94	29.61	5.47	10.05	12.38	13.66	19.29
TCT-Colbert	6.27	15.44	23.59	30.95	29.31	25.73	26.08	27.95	4.58	8.64	11.02	12.39	18.02
CoCondenser-Hybrid	5.28	14.81	24.25	32.88	22.13	21.87	23.96	26.89	3.41	6.82	9.10	10.63	16.78
QLD	5.49	13.96	23.56	31.96	24.54	21.71	23.63	26.55	3.77	7.07	9.51	11.13	16.56
CoCondenser	4.87	13.75	23.02	31.52	20.71	20.42	22.64	25.54	3.14	6.20	8.35	9.77	15.69
Unicoil	4.47	10.95	17.27	23.28	20.86	17.96	18.70	20.49	3.25	6.05	7.72	8.83	13.19
DPR	3.90	9.62	15.99	21.72	18.51	15.90	16.64	18.41	2.63	4.48	5.67	6.37	10.89

Table 5: Performance of a variety of baselines on D-MERIT. Recall, NDCG, and MAP are evaluated over four k values: 5, 20, 50, and 100. The k value in R-precision is the total number of evidence of a query, which changes from query to query.

the group defined by the query. For example, in a query aimed at identifying names of Cambodian cities, the passage must either explicitly state or strongly suggest that a particular city belongs in Cambodia to be considered relevant. Our prompts follow our definition of relevance from Section 2.2:

```
If you were writing a report on member being part of article-name, and would like to gather *all* the documents that directly confirm member is part of article-name, in the category hierarchy article-name » section-name » column-name, will you add the following document to the collection? Answer with “yes” or “no”.
```

Natural-language query generation prompt. To translate a structured query to its natural-language variant, we prompt GPT-4 using the template below. Examples of input and output can be viewed in Table 6.

```
Please pretend you are a typical Google Search user, show me what you would write in the search bar. For example: cultural property of national significance in Switzerland:Zurich » Richterswil » Name, where » indicates a hierarchy, a typical search would be: names of cultural properties of national significance in Richterswil, Zurich, Switzerland.

Here, try this one: {input}
```

D Concordance

Kendall-τ (Kendall, 1938) is a popular metric for evaluating rank correlation between rankings. This is done by comparing the number of concordant and dis-concordant elements between two ranks over a set of elements. More general variants of Kendall-τ (Kendall, 1945; Stuart, 1953) address cases where ties exist (i.e., in one ranking two elements received an identical score).

Structured Query	Natural-language Query
List of Zhejiang University alumni » Politics & government » Name	names of Zhejiang University alumni in politics and government
List of Wisconsin state forests » Forest name	names of Wisconsin state forests
List of World War I flying aces from the United States » Served with the Aéronautique Militaire » Name	names of US World War I flying aces who served with the Aéronautique Militaire
List of LGBT classical composers » 20th century » Name	names of 20th century LGBT classical composers
List of Eliteserien players » Name	names of Eliteserien football players
List of National Monuments in County Sligo » National Monuments » Monument name	names of National Monuments in County Sligo, Ireland

Table 6: Examples of structured queries and their corresponding natural-language form.

The simplicity of Kendall-τ makes it tempting to utilize it to compare the ranking of retrieval systems. However, it fails to capture some of the intricacies of this comparison due to several reasons. First, simply comparing system scores is insufficient, as an additional verification using a significance test is necessary. Ties can be defined (i.e., system A is tied with system B if $p > 0.05$), but the relation is not transitive (A tied with B and B tied with C does not imply that A is tied with C), as required by variants of Kendall-τ that support ties. Second, some ranking errors are more troublesome than others. Finding that a new system is “tied” with the baseline system when in fact it is worse might be undesirable. However, incorrectly reporting that it is better is improper.

Even though Kendall-τ suffers from the shortcomings above, we hypothesize that it is still a good metric for comparing performance rankings. To validate this we propose a new metric, *concordance*,

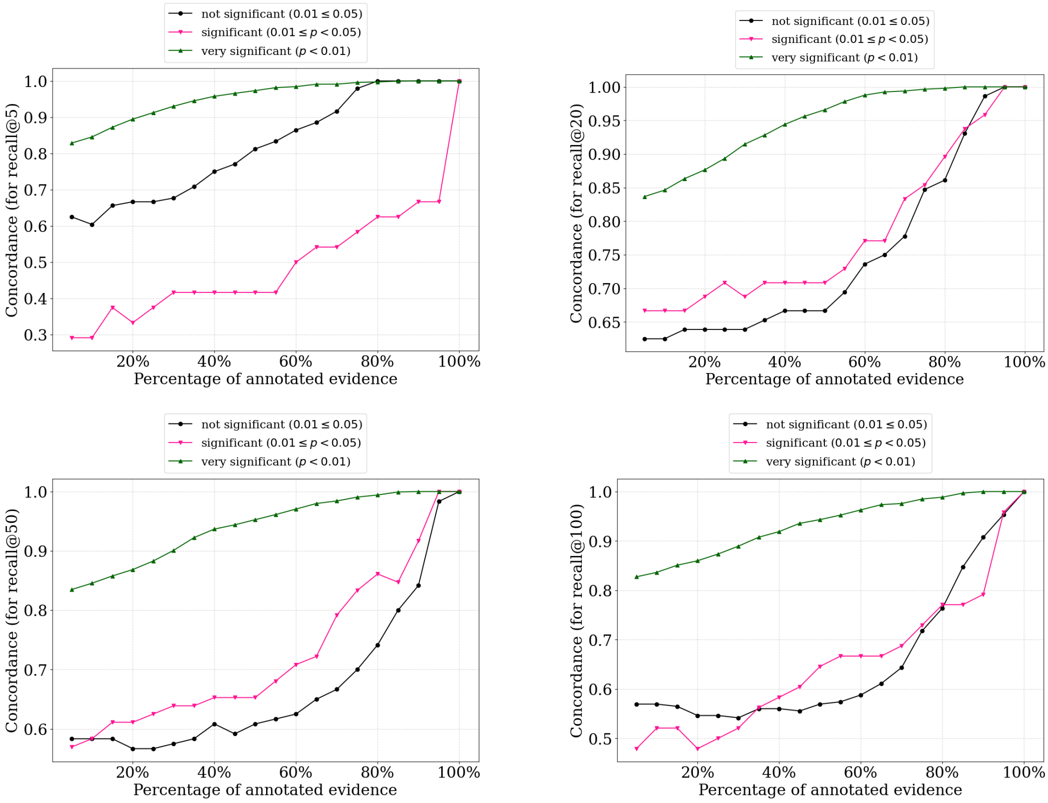

Figure 4: Concordance between rankings of systems with varying percentages of evidence and ranking with all evidence, using *recall@5*, *recall@20*, *recall@50*, and *recall@100*. System pairs are divided into 3 buckets as described in Section 3.3.

that addresses these shortcomings of Kendall-τ and its variants. This is done by considering the relations $A > B$ and $A < B$ for a pair of systems A and B. This way if in the ground truth A is significantly better than B and in the compared ranking A is tied with B, the two rankings will agree on the relation $A < B$ (will be false in both) and disagree on the relation $A > B$. In a more troublesome error, where $A < B$ in the compared ranking, the two rankings will disagree on both relations. Formally, let π_1 and π_2 be two rankings of a set of retrieval systems S. For each pair of systems s_1, s_2 and ranking π we define

$$\pi(s_1, s_2) = \begin{cases} 1, & s_1 \text{ is significantly better than } s_2 \\ 0, & \text{otherwise.} \end{cases}$$

Then concordance is defined as the agreement between the rate of agreement over all ordered pairs of systems between two rankings:

$$\text{conc}(\pi_1, \pi_2) = \frac{1}{P(|S|, 2)} \sum_{s_1} \sum_{s_2 \neq s_1} \pi_1(s_1, s_2) \odot \pi_2(s_1, s_2),$$

where $P(n, r)$ is the number of permutations of size r from a set of size n, and $\odot$ is the XNOR operator (equals to 1 if both inputs equal).

Using concordance, we validate the results found in Section 3.3 and Section 3.4 using Kendall-τ. This is done by repeating the experiment and calculating the mean concordance of system rankings given evidence found by different systems with the ground truth ranking (in which all evidence are annotated). We run this experiment for a single annotated evidence and different percentiles of annotated evidence.

In Table 7 and Figure 4 we see that pairs of systems with a very significant difference between them (i.e., $p < 0.01$) are evaluated with higher accuracy than systems falling in the other two buckets. This validates the results found in Section 3.3 and Section 3.4 and shows that Kendall-τ is a good proxy for evaluating the rankings of IR systems.

E TREC Coverage

TREC (Craswell et al., 2020, 2021, 2022, 2023, 2024), a popular retrieval competition, also tries to deal with the problem of partial annotated retrieval

k	p_{min}	p_{max}	Concordance
5	0.0	0.01	0.809
5	0.01	0.05	0.292
5	0.05	1.0	0.646
20	0.0	0.01	0.823
20	0.01	0.05	0.708
20	0.05	1.0	0.611
50	0.0	0.01	0.821
50	0.01	0.05	0.556
50	0.05	1.0	0.592
100	0.0	0.01	0.813
100	0.01	0.05	0.500
100	0.05	1	0.583

Table 7: Concordance computed only on pairs of systems that fall within the $[p_{min}, p_{max})$ bucket. k is the *recall@k* used.

datasets. In this section we compare our approach for collecting multiple evidence for queries with their approach. This is done by applying TREC's approach to our dataset and testing its coverage. This will reveal, even though anecdotally, the ability of TREC's approach to find numerous evidence. The approach in TREC does not utilize a structured data source for the creation of the judgement set. Instead, they create a pool of candidates from the set of passages retrieved by a large set of systems. Specifically, TREC runs a competition and publishes a query set and a corpus. Any participant team executes their system and submits a retrieved list. Then, TREC pools top-k passages from each participant and sends them for human annotation, annotating for relevancy. Before applying the approach used by TREC to our dataset we first formally define this process. Let Q be the set of queries and E_q the evidence set of query $q \in Q$. In addition, let S be the set of systems and $E_{q,s}$ be the evidence set found in the top-10 passages retrieved by system $s \in S$ for query $q \in Q$. Then, the judgement set of query q is defined as $J_q(S) = \cup_{s \in S} E_{q,s}$. We denote the coverage of S on Q as:

$$C_Q(S) = \frac{1}{|Q|} \sum_{q \in Q} \frac{|J_q(S)|}{|E_q|}.$$

When fixing the number of passages retrieved by each system to $k = 10$, as done in TREC, and given the 12 systems considered in this paper (see Section 3.1), we can compute their coverage on D-MERIT which is equal to 31.7%. While this may be low, we only consider a small number of systems, as it is typical to use around 100 systems. Also, increasing k is expected to increase the coverage. Following, we use extrapolation techniques to estimate the affect of both.

E.1 Extrapolating Number of Systems

Due to time and compute constraints using 100 systems, as typically done in the TREC competition, is unrealistic. This leads us to approximate the coverage instead. In order to approximate the coverage of a larger number of systems we first fix $k = 10$, and compute the expected coverage of a random subset of systems of size t uniformly sampled from S. That is,

$$C^*_Q(S,t) = \mathop{\mathbb{E}}_{S' \sim U(S),\ |S'|=t} [C_Q(S')].$$

Given the values of $C^*_Q(S,t)$ for $t = 1, \ldots, 12$, we fit a logarithmic curve (as coverage is both concave and monotonically-increasing) to these observations and observe a root mean-squared-error (RMSE) of 0.16% and a maximum error of 0.31%. Finally we extrapolate to predict the coverage for $t = 13, \ldots, 100$. The results of the experiment is presented in Fig. 5. As can be seen, we predict that broadening the judgement sets by retrieving with as many as 100 systems only increases the coverage from 31.7% to 47.1%. This result further corroborates the finding by (Zobel, 1998), which states that the pooling approach used in TREC finds, at best, 50-70% of the evidence. We conclude that our approach is able to achieve a much higher coverage. This is expected to improve the correctness of our evaluation. Note that our approach depends on structured data in Wikipedia. On the other hand, the approach utilized in TREC is universal as it can be applied to any corpus and query.

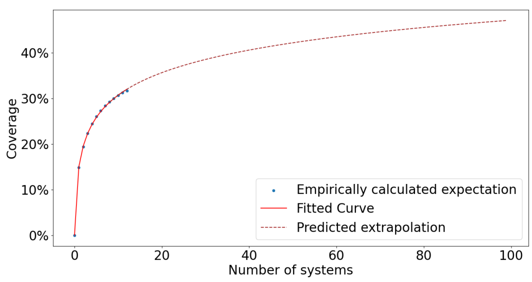

Figure 5: Fraction of relevant passages covered by top-10 passages for s systems.

E.2 Extrapolating Number of Retrieved Documents per System

Increasing the pool size can uncover additional positive results, but will result in a significantly larger annotation pool size. We adopt a similar method to extrapolating the coverage by increasing the number of systems, and but focus instead on the size of the pool.

We use the coverage evaluation dataset described in section 2.4 which takes a the top-20 pool from 12 systems and uses human annotators to label the relevancy of each entry in the pool. Next, we assign each relevant entry in the pool its minimum rank from all systems and construct pools for each depth size. For example, for k=10, we take all documents that were ranked at the top-10 by at least a single system.

Finally, we extrapolate to predict for the number of newly identified evidence (Figure 6) and the overall documents found by the pooling approach (Figure 7) for $t = 21, \ldots, 100$. The results show that even for a pool-depth of $k = 100$, we estimate that only 60 new evidences will be identified. This means that the coverage of our method is estimated to be $\sim 94.5\%$ out of all identified evidence. In addition, we see that the pooling approach for $k = 100$ is estimated to retrieve 638 evidence (578 already found by our method) covering only 60.8% with a significant increase of annotation overhead.

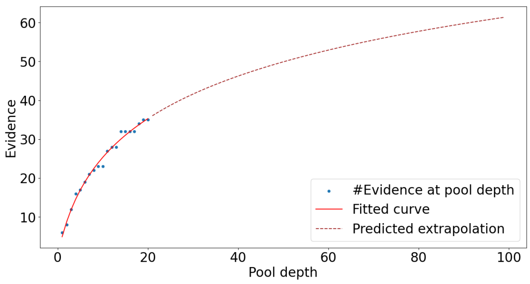

Figure 6: Number of newly identified evidence by pool depth k.

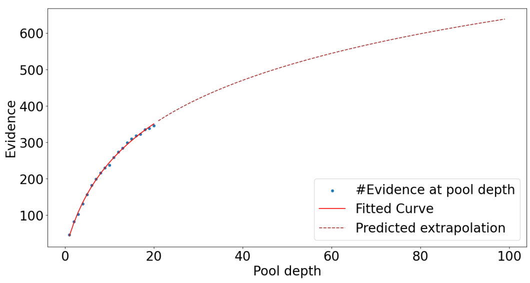

Figure 7: Number of identified evidence by pool depth.

F Extended Results

In the main paper we focused on *recall@20* for brevity when reporting results. Here, we report experiments shown in Section 3 measuring also *recall@5/50/100*. Conclusions pointed out in the main paper hold for all values of k.

k	p_{min}	p_{max}	**partial-τ**	**Error-rate (%)**
5	0.0	0.01	0.654	17.30
5	0.01	0.05	-0.583	79.15
5	0.05	1.0	-0.125	56.25
20	0.0	0.01	0.658	17.10
20	0.01	0.05	0.333	33.35
20	0.05	1.0	0.000	50.00
50	0.0	0.01	0.658	17.10
50	0.01	0.05	0.167	41.65
50	0.05	1.0	0.200	40.00
100	0.0	0.01	0.642	17.90
100	0.01	0.05	-0.083	54.15
100	0.05	1	0.185	40.75

Table 8: partial-Kendall-τ similarity (as defined in Section 3.3, denoted here as partial-τ) and Error-rate computed only on pairs of systems that fall within the [p_{min}, p_{max}) bucket. k is the *recall@k* used.

k	**Selection**	**τ-similarity**	**Error-rate (%)**
5	Random	0.815	9.25
5	Most popular	0.727	13.65
5	Longest	0.462	26.90
5	Shortest	0.585	20.75
5	System-based	0.587	80.65
20	Random	0.936	3.20
20	Most popular	0.697	15.15
20	Longest	0.545	22.75
20	Shortest	0.697	15.15
20	System-based	0.616	19.20
50	Random	0.916	4.20
50	Most popular	0.687	15.65
50	Longest	0.606	19.70
50	Shortest	0.576	21.20
50	System-based	0.596	20.20
100	Random	0.894	5.30
100	Most popular	0.818	9.10
100	Longest	0.697	15.15
100	Shortest	0.545	22.75
100	System-based	0.523	23.85

Table 9: Kendall-τ similarities and error for different biases, in a single-annotation setup. k is the *recall@k*.

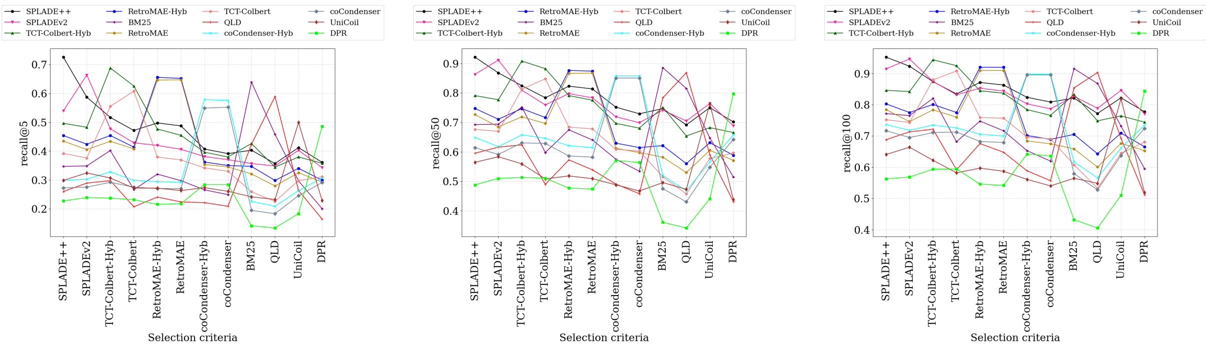

Figure 8: Single-annotation per query datasets with varying selection methods. Left to right: *recall@5/50/100*.

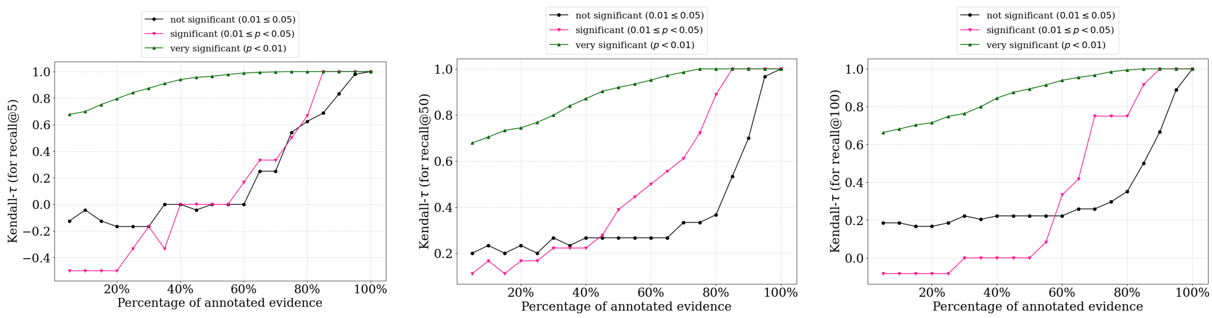

Figure 9: Kendall-τ between rankings of systems with varying percentages of evidence and ranking with all evidence, using *recall@5/50/100*. System pairs are divided into 3 buckets as described in Section 3.3.

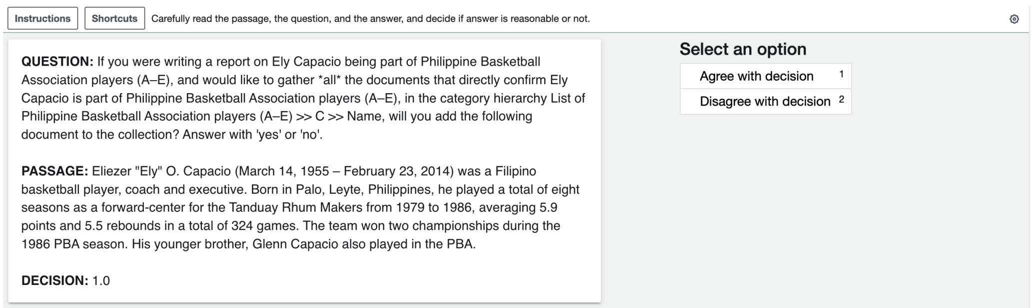

Figure 10: The human evaluation task detailed in Section 2.4.

List of Zhejiang University alumni

1 language

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

This is a list of notable graduates as well as non-graduate former students, academic staff, and university officials of Zhejiang University and its predecessors in China. It also includes those who may be considered alumni by extension, having studied at institutions that later merged with Zhejiang University.

This article contains dynamic lists that may never be able to satisfy particular standards for completeness. You can help by adding missing items with reliable sources.

Politics & government [edit]

Name	Known as	Known for	Links to Zhejiang University
Chen Duxiu	political leader, writer	• Published *La Jeunesse* which led to the New Culture Movement • Founded the Chinese Communist Party and served as its leader	Studied shipbuilding and French at Qiushi Academy during 1897-1899 until he was expelled due to anti-government speech.
Jiang Baili	military writer, strategist, trainer	• Served as the acting principal of the Whampoa Military Academy • Wrote *Treatise on National Defence*	Studied at Qiushi Academy during 1899-1901
Chen Yi	military leader, politician	• Led the 19th Route Army to fight against Japan in January 28 incident • Served as the Governor of Taiwan Province from 1947-to 1949, during which the February 28 incident occurred	Studied at Qiushi Academy
Huang Fu	politician	• Led and participated in the 1924 Beijing Coup • Served as the Acting President of the Republic of China	Studied at Qiushi Academy

Figure 11: A screenshot of the Wikipedia article corresponding to the first query in Table 6.