The Viability of Crowdsourcing for RAG Evaluation

Lukas Gienapp
Leipzig University and ScaDS.AI
Leipzig, Germany

Tim Hagen
University of Kassel
and hessian.AI
Kassel, Germany

Maik Fröbe
Friedrich-Schiller-
Universität Jena
Jena, Germany

Matthias Hagen
Friedrich-Schiller-
Universität Jena
Jena, Germany

Benno Stein
Bauhaus-Universität Weimar
Weimar, Germany

Martin Potthast
University of Kassel,
hessian.AI, and ScaDS.AI
Kassel, Germany

Harrisen Scells
University of Kassel
and hessian.AI
Kassel, Germany

Abstract

How good are humans at writing and judging responses in retrieval-augmented generation (RAG) scenarios? To answer this question, we investigate the efficacy of crowdsourcing for RAG through two complementary studies: response writing and response utility judgment. We present the Crowd RAG Corpus 2025 (CrowdRAG-25), which consists of 903 human-written and 903 LLM-generated responses for the 301 topics of the TREC RAG'24 track, across the three discourse styles 'bulleted list', 'essay', and 'news'. For a selection of 65 topics, the corpus further contains 47,320 pairwise human judgments and 10,556 pairwise LLM judgments across seven utility dimensions (e.g., coverage and coherence). Our analyses give insights into human writing behavior for RAG and the viability of crowdsourcing for RAG evaluation. Human pairwise judgments provide reliable and cost-effective results compared to LLM-based pairwise or human/LLM-based pointwise judgments, as well as automated comparisons with human-written reference responses. All our data and tools are freely available.[1]

CCS Concepts

• **Information systems → Evaluation of retrieval results**; **Language models**.

Keywords

Retrieval-Augmented Generation, Evaluation, Crowdsourcing

ACM Reference Format:
Lukas Gienapp, Tim Hagen, Maik Fröbe, Matthias Hagen, Benno Stein, Martin Potthast, and Harrisen Scells. 2025. The Viability of Crowdsourcing for RAG Evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

[1] Data: https://zenodo.org/records/14748980
Code: https://github.com/webis-de/sigir25-rag-crowdsourcing

SIGIR '25, Padua, Italy.
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1592-1/25/07
https://doi.org/10.1145/XXXXXX.XXXXXX

Table 1: Key figures of CRAGC-25 and the two crowdsourcing studies carried out to compile the corpus.

The Crowd RAG Corpus 2025			
RAG Response Writing		RAG Response Judgment	
Topics (TREC RAG'24 [30])	301	Topics (TREC RAG'24 [30])	65
RAG responses	1,806	Quality criteria [12]	7
↳ Bullet list style	301 🧑 + 301 🤖	Response pairs	1,352
↳ Essay style	301 🧑 + 301 🤖	🧑 pairwise judgments	47,320
↳ News style	301 🧑 + 301 🤖	🤖 pairwise judgments	10,556
Human workers	34	Crowd judges	420
Words per response	≈ 270	Judgments per judge	112
Time per response	≈ 11 min.	Time per judgment	≈ 0.6 min.
Hourly rate	≈ $14.40	Hourly rate	≈ $13.20
Cost per response	$2.89	Cost per (gold) judgment	($0.30) $0.06

1 Introduction

The introduction of large language models (LLMs) caused a paradigm shift in information retrieval (IR). LLMs enabled the development of a new generation of search engines that implement retrieval-augmented generation (RAG) [20]: Instead of the traditional search engine results page in the form of a ranked list of retrieved documents (list SERP), RAG systems return a text response in the style of a direct answer [29] (text SERP [12]). The goal of RAG systems is to relieve users from browsing search results by synthesizing direct, coherent, and satisfactory answers to their queries, based on information from documents retrieved by a backend search engine.

Unlike for traditional search engines, there is currently no community-wide standard for the evaluation of RAG systems in laboratory settings. Therefore, many new evaluation approaches have been proposed in the last two years (see Section 2). They can be grouped into two basic paradigms [12]: *judgment-based* evaluation, inspired by traditional IR evaluation, which ranks systems by assigning explicit utility judgments to their responses, and *reference-based* evaluation, inspired by summarization research, which ranks systems by comparing their generated response to a ground-truth reference using a similarity measure. Despite their past successful use in natural language generation tasks, the former is not very scalable, whereas the latter, apart from a high up-front overhead, still lacks similarity measures that correlate with user preferences [14, 46].

In both cases, LLMs have been proposed and used to replace human judges [8, 10, 30, 41], to generate reference texts [18], and as new, advanced similarity measures [21]. The use of LLMs to judge LLM-generated RAG responses, however, has been criticized [6, 36], arguing for humans as the only valid source of utility evidence, possibly supported by LLMs [7, 9].

To our knowledge, no systematic investigation of the capabilities and limitations of human-sourced ground-truth data for retrieval-augmented generation has not done so far. We therefore investigate the viability of crowdsourcing as a scalable alternative for RAG evaluation. With cost-effectiveness in mind, we ask how well humans can write reference RAG responses, and how reliable and valid crowdsourced RAG response judgments are, using LLM-based responses, and judgments, respectively, for comparison. As a result, we compile the Crowd RAG Corpus 2025 (CrowdRAG-25) summarized in Table 1. First, for reference-based evaluation, we collect 903 human-written and 903 LLM-generated RAG responses for all 301 topics of the TREC RAG'24 track [30], encompassing 301 responses for each of three potential RAG discourse styles: bulleted lists, essays, and news. Second, all 1,806 responses are judged by a different group of human crowd workers, and an LLM. We collect a total of 47,320 pairwise human judgments and 10,556 pairwise LLM judgments across seven RAG-specific utility dimensions [12]. Our analysis compares the writing of humans and LLMs for RAG, and the efficacy of both evaluation paradigms in obtaining reliable and valid results.

2 Related Work

We review evaluation approaches for retrieval-augmented generation systems and then discuss related work on using crowdsourcing and LLMs as sources of ground-truth in IR and beyond.

Retrieval-Augmented Generation. Retrieval-augmented generation (RAG) has been originally proposed as a means to reduce confabulations in pre-trained LLMs by conditioning them on documents relevant to a given prompt that are retrieved during generation [20]. However, this implicit form of augmentation at the attention level was quickly complemented by an explicit augmentation at the instruction level [12]: Rather than just producing statements that correctly reproduce retrieved sources with a high probability, RAG systems are now expected to explicitly cite them. Today's RAG systems replace the traditional listing of sources on a search results page (list SERP) with a summary referencing the retrieved sources (text SERP). Gao et al. [11] provide an overview of the state of the art in RAG approaches. The shape of RAG responses has quickly arrived at what amounts to a new industry standard. However, no consensus has yet been reached on how RAG systems should be evaluated. In the literature on RAG evaluation, two basic paradigms can be distinguished: reference-based evaluation and judgment-based evaluation. For both, a key point of discussion is to what extent the evaluation can be automated.

Reference-based Evaluation. Since RAG responses are a kind of multi-document summary, following summarization evaluation, reference-based evaluation measures the utility of a RAG response by comparing it with a ground-truth response on the same topic [12]. Corresponding RAG benchmarks have been developed [4, 10, 24, 25, 37, 43], where the comparison is made using a similarity measure such as BLEU [28], ROUGE [23], BERTScore [45], exact matching, or fine-tuned language models [22]. However, reference-based evaluation has been shown to not accurately differentiate the actual effectiveness of systems: For example, for news summarization, Zhang et al. [46] and Goyal et al. [14] show that human preferences often do not match reference-based results. For RAG, these measures have not been validated against human preferences.

Judgment-based Evaluation. Following traditional IR evaluation, judgment-based evaluation assigns explicit utility judgments to RAG responses. Gienapp et al. [12] review existing work and compile an overview of the different utility dimensions for RAG, grouped under the five top-level dimensions of coherence, coverage, consistency, correctness, and clarity. Hosking et al. [16] collected human judgments for a generic language generation task and found that undifferentiated judgments tend to be biased against certain utility dimensions. Other studies relying on human judgments focus only on subsets of the utility dimensions [18, 47].

LLM-based Evaluation Automation. As of recently, LLMs are being used to collect query relevance judgments (qrels) [9, 26, 31, 32, 38, 40, 41]. Moreover, LLMs have been used for 'query utility judgments' (qutils) for RAG responses as well [8, 10, 21]. Reference-based ground-truth data is also increasingly produced using LLMs across disciplines, including generating documents [39] and ground-truth responses [18]. LLMs are also used to generate synthetic training data to fine-tune evaluation models [34]. For text generation, LLMs have produced output that human test subjects found hard to distinguish from human-written texts [5], surpassing the latters' quality on specific tasks [46]. However, the use of LLMs to evaluate LLM output has been criticized as circular [6, 36].

Crowdsourcing for Evaluation. Crowdsourcing has been successfully used as a source of query relevance judgments in IR. It has been shown to produce reliable judgments [33] in a scalable manner [2] that can be validated with expert judgments [3]. Quality dimensions beyond relevance have also been successfully measured in this way [13]. However, given that Hosking et al.'s [16] results suggest that asking a human for a single, overall judgment of a RAG response tends to introduce bias, e.g., against factuality and consistency in responses, differentiated utility judgments may be necessary. Moreover, Hosking et al. corroborate the findings of multiple related studies [13, 14, 27, 46] that show that pairwise judgments produce reliable outcomes with respect to fine-grained text quality dimensions, contrary to pointwise judgments. Crowdsourcing has also been employed to collect human-written text. For example, Zhang et al. [46] conducted a crowdsourcing study to evaluate single-document news summarization, where crowd workers formulated reference summaries. Verroios and Bernstein [42] collected human-written summaries in a multi-stage writing process. Hagen et al. [15] analyzed writing progress and source usage for open-ended writing tasks with multiple source documents. The viability of crowdsourcing for collecting judgments and text has thus been extensively demonstrated, both in IR and beyond. Crowdsourcing will therefore also be useful for RAG evaluation.

3 The Crowd RAG Corpus 2025

Following established best practices for crowdsourcing [1, 44], we design two crowdsourcing studies to gather RAG responses, query utility judgments, and sufficient evidence to verify the study design.

3.1 RAG Response Writing

With the first crowdsourcing study, we gather human-written RAG responses for a set of topics, written in three different discourse styles. Similarly, LLM responses are compiled for comparison.

3.1.1 Topics, Retrieval Results, and Preliminary Steps. To crowdsource RAG responses, a set of topics is required, where for each topic a ranked list of relevant retrieval results is available. To maximize synergies with existing and future research, we reuse the 301 topics of the recent TREC RAG'24 track [30]. The track also supplies a document collection, where each document has been preprocessed to extract a total of 113M passages. One of the tasks at TREC RAG'24 was to retrieve relevant passages for RAG response generation, and we reused the system `webis-01` for our study, one of the top most effective at TREC RAG'24 with a focus on recall.

3.1.2 Study Design. For every topic and its top-20 most relevant passages, a crowd worker was tasked with composing a 250 word RAG response. The web-based writing interface we implemented shows a text editor and next to it the 20 retrieved passages as a ranked list. Since this list exceeds the screen height, scrolling through it does not move the editor out of view. A basic JavaScript-based search allows a worker to filter the passages. Workers received writing instructions, an explanation of our study, and guidelines to cite claims taken from passages using a prescribed citation format, with multiple references where applicable. Workers were also prohibited from using language models to complete the task. In this respect, we informed them that we tracked their interactions with our interface including saving the current text version at 300ms intervals until completion, as well as clicks, copy/paste events, dwell times, key presses, etc., via the BigBro library [35].

As part of our research, we were interested in studying alternative discourse styles of RAG responses. For each topic, we asked three different workers to compose their RAG response in one of the following discourse styles: (1) *bullet* list style, listing all the points relevant to the topic as a bulleted list; (2) *essay* style, starting with a clear thesis, then providing arguments, and finishing with a conclusion; and (3) *news* style, starting with the lead, then providing the important details, and lastly adding background information (i.e., the "inverted pyramid" scheme of news article).

3.1.3 Worker Recruitment. We recruit workers on the Upwork platform.[2] A detailed job ad was posted to which interested workers could apply. They were hired after manual review of their worker profile, their previous work, and successfully writing a paid response for an example topic. A total of 34 workers were recruited. On average, each worker composed 26 responses.

3.1.4 Review Process. We identify four factors that could impact the quality of the collected text responses: (1) bad input data, i.e., the topic or the retrieved passages not being suitable to formulate high-quality responses, (2) workers misunderstanding assignments, (3) usage of generative models, and (4) spam. We account for these factors through a combination of entry and exit questionnaires, manual checks, and user interaction analysis.

To check for bad input data and task understanding, workers were asked to fill out entry and exit questionnaires. The entry questionnaire asks for a self-assessment on topic knowledge and expertise, and whether a worker considers the topic answerable as well as whether it is controversial. The exit questionnaire asks for how satisfied a worker was with their response, if the provided passages were adequate sources of knowledge on the topic, and whether own prior knowledge was introduced. All questions were answered on either a 1-5 Likert scale or a ternary 'yes'/'maybe'/'no' scale. Self-assessed worker knowledge and expertise were highly correlated ($\rho = 0.88$), with expertise (median of 1/5) slightly lower than knowledge (median of 2/5). The quality of passages showed no significant correlation with the self-assessed use of own prior knowledge ($\rho = -0.04$). Prior knowledge use was generally low (assessed 1/5 for 66% of responses), while result passage quality was consistently judged high (median judgment of 4/5). In 60% of the responses, workers reported that some of the given passages were omitted. Omissions were not contingent on passage quality, as the proportion of omissions is approximately the same for each quality level. Workers are largely satisfied with their responses (87% positive). These findings indicate that (1) workers have confidence in their work, despite low initial self-assessed expertise; (2) the provided passages were considered sufficient; (3) workers relied heavily on the passages to formulate their responses; and (4) workers tend to curate the passages.

The reliability of this self-assessment, however, depends on worker faithfulness. We therefore conducted manual checks to judge overall response quality. Before hiring a worker, we provided an initial set of three topics. The submitted responses were manually screened and feedback was given to workers on assignment adherence and writing quality. If successful, a contract for additional batches of topics was drawn up. Our screening ensured that workers understand the assignment and drastically reduced spam.

To further ensure that responses were genuine human writing, we manually reviewed every response via a web interface that allows an accelerated replay of user interface interactions. We also watched out for common signs of LLM usage, such as copy-paste interactions of non-source material, sudden jumps in character count, completely homogeneous keyboard inputs, and absence of typing error corrections. This hybrid approach was trialed in a pilot study, where multiple instances of LLM usage were successfully spotted. Hired workers were made aware of our checks.

3.1.5 Scale and Cost. In total, 903 responses were collected; 86 responses were repeated due to workers failing our spam detection or screening for writing assistance tools. Some workers, who submitted responses that were rejected, were still compensated for their effort upon manual review, but all were discarded from the final response pool. A flat amount of $2.50 was paid per individual response, resulting in an hourly wage of $14.40 at an average work time of 10:22 minutes per response. The total cost for collecting the human RAG responses was $2,610.32, which includes pilot study payments, platform fees, and tax.

[2] www.upwork.com

Table 2: Pairwise operationalization of utility dimensions proposed by Gienapp et al. [12], as well as overall quality.

Dimension	**Question** (Which response...)
Topical Correctness	... better answers the query?
Logical Coherence	... is easier to follow?
Stylistic Coherence	... has a more consistent writing style?
Broad Coverage	... looks at more aspects of the topic?
Deep Coverage	... explains things more in detail?
Internal Consistency	... is clearer about how different views fit together?
Overall Quality	... is better overall?

3.1.6 LLM Responses. Using OpenAI's GPT-4o model, version 2024-08-06, we generated responses for each topic and discourse style, prompting it with the same instructions and passages given to the human workers. The total inference cost was $23.32.

3.2 Pairwise Utility Judgment

With the second crowdsourcing study, we gather pairwise judgments for seven RAG utility dimensions.

3.2.1 Study Design. We sampled 65 of the 301 TREC topics (60 topics as a 20% test split, 5 topics for pilot experiments). Each topic has six responses (three human-written, three LLM-generated). We thus add to the comparison pool per topic all 15 unique pairings of responses (975 pairs total) to be judged, and for a third of them (stratified by length and origin) the reverse pairing as well. The comparison pool comprises a total of 1,352 pairs. For each pair, utility is judged according to seven dimensions. Following Gienapp et al. [12], we adapt six of their ten dimensions, omitting external consistency and factual correctness, which require an in-depth comparative text analysis between RAG response and its sources, and both clarity dimensions, as they depend on the user information need, which has not been explicitly described for the TREC topics. Like Hosking et al. [16], we add an overall quality judgment as a baseline. All pairwise utility judgments for each of the utility dimensions permit a neutral response option, with the exception of overall quality, which mandates a definitive answer.

A single questionnaire consists of 15 individual response pairs, each with its topic, randomly sampled and stratified by response length to ensure equal workload. We employ five different workers for each questionnaire. Workers receive preliminary instructions covering the study objective, interface guide, data presentation details, and judgment protocol. We track user interaction data for spam detection via the BigBro library [35], including clicks and dwell time.

3.2.2 Worker Recruitment. We recruit workers via the Prolific platform,[3] applying selection criteria to only recruit workers whose primary language is English, who live in countries with English as official language, who have an approval rate higher than 99%, and who have previously completed at least 500 tasks. In total, 420 individual workers were recruited. On average, a worker contributed judgments for 16 response pairs, with a minimum of eleven and a maximum of 46.

3.2.3 Review Process. We identify three factors that could impact the quality of the collected judgments: (1) spam, (2) misunderstood assignment, and (3) under-performing workers. We address spam by screening submissions by total time taken, with a lower acceptance limit of 10 minutes per questionnaire, i.e., 20 seconds per text in each response pair, corresponding to fast read time of 120 words (half of average text length). Submissions faster than that were rejected and re-judged by different workers. We address the other two factors through replication. Since response pairs are distributed randomly, i.e., a different set of workers judged each pair, the probability of the majority of workers simultaneously misunderstanding or under-performing is low per response pair. Further, we use MACE [17] to estimate worker competency scores, inducing a trustworthy gold label for each pairwise comparison as a competence-weighted majority vote. Both these measures do not prevent systematic task failure, i.e., the majority of workers not being able to complete the task. We investigate this by cross-checking a sample of items with expert judgments (Section 3.3.2), and find no evidence for systematic failure (Section 5.1).

3.2.4 Scale and Cost. In total, 47,320 unique pairwise judgments were collected, 6,760 for each of the seven utility dimensions. A flat amount of $5.50 was paid per individual questionnaire, resulting in an hourly wage of $13.20, with an average time of 25:00 minutes spent. The total costs were $3,672.54, including pilot study costs, platform fees, and tax.

3.2.5 LLM Judgments. We use OpenAI's GPT-4o model, version 2024-08-06, to collect pairwise utility judgments automatically. The model received the same instructions as given to the human judges, experimenting with two setups: judging all utility dimensions simultaneously for a given topic and response pair, and separately per dimension. We collect LLM judgments on all response pairs where humans judged both directions, for a total of 10,556 individual judgments. Inference was conducted with temperature zero to ensure consistent results. The total cost was $44.70.

3.3 Verification Studies

3.3.1 Pointwise Judgments. As baseline for pairwise utility judgment, we also attempted to judge the seven utility dimensions in a pointwise manner. Each dimension is judged given (1) a short question text, prompting for a judgment between two given extremes A and B of that utility dimension (e.g., 'basic' and 'detailed' for deep coverage); (2) a description, illustrating the judged concept and providing exemplary descriptions of both extremes; and (3) a four-point Likert scale, (e.g., 'very basic', 'somewhat basic', 'somewhat detailed', and 'very detailed'). In total, 1,645 judgments were collected for a set of 47 unique, randomly chosen responses, each judged by five different workers. Workers were recruited and their work reviewed as described above (Section 3.2.2). A flat amount of $10.50 was paid per questionnaire, with an hourly wage of $13.40, for a total of $204.75, including platform fees and tax.

3.3.2 Expert Judgments. We sample a set of 30 response pairs deemed hard to decide for crowd workers (see Section 5.1). Three persons from the author team served as expert judges. They used the same study design as the crowd workers.

[3] www.prolific.com

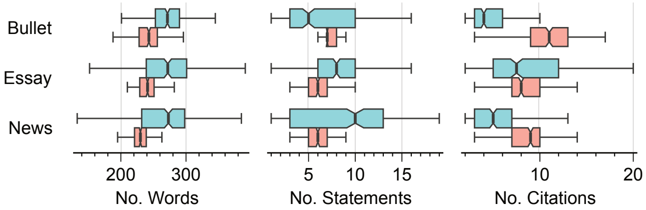


Figure 1: Distribution of space-separated words, citation-separated statements, and unique cited documents, per type and origin (🟥 LLM, 🟦 Human).

4 Analysis of RAG Response Writing

Humans and LLMs reveal distinct patterns in their writing and use of documents when formulating responses. Our investigation into these patterns is structured around five key questions: (1) How many and which of the documents are used by workers? (2) Does the given ranking order matter, or do workers treat the retrieved documents as a set? (3) Is information transferred verbatim, or paraphrased? (4) Where and how are citations located within the text? (5) How accessible is the final response to readers?

This section thus presents a systematic examination of response characteristics through descriptive statistics. In congruence with the posed questions, it analyses document selection, citation order, text reuse, attribution position and granularity, and readability, for both human workers and LLMs. We segment the raw text responses into individual statements using the explicit citation markers as delimiters, with both human workers and LLMs having received identical citation formatting instructions (bracketed numbers).

4.1 General Characteristics

Figure 1 presents the distribution of word count, statement count, and citation count over responses across writing styles and origins, excluding non-responses below 50 words. Both human workers and LLMs remain close to the 250-word target, with LLM responses exhibiting lower variance and marginally shorter lengths. They also contain fewer statements with reduced variation in count, yet employ more unique documents: humans use an average of 6.3 documents of the available 20 per response, while LLMs draw from 9.2. Moreover, human workers cite more individual documents (2.1 citations per document versus LLMs' 1.35) but incorporate fewer documents per statement (1.62 versus LLMs' 1.93). This pattern, combined with the humans' shorter statement length, indicates more granular information attribution: humans segment information into smaller units with precise document attribution, while LLMs favor broader statements, synthesizing multiple documents.

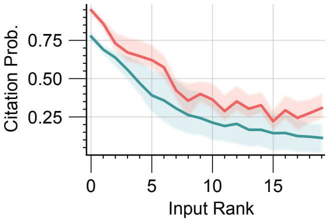
4.2 Document Selection

To more closely examine document selection criteria, we calculate the Jaccard coefficient between cited and available documents. In an extension of the overall citation counts as shown in Figure 1, the first partition of Table 3 breaks down document selection by text style. The disparity in document count is most pronounced in bullet-style responses, where humans cite their minimum of 4.4 documents against LLMs' maximum of 10.8. For essay-style responses, almost equal counts of documents are used, while news-style responses again exhibit a slightly higher count for LLM-written responses. The specific documents used differ, too, with a Jaccard overlap of only 0.33 on average between human-cited and LLM-cited documents. This selection in both cases appears to be influenced by the order of documents. In Figure 2 (left), the probability of a document being cited decreases with ranking position. In Figure 2 (right), the table shows that this effect is visible across all writing styles, as the median rank of documents (averaged across all topics) cited in the response is much lower than for uncited documents.

Table 3: Mean and 95% CI of Jaccard coefficient and Spearmans' ρ correlation, for different pairings of citation sets.

Measure	Pair		Bullet	Essay	News	Avg.
Jaccard-Coeff.	👦	🗃	0.22 ± 0.12	0.43 ± 0.25	0.29 ± 0.22	0.31
	🤖	🗃	0.54 ± 0.14	0.41 ± 0.11	0.43 ± 0.11	0.46
	👦	🤖	0.30 ± 0.17	0.38 ± 0.18	0.31 ± 0.16	0.33
Spearmans' ρ	👦	🗃	0.51 ± 0.61	0.35 ± 0.52	0.33 ± 0.64	0.39
	🤖	🗃	0.24 ± 0.39	0.34 ± 0.44	0.30 ± 0.48	0.29
	👦	🤖	0.39 ± 0.70	0.36 ± 0.62	0.32 ± 0.71	0.36

👦= Human, 🤖= LLM, 🗃= Document Ranking



Style	Cited		Uncited	
	👦	🤖	👦	🤖
Bullet	3.7	7.2	11.2	11.8
Essay	6.5	6.3	11.6	11.5
News	4.7	6.6	11.1	11.5
Avg.	5.0	6.7	11.3	11.6

Figure 2: Left: probability of a document being cited by rank position; Right: median rank of cited and uncited documents, per response style and origin (🤖/🟥 LLM, 👦/🟦 Human).

4.3 Citation Order

While rank position influences document selection, it might also influence the structure of the text, evident if the input ranking order is similar to the documents' citation order in-text. We calculate Spearman's ρ correlation coefficient between a documents' rank position and its relative position in the response, excluding uncited documents. The second partition of Table 3 details the resulting distribution of ρ. Both human workers and LLMs demonstrate weak positive correlations between citation and ranking order (first two lines), with humans exhibiting consistently higher correlations across all response styles. Bullet-style responses show the strongest adherence to initial ranking, while news-style responses score lowest. Human responses display greater variance in correlation values suggesting that they make more pronounced deviations when choosing alternative orderings. However, an average correlation of only 0.39 for humans and 0.29 for LLMs indicates that both primarily treat documents as unordered sets. When examining the overlapping subset of documents cited by both groups (last line), the low correlation (0.37) indicates that both re-rank differently.

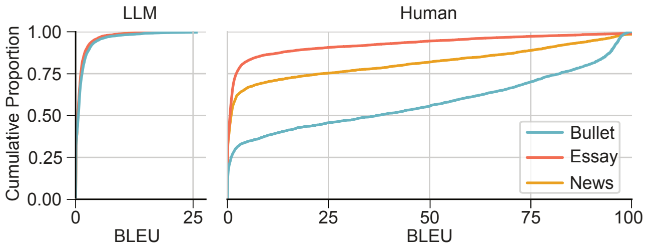


Figure 3: Cumulative proportion of statement/reference pairs per sentence-BLEU score.

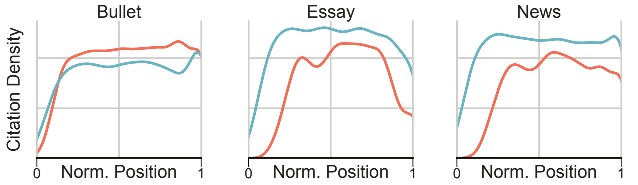


Figure 4: Relative density of citations over normalized text position per style and origin (🟥 LLM, 🟦 Human).

4.4 Text Reuse

To quantify the occurrence of text reuse, i.e., copying between documents and response, we compute sentence-BLEU [28] between each statement of a response and its cited documents, using a maximum ngram-order of 8. Figure 3 displays the cumulative score distributions per response style and origin. LLM responses consistently show sentence-BLEU scores below 25 with no deviation across styles, while human responses exhibit style-dependent variation. For news and essay styles, 75% and 88% of human responses respectively score below the maximum LLM value of 25. Bullet-style responses demonstrate substantially higher verbatim copying. This style-specific divergence reflects distinct writing strategies. A cursory qualitative analysis of bullet-style responses reveals that LLMs primarily organize responses by answer aspects, each citing multiple documents and building an abstractive micro-summary around it, while humans tend to organize responses by document rank position, favoring passage reuse and creating bullet lists that mirror list-SERP anchor texts.

4.5 Attribution Position & Granularity

Figure 4 illustrates the relative citation density across normalized text positions, revealing distinct patterns in reference distribution. All response styles exhibit an initial lag, as there is always text preceding the first citation, followed by consistent citation density throughout the text. This onset is more pronounced for LLM-written responses, indicating that their first citation occurs later on average. Additionally, essay-style responses uniquely show decreased citation density in their concluding segments, suggesting unreferenced summary statements. Humans demonstrate higher overall citation density in essay and news styles, indicating more granular source attribution. Bullet-style responses, however, show comparable citation densities between humans and LLMs.

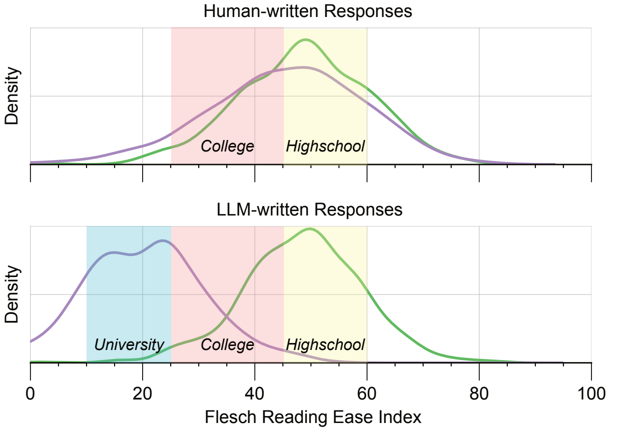


Figure 5: Distribution of Flesch reading ease index scores for responses 🟪 and their cited documents 🟩, per origin.

4.6 Readability

To judge and compare the accessibility of responses, we compare text readability scores between human and LLM-generated content, as well as the documents they cite. We calculate Flesch reading ease indices [19] at the statement level, for both the statement text and the document it cites. An aggregate score pair for each response and its sources is built using the median of its respective statements' and the median of the cited documents' scores. Figure 5 plots the distribution of both readability scores for human- and LLM-written texts, respectively. Negligible deviation was apparent between text styles in both cases, hence their distinction is omitted from the plot. Human-authored responses maintain readability levels matching their cited documents, both exhibiting normally distributed scores centered around a score of 50, corresponding to high school reading level. In contrast, while the cited documents used by LLMs follow a similar normal distribution centered around a score of 50, their generated text shows markedly lower readability with a distribution centered at slightly above 20, corresponding to college- to university-level reading ease. This suggest that LLMs tend to increase text complexity when reformulating content, potentially due to their tendency to combine aspects from multiple documents into densely packed statements, while humans, as shown in previous subsections, align their writing more closely to the cited documents.

4.7 Summary

Human and LLM writing behaviors differ in several aspects. LLMs produce responses with lower variance in length and fewer statements, but draw from a broader set of documents. Document selection differs, though both LLMs and human workers are influenced by input ranking order. Citation sequencing demonstrates weak correlations with input order, indicating treatment of the top documents as unordered sets. Text reuse analysis shows LLMs exhibit low source overlap, while humans tend to include more verbatim short-span reuse. Readability judgment demonstrates humans maintain reading ease, while LLMs generate more complex text.

Table 4: Krippendorff's α agreement for pairwise RAG utility judgment. The minimal differentiability split includes items where for a majority of dimensions, no majority vote is established by the five crowd judges. 'Pntw' denotes pointwise judgments from the verification study, $n_{\text{Items}} = 410$.

Util. D.	Complete Data $n_{\text{Items}} = 1352$		Min.-Diff. Split $n_{\text{Items}} = 290$		Diff. Split $n_{\text{Items}} = 1062$		Pntw.
	α (C)	α (C*)	α (C)	α (E)	α (C)	α (C*)	α (C*)
Top. Cor.	0.19	0.43	−0.11	−0.01	0.28	0.48	0.22
Log. Coh.	0.18	0.39	0.03	0.09	0.23	0.46	0.21
Styl. Coh.	0.11	0.38	−0.08	0.08	0.17	0.44	0.21
Brd. Cov.	0.28	0.44	−0.07	0.01	0.37	0.51	0.23
Dp. Cov.	0.28	0.45	−0.04	0.10	0.36	0.51	0.21
Int. Con.	0.14	0.42	−0.12	−0.14	0.22	0.49	0.21
Ovr. Qu.	0.17	0.39	−0.12	0.08	0.24	0.47	0.19
Mean	0.19	0.41	−0.07	0.03	0.27	**0.48**	0.21

E = Expert, C = Crowd, C* = Competency-corrected Crowd

5 Crowdsourcing for RAG Evaluation

We investigate the suitability of crowdsourcing to gather data for both reference-based and judgment-based RAG evaluation guided by three central questions: (1) Is crowdsourcing suitable for utility judgments? (2) How do human-written and LLM-generated RAG responses compare with respect to the different utility dimensions in judgment-based evaluation? (3) Do reference-based evaluation metrics succeed in reproducing the system ranking given by preference data? (4) Can LLMs successfully serve as utility judges, reproducing crowd judgment? This section thus present a systematic examination of the reliability of crowdsourced data, the two RAG evaluation paradigms, and the competing LLM-as-judge approach.

5.1 Crowdsourcing Reliability

We establish the reliability of the collected data by investigating worker agreement, presence of order bias, and interdependence of utility dimensions.

Agreement & Competency. We assess crowd judgments reliability using Krippendorff's α for ordinal judgments. Table 4 shows agreement values for the complete set of collected judgments (C). Initial agreement appears low, averaging 0.19 with substantial variance across utility dimensions. Yet, two primary sources of low agreement emerge: worker competency limitations and pairs with minimal differentiability, where both responses are too similar to enable consistent comparative judgments. To address these challenges, we employ two complementary strategies. First, we use MACE [17] to estimate worker competency scores and generate competency-corrected gold labels. Judgments submitted by the lowest-scoring workers in terms of competency were subsequently removed from the judgment pool, ensuring that at least three judgments remain for all pairs. This excludes approximately the lower quarter of the workers from consideration (C*). This substantially improves agreement, more than doubling the average to 0.41 with reduced variance. Second, we identify minimally differentiable item pairs through voting behavior: pairs lacking a majority vote (3 out of 5) across a majority of utility dimensions (4 out of 7). Approximately 20% of pairs fall into this category. A targeted expert validation of a 30-item sub-sample reveals that while experts demonstrate higher absolute agreement, both crowd and expert workers fail to establish a reliable vote. Notably, experts show a higher tendency for neutral choices, which crowd workers tend to avoid, thereby amplifying disagreement. We thus conclude that agreement is negatively impacted by both a small set of low-competency crowd workers and a small set of minimally differentiable items. After applying correction for both factors, the resulting gold labels are subject to agreement levels (marked bold in Table) comparable and even exceeding previous IR and NLP literature [14, 46] and we deem them reliable. Additionally, investigate the agreement attainable in a pointwise study design (Section 3.3.1) for comparison. After applying the same MACE competency correction, the resulting agreement is still very low at 0.21, not better than the uncorrected pairwise scores. The pairwise study design, while more expensive, thus offers a substantial improvement in reliability. In future studies, more rigorous screening for worker competency could help improve cost efficiency.

Table 5: Krippendorff's α agreement and maximum delta between individual directions (←/→) and combined set (↔) for within-item order effects, and 1^{st} and 2^{nd} half of items in a questionnaire for across-item order effects.

Util. D.	Within-item Order $n_{\text{Items}} = 377$				Across-item Order $n_{\text{Items}} = 1352$		
	$\alpha(\leftarrow)$	$\alpha(\rightarrow)$	$\alpha(\leftrightarrow)$	$\Delta_{\max}$	$\alpha(1^{\text{st}})$	$\alpha(2^{\text{nd}})$	Δ
Top. Cor.	0.18	0.19	0.17	0.02	0.18	0.20	0.01
Log. Coh.	0.18	0.19	0.18	0.01	0.18	0.18	0.00
Styl. Coh.	0.08	0.14	0.10	0.04	0.11	0.11	0.00
Brd. Cov.	0.27	0.28	0.26	0.02	0.28	0.28	0.00
Dp. Cov.	0.27	0.27	0.26	0.01	0.28	0.27	0.01
Int. Con.	0.13	0.15	0.14	0.01	0.13	0.16	0.02
Ovr. Qu.	0.16	0.17	0.15	0.02	0.15	0.18	0.03
Mean	0.18	0.20	0.18	0.01	0.19	0.20	0.01

Order Bias. We distinguish two types of order effects: within-item and across-item. Within-item order effects occur when the presentation sequence of two responses within a single questionnaire item influences judge decisions, such as a preference for the first option. Across-item order effects occur when the sequence of items in a questionnaire systematically impacts judgments, potentially due to learning or fatigue. As presented in Table 5, for within-item order, we examine the subset of comparisons where both response directions were independently judged. Under order bias, we would expect higher agreement among workers when analyzing each direction (← /→) separately compared to pooled results (↔). For across-item order, we split the dataset by item position in each worker's questionnaire, i.e., the first seven response pairs (1^{st}) and the remaining response pairs (2^{nd}) of each questionnaire. As questionnaire order is randomized, any difference in agreement between the two halves would originate from study design, not data characteristics. For both effect types, the agreement values remain nearly identical, both overall as well as for individual utility dimensions, suggesting that judge decisions are robust to presentation order.

Table 6: Kendall's τ correlation cross-tabulation between gold labels of all quality dimensions. Columns use initial letters to designate utility dimensions.

Utility Dimension	T	L	S	B	D	I	O
Topical Correctness		0.30	0.35	0.42	0.46	0.45	0.52
Logical Coherence	0.30		0.34	0.18	0.20	0.30	0.37
Stylistic Coherence	0.35	0.34		0.23	0.29	0.36	0.36
Broad Coverage	0.42	0.18	0.23		0.57	0.38	0.42
Deep Coverage	0.46	0.20	0.29	0.57		0.40	0.43
Internal Consinstency	0.45	0.30	0.36	0.38	0.40		0.45
Overall Quality	0.52	0.37	0.36	0.42	0.43	0.45	
Mean	0.42	0.28	0.32	0.37	0.39	0.39	0.43

Interdependence of Utility. To assess crowd workers' ability to discriminate utility dimensions, we investigate their interdependence by computing Kendall's τ cross-correlation between gold labels across dimensions as shown in Table 6. Correlation values range from 0.18 to 0.57, with a mean of 0.37. While overall quality exhibits the highest average correlation (0.42), the low inter-dimensional correlations of other utility types suggest successful differentiation. We further find indications of thematic clustering: coverage dimensions and topical correctness – addressing response groundedness – demonstrate higher inter-correlation while, conversely, exhibiting distinctly low correlation with stylistic and logical coherence dimensions, which address response presentation.

5.2 Judgment-based Evaluation

With the reliability of the preference data established, we operationalize the judgment-based evaluation approach, ranking the given responses per topic based on the pairwise comparisons, in order to comparatively quantify the utility of human-written and LLM-generated RAG responses. To rank the six different responses within each topic based on the pairwise preferences, we use a variant of the probabilistic Bradley-Terry model adapted to handle ties and contradictory comparisons [13]. It has been successfully applied to infer robust scalar scores from crowdsourced pairwise labels on text quality before [13]. We compute ranks independently for each topic and utility dimension, and assign based on them descending 'grades' between 6 and 1 (higher is better). This allows us to compare grades across topics. Table 7 shows the mean grade for each combination of text style, origin, and utility dimension. Further, it includes averages per style without distinguishing origin (columns 'Both'), per origin without distinguishing style (column 'All Styles'), as well as across all utility dimensions (last row).

LLMs consistently received higher grades than human workers across most evaluation dimensions. This difference is statistically significant (Wilcoxon signed rank test, $\alpha = 0.05$, Benjamini-Hochberg correction) for most style-dimension combinations, with two notable exceptions: logical coherence in essay-style responses, both coverage dimensions (broad and deep) in news-style responses, and deep coverage for bullet-style responses. When aggregating across styles, LLMs higher grading remained significant for all utility dimensions except coverage. In a style-based analysis independent of authorship, bullet-style responses significantly outperformed both essay and news styles, which are graded on-par (Wilcoxon signed rank test, $\alpha < 0.05$, Benjamini-Hochberg correction). Bullet-style responses showed particular strength in logical coherence, coverage metrics, and overall quality, while stylistic coherence and internal consistency showed no significant style-based differences. Between essay and news styles, the only significant difference emerged in deep coverage, favoring essays.

Table 7: Mean response grade per style, origin, and utility dimension. 👱🤖 designates union of both origins. Grade is on a scale of 1–6, higher is better.

Util.	Bullet			Essay			News			All Styles	
	👱	🤖	👱🤖	👱	🤖	👱🤖	👱	🤖	👱🤖	👱	🤖
Top.	2.8	4.5	3.7	3.1	4.1	3.6	2.7	3.8	3.2	2.9	4.2
Log.	4.2	5.4	4.8	2.7	3.1	2.9	2.4	3.3	2.8	3.1	3.9
Sty.	3.2	4.3	3.7	2.6	4.2	3.4	2.6	4.2	3.4	2.8	4.2
Bro.	3.5	4.6	4.0	2.9	3.8	3.3	2.8	3.4	3.1	3.1	3.9
Dee.	3.6	4.3	3.9	3.1	4.1	3.6	2.7	3.3	3.0	3.1	3.9
Int.	3.0	4.3	3.6	2.9	4.2	3.6	2.6	3.9	3.3	2.9	4.1
Ovr.	3.3	4.9	4.1	2.7	4.0	3.4	2.5	3.6	3.1	2.8	4.2
Avg.	3.4	4.6	3.9	2.9	3.9	3.4	2.6	3.6	3.1	3.0	4.1

Table 8: Spearman rank correlation between label-induced ranking and ranking by content overlap metrics. Best/Worst only compares relative order of highest and lowest-ranked according to label ranking.

Util. Dim.	Best/Worst			Full Ranking		
	BERTS.	BLEU	RougeL	BERTS.	BLEU	RougeL
Topical Cor	0.354	0.354	0.169	0.172	0.211	0.094
Logical Coh.	0.354	0.477	0.446	0.288	0.337	0.294
Stylistic Coh.	0.385	0.385	0.385	0.263	0.315	0.243
Broad Cov.	0.231	0.354	0.231	0.183	0.254	0.228
Deep Cov.	0.262	0.415	0.385	0.211	0.269	0.223
Internal Con.	0.323	0.323	0.323	0.191	0.208	0.180
Overall Qu.	0.385	0.323	0.292	0.292	0.268	0.255

5.3 Reference-based Evaluation

To investigate the efficacy of content-overlap and similarity metrics for RAG evaluation, we leverage the pairwise-judgment-based response rankings established per topic previously. For each topic, we designated the highest-ranked response as the reference and generated rankings for the remaining five candidate responses. We use three metrics to rank candidates with respect to the reference: the contextualized-embedding-based BERTScore [45], and two ngram overlap measures - sentence-BLEU [28] and RougeL [23]. We then compute Spearman's rank correlation between these metric-induced rankings and the ground-truth rankings derived from human judgments. This approach tests the hypothesis that effective reference-based evaluation produces rankings consistent with human judgment. Table 8 presents correlation values across utility dimensions, examining both binary discrimination (correct ordering of best and worst candidate according to ground-truth ranking) and full ranking correlation. Across all metrics and utility dimensions, correlation remains low. The best/worst correlation is slightly better than full ranking correlation, indicating that higher differentiability

Table 9: Krippendorff α agreement between LLM prompt variations (α, 🤖/🤖) / LLM and human gold judgment (α, 🤖/👦), and mean judgment correlation to other dimensions ($\bar{\rho}$, 🤖) for different prompting strategies by utility dimensions.

Utility Dimension	Combined Inference			Individual Inference			Both
	α		$\bar{\rho}$	α		$\bar{\rho}$	α
	🤖/🤖	🤖/👦	🤖	🤖/🤖	🤖/👦	🤖	🤖/🤖
Topical Correctness	0.65	0.10	0.80	0.87	0.12	0.86	0.48
Logical Coherence	0.78	0.16	0.74	0.84	0.16	0.82	0.53
Stylistic Coherence	0.82	0.14	0.71	0.88	0.13	0.82	0.53
Broad Coverage	0.65	0.08	0.64	0.58	0.06	0.74	0.40
Deep Coverage	0.38	0.01	0.72	0.37	0.00	0.74	0.32
Internal Consinstency	0.66	0.12	0.77	0.90	0.13	0.84	0.49
Overall Quality	0.64	0.06	0.81	0.86	0.09	0.86	0.47

of responses within the respective quality dimension corresponds to better efficacy of reference-based evaluation; yet, fine-grained accurate system rankings, as in the full correlation setup, remain problematic. BLEU shows highest correlation for all utility dimensions except overall quality, where BERTScore performs best. The highest correlation in all metrics and both is attained for logical coherence. Yet, given an absolute maximum value of merely 0.477 (BLEU for logical coherence), none of the utility dimensions are accurately measurable by any of the three metrics.

5.4 LLMs as Judges of Utility

To investigate the ability of LLMs as utility judges, we assess three critical properties: (1) consistency in judgment across conditions, (2) correctness relative to human gold labels, and (3) dimensional differentiation in utility judgment. Table 9 quantifies these properties using Krippendorff's α agreement and Spearman's ρ correlation across the two inference settings (Section 3.2.5). For combined inference—simultaneously judging all utility dimensions—we first analyze bidirectional LLM-to-LLM agreement (first column) to detect prompt order effects. While substantial agreement exists, prompt-induced inconsistencies are present. Further, LLM-to-human agreement (second column) shows only little overlap with gold labels, calling into question the correctness of LLM judgments. This could be due to their high cross-dimensional correlation (third column), indicating detection of a general preference in each comparison, rather than a fine-grained view of individual dimensions. Individual inference, where dimensions are evaluated separately, exhibits similar patterns: while improved LLM-to-LLM agreement suggests improved consistency, LLM-to-human agreement remains poor. Increased dimensional correlation in the individual setting indicates that combined prompting, despite lower consistency, better facilitate dimensional differentiation, possibly due to the model being explicitly made aware of other dimensions to be judged.

5.5 Summary

Crowdsourcing can be a reliable source of judgment data for RAG evaluation when controlling for worker competence. Analysis of order effects and utility dimension interdependence confirmed judgments robustness. Using these validated judgments, we find that LLM-generated responses exhibit significantly higher quality than human-written ones across most utility dimensions. Bullet-style responses are preferred to essay and news styles. However, contrary to judgment-based evaluation, reference-based evaluation is not easily operationalizable, as all tested evaluation metrics demonstrated severe limitations. This suggests that judgment-based approaches are a more promising approach to RAG evaluation. Yet, using LLMs as utility judges, in an effort to improve the efficiency of judgment-based evaluation, fails to produce consistent and correct judgments across all tested settings.

6 Conclusion

We summarize our main findings below, grouped into those relevant to RAG model development, and those relevant to RAG evaluation. We end our paper with a reflection on its limitations and include ethical considerations on study design and attained results.

Findings for RAG Model Development. We find response style exerts significant influence on perceived utility of responses; while most current RAG models focus on continuous text, akin to the essay style, bullet-style responses exhibit higher preference judgments. This prompts future work on style adaption for RAG, as specific style instructions can yield stronger preferences from the same model. Furthermore, readability of LLM-generated responses could be improved, scoring worse than human-written responses.

Findings for RAG Evaluation. Reference-based evaluation fails to accurately operationalize any of the evaluated utility dimensions. This is in line with previous findings [14, 46]. We thus discourage its use and instead demonstrate that judgment-based evaluation is a feasible and accurate alternative when using crowdsourced data. The increasingly adopted practice of using LLMs as judges of utility, however, is concerning, as they fail to produce consistent and correct judgments in a zero-shot setting. For future work, we will study label transfer from a pool of existing human-judged responses to a new, unjudged responses, akin to a few-shot setting. The utility dimensions previously motivated theoretically [12] have been shown empirically to be distinguishable by human judges.

Limitations & Ethical Considerations. We acknowledge methodological constraints associated with our study. The exclusive use of TREC RAG 2024 topics may limit generalizability. While the utility dimensions we employ aim for comprehensiveness, they may not capture all aspects of RAG system effectiveness relevant to practitioners. The crowdsourcing setup may introduce worker quality variations and cultural biases in both tasks. Furthermore, our investigation of LLM-written responses relies on a single model configuration, leaving an investigation of other LLM configurations and prompting strategies for future work, which can be compared or validated against our collected data.

The collection of human-written responses and human judgments raises privacy concerns, which we mitigated through informed consent and comprehensive data anonymization. Crowdsourcing raises concerns about fair labor practices, which we addressed by implementing above-market compensation rates, flexible time allocations, and transparent task descriptions. Overall, our work contributes to a broader discussion on the socio-technical implications of AI-generated responses for search, and highlights the need for RAG evaluation practices grounded in human judgment.

Acknowledgments

This publication has been partially supported by the ScaDS.AI Center for Scalable Data Analytics and Artificial Intelligence, funded by the Federal Ministry of Education and Research of Germany and by the Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus; by a Research Fellowship for Harrisen Scells from the Alexander von Humboldt Foundation; and by the OpenWebSearch.eu project, funded by the European Union (GA 101070014).

References

[1] Omar Alonso. 2015. Practical Lessons for Gathering Quality Labels at Scale. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto (Eds.). ACM, 1089–1092. https://doi.org/10.1145/2766462.2776778

[2] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Inf. Process. Manag.* 48, 6 (2012), 1053–1066. https://doi.org/10.1016/J.IPM.2012.01.004

[3] Omar Alonso, Stefano Mizzaro, et al. 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, Vol. 15. 16.

[4] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking Large Language Models in Retrieval-Augmented Generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 17754–17762. https://doi.org/10.1609/AAAI.V38I16.29728

[5] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 7282–7296. https://doi.org/10.18653/V1/2021.ACL-LONG.565

[6] Charles L. A. Clarke and Laura Dietz. 2024. LLM-based Relevance Assessment Still Can't Replace Human Relevance Assessment. *CoRR* abs/2412.17156 (2024). https://doi.org/10.48550/arXiv.2412.17156 arXiv:2412.17156

[7] Laura Dietz. 2024. A Workbench for Autograding Retrieve/Generate Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 1963–1972. https://doi.org/10.1145/3626772.3657871

[8] Shahul ES, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - System Demonstrations, St. Julians, Malta, March 17-22, 2024*, Nikolaos Aletras and Orphée De Clercq (Eds.). Association for Computational Linguistics, 150–158. https://aclanthology.org/2024.eacl-demo.16

[9] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *13th ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR 2023)*, Masaharu Yoshioka, Julia Kiseleva, and Mohammad Aliannejadi (Eds.). ACM, Taipei, Taiwan, 39–50. https://doi.org/10.1145/3578337.3605136

[10] Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems. *CoRR* abs/2407.11005 (2024). https://doi.org/10.48550/ARXIV.2407.11005 arXiv:2407.11005

[11] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *CoRR* abs/2312.10997 (2023). https://doi.org/10.48550/ARXIV.2312.10997 arXiv:2312.10997

[12] Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Evaluating Generative Ad Hoc Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 1916–1929. https://doi.org/10.1145/3626772.3657849

[13] Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient Pairwise Annotation of Argument Quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5772–5781. https://doi.org/10.18653/V1/2020.ACL-MAIN.511

[14] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. *CoRR* abs/2209.12356 (2022). https://doi.org/10.48550/ARXIV.2209.12356 arXiv:2209.12356

[15] Matthias Hagen, Martin Potthast, Michael Völske, Jakob Gomoll, and Benno Stein. 2016. How Writers Search: Analyzing the Search and Writing Logs of Nonfictional Essays. In *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13-17, 2016*, Diane Kelly, Robert Capra, Nicholas J. Belkin, Jaime Teevan, and Pertti Vakkari (Eds.). ACM, 193–202. https://doi.org/10.1145/2854946.2854969

[16] Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human Feedback is not Gold Standard. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=7W3GLNImfS

[17] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning Whom to Trust with MACE. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (Eds.). The Association for Computational Linguistics, 1120–1130. https://aclanthology.org/N13-1132/

[18] Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. HAGRID: A Human-LLM Collaborative Dataset for Generative Information-Seeking with Attribution. *CoRR* abs/2307.16883 (2023). https://doi.org/10.48550/ARXIV.2307.16883 arXiv:2307.16883

[19] J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (Automated Reliability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report 8-75). Memphis, TN: Naval Air Station; 1975. *Naval Technical Training, US Naval Air Station: Millington, TN* (1975).

[20] Patrick S. H. 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. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[21] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. *CoRR* abs/2411.16594 (2024). https://doi.org/10.48550/ARXIV.2411.16594 arXiv:2411.16594

[22] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. *CoRR* abs/2412.05579 (2024). https://doi.org/10.48550/ARXIV.2412.05579 arXiv:2412.05579

[23] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[24] Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. RECALL: A Benchmark for LLMs Robustness against External Counterfactual Knowledge. *CoRR* abs/2311.08147 (2023). https://doi.org/10.48550/ARXIV.2311.08147 arXiv:2311.08147

[25] Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2024. CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models. *CoRR* abs/2401.17043 (2024). https://doi.org/10.48550/ARXIV.2401.17043 arXiv:2401.17043

[26] Sean MacAvaney and Luca Soldaini. 2023. One-Shot Labeling for Automatic Relevance Estimation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2230–2235. https://doi.org/10.1145/3539618.3592032

[27] Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2018. RankME: Reliable Human Ratings for Natural Language Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 72–78. https://doi.org/10.18653/V1/N18-2012

[28] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, 186–191. https://www.aclweb.org/anthology/W18-6319

[29] Martin Potthast, Matthias Hagen, and Benno Stein. 2020. The dilemma of the direct answer. *SIGIR Forum* 54, 1 (2020), 14:1–14:12. https://doi.org/10.1145/3451964.3451978

[30] Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Ragnarök: A Reusable RAG Framework and Baselines for TREC 2024 Retrieval-Augmented Generation Track. *CoRR* abs/2406.16828 (2024). https://doi.org/10.48550/ARXIV.2406.16828 arXiv:2406.16828

[31] Hossein A. Rahmani, Xi Wang, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Paul Thomas. 2024. SynDL: A Large-Scale Synthetic Test Collection for Passage Retrieval. *CoRR* abs/2408.16312 (2024). https://doi.org/10.48550/ARXIV.2408.16312 arXiv:2408.16312

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

[33] Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Falk Scholer. 2021. On the effect of relevance scales in crowdsourcing relevance assessments for Information Retrieval evaluation. *Inf. Process. Manag.* 58, 6 (2021), 102688. https://doi.org/10.1016/J.IPM.2021.102688

[34] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (Eds.). Association for Computational Linguistics, 338–354. https://doi.org/10.18653/V1/2024.NAACL-LONG.20

[35] Harrisen Scells, Jimmy, and Guido Zuccon. 2021. *Big Brother:* A Drop-In Website Interaction Logging Service. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2590–2594. https://doi.org/10.1145/3404835.3462781

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

[37] Yixuan Tang and Yi Yang. 2024. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. *CoRR* abs/2401.15391 (2024). https://doi.org/10.48550/ARXIV.2401.15391 arXiv:2401.15391

[38] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models can Accurately Predict Searcher Preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 1930–1940. https://doi.org/10.1145/3626772.3657707

[39] Mehmet Deniz Türkmen, Mucahid Kutlu, Bahadir Altun, and Gokalp Cosgun. 2025. GenTREC: The First Test Collection Generated by Large Language Models for Evaluating Information Retrieval Systems. *CoRR* abs/2501.02408 (2025). https://doi.org/10.48550/ARXIV.2501.02408 arXiv:2501.02408

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

[41] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2024. A Large-Scale Study of Relevance Assessments with Large Language Models: An Initial Look. *CoRR* abs/2411.08275 (2024). https://doi.org/10.48550/ARXIV.2411.08275 arXiv:2411.08275

[42] Vasilis Verroios and Michael S. Bernstein. 2014. Context Trees: Crowdsourcing Global Understanding from Local Views. In *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*, Jeffrey P. Bigham and David C. Parkes (Eds.). AAAI, 210–219. https://doi.org/10.1609/HCOMP.V2I1.13149

[43] Shuting Wang, Jiongnan Liu, Shiren Song, Jiehan Cheng, Yuqi Fu, Peidong Guo, Kun Fang, Yutao Zhu, and Zhicheng Dou. 2024. DomainRAG: A Chinese Benchmark for Evaluating Domain-specific Retrieval-Augmented Generation. *CoRR* abs/2406.05654 (2024). https://doi.org/10.48550/ARXIV.2406.05654 arXiv:2406.05654

[44] Meng-Han Wu and Alexander J. Quinn. 2017. Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2017, 23-26 October 2017, Québec City, Québec, Canada*, Steven Dow and Adam Tauman Kalai (Eds.). AAAI Press, 206–215. https://doi.org/10.1609/HCOMP.V5I1.13317

[45] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr

[46] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen R. McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Trans. Assoc. Comput. Linguistics* 12 (2024), 39–57. https://doi.org/10.1162/TACL_A_00632

[47] Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-Hong Huang, and Evangelos Kanoulas. 2024. Towards Fine-Grained Citation Evaluation in Generated Text: A Comparative Analysis of Faithfulness Metrics. In *Proceedings of the 17th International Natural Language Generation Conference, INLG 2024, Tokyo, Japan, September 23 - 27, 2024*, Saad Mahamood, Minh Le Nguyen, and Daphne Ippolito (Eds.). Association for Computational Linguistics, 427–439. https://aclanthology.org/2024.inlg-main.35