Perspectives on Large Language Models for Relevance Judgment

Guglielmo Faggioli
University of Padova

Laura Dietz
University of New Hampshire

Charles L. A. Clarke
University of Waterloo

Gianluca Demartini
University of Queensland

Matthias Hagen
Friedrich-Schiller-Universität Jena

Claudia Hauff
Spotify

Noriko Kando
National Institute of Informatics (NII)

Evangelos Kanoulas
University of Amsterdam

Martin Potthast
Leipzig University and ScaDS.AI

Benno Stein
Bauhaus-Universität Weimar

Henning Wachsmuth
Leibniz University Hannover

ABSTRACT

When asked, large language models (LLMs) like ChatGPT claim that they can assist with relevance judgments but it is not clear whether automated judgments can reliably be used in evaluations of retrieval systems. In this perspectives paper, we discuss possible ways for LLMs to support relevance judgments along with concerns and issues that arise. We devise a human–machine collaboration spectrum that allows to categorize different relevance judgment strategies, based on how much humans rely on machines. For the extreme point of 'fully automated judgments', we further include a pilot experiment on whether LLM-based relevance judgments correlate with judgments from trained human assessors. We conclude the paper by providing opposing perspectives for and against the use of LLMs for automatic relevance judgments, and a compromise perspective, informed by our analyses of the literature, our preliminary experimental evidence, and our experience as IR researchers.

CCS CONCEPTS

• **Information systems → Relevance assessment**.

KEYWORDS

large language models, relevance judgments, human–machine collaboration, automatic test collections

1 INTRODUCTION

Evaluation is very important to the information retrieval (IR) community and the difficulty of proper evaluation setups is well-known and often discussed (e.g., [40, 54, 68, 70]). Many long-standing evaluation campaigns like TREC, NTCIR, CLEF, or FIRE [15, 42, 47, 56] trace their roots back to the Cranfield paradigm [20], which relies on test collections that consist of (i) a document corpus, (ii) a set of information needs or topics, and (iii) relevance judgments for documents on the topics. Critically, according to the Cranfield paradigm, human assessors are needed for the relevance judgments—a time-intensive and costly procedure.[1]

However, over the past decades, in IR more and more tasks have been delegated to machines that were traditionally performed by humans, starting with indexing and retrieval. While the idea of automatically generated judgments has been considered before [77], it has not found widespread use in the IR community. Other previous ideas to minimize the cost of collecting relevance judgments include judging text nuggets instead of documents [66], using crowdworkers [3, 14] (though this comes with its own set of problems [63]), cleverly selecting which documents to judge [17, 55], constructing test collections from Wikipedia [30], or automating parts of the judgment process via a QA system [69].

> Dear LLM, can you assist us with relevance judgments?
>
> Yes, I can certainly provide assistance with relevance judgments. Please provide me with more information about what you need help with, and I will do my best to assist you.
> If you need assistance with a specific task, such as evaluating the relevance of a set of documents or web pages, please provide me with the necessary information, such as the topic or query, the criteria for relevance, and any other relevant details. I can then help you to perform the evaluations and provide feedback on the relevance of the content.

Figure 1: Asking ChatGPT for assistance on February 15, 2023.

[1]As a concrete example, for the 50 topics in the TREC-8 Ad Hoc track [81], 129 participating systems led to more than 86,000 pooled documents to judge, requiring more than 700 assessor hours at a cost of about USD 15,000.

We thank Ian Soboroff for his ideas, comments, and other contributions.

ICTIR '23, July 23, 2023, Taipei, Taiwan

ACM ISBN 979-8-4007-0073-6/23/07…$15.00
https://doi.org/10.1145/3578337.3605136

Figure 1 shows the response of ChatGPT[2] when asked whether it can assist with relevance judgments. ChatGPT suggests that it is able to carry out relevance judgments, but it is unclear how well such judgments align with those made by human annotators. In this perspectives paper, we explore whether we are on the verge of being able to delegate relevance judgments to machines—either fully or partially—by employing large language models (LLMs). We aim to provide a balanced view on this contentious question by presenting both consenting and dissenting voices in the scientific debate surrounding the use of LLMs for relevance judgments. Although a variety of document modalities exist (audio, video, images, text), we here focus on text-based test collections. The consolidated methodology for assessing the relevance of textual documents, which dates back to the Cranfield paradigm, enables us to carry out a grounded comparison between LLMs and human assessors. While the technology might not be ready yet to provide fully automatic relevance judgments, we argue that LLMs are already able to help humans in relevance assessment—to various extents. To model the range of automation options, we propose and discuss a spectrum of collaboration between humans and LLMs (cf. Table 1): from manual judgments, the current setup, to fully automated judgments that are carried out solely by LLMs as a potential future option.

Some of the spectrum's scenarios have already been studied (cf. Section 2), while others are currently emerging. We describe risks as well as open questions that require further research and we conduct a pilot feasibility experiment where we assess how well judgments generated by LLMs agree with humans, including an analysis of LLM-specific caveats. To conclude our paper, we provide two opposing perspectives—for and against the use of LLMs as relevance "assessors"—, as well as a compromise between them. All of the perspectives are informed by our analyses of the literature, our pilot experimental evidence, and our experience as IR researchers.

2 RELATED WORK

Following the Cranfield paradigm, a test collection-based approach to IR evaluation requires documents, queries, and relevance judgments for query–document pairs. The traditional approach to acquire relevance judgments is to hire human assessors. However, the judgment effort is staggering, leading to a range of approaches to assist the assessors or to automate tedious tasks. Below, we describe existing approaches and relate them to our human–machine collaboration spectrum (cf. Table 1).

2.1 Human Judgment

As document collections kept growing in size, the ratio of documents that could practically be judged by human assessors kept getting smaller. This triggered the IR community to look for ways to scale-up human-generated relevance judgments. Around 2010, replacing trained human assessors by micro-task crowdsourcing became an option [3] so that the community started to study the reliability of crowdsourced relevance judgments [14] and questions related to cost and quality management [63].

The workforce increased via crowdsourcing usually comes with a decreased reliability, often due to the complicated interactions between crowdworkers and task requesters [65]. Still, before the advent of large language models, several studies showed that crowdsourcing is a viable alternative to scale-up relevance judgments compared to the "classic" hiring of trained human assessors—as long as the domain is accessible to non-experts and quality control mechanisms are put in place [79]. Quality control mechanisms may include label aggregation methods [75], task design strategies [2, 58], and crowdworker selection strategies [41].

Some studies have tried to increase the judgment efficiency of crowdworkers by adding machine-generated information (e.g., metadata) [85] but recent findings suggest that LLMs alone are even better at several text annotation tasks than crowdworkers [43].

2.2 Human Verification and AI Assistance

In this scenario, humans partially relinquish control over which documents will be assessed or how machine assessments will be derived but humans remain in control of defining relevance.

For example, some studies suggest to adjust evaluation metrics to be able to deal with incomplete judgments (e.g., [39, 87]). This way, judgment costs can be reduced by reducing the number of assessments needed for evaluating retrieval systems.

Alternatively, Keikha et al. [59] suggest to automatically transfer manual relevance judgments in the context of passage retrieval: any unjudged passage that has a high similarity to a judged passage will inherit the judged passage's relevance label on a given topic. In the original setup, the authors used ROUGE as the similarity measure but also "modern" alternatives like BertScore [90] could be tried—as transferring relevance judgments between corpora without proper similarity checking is problematic [38].

Other ideas for semi-automatic relevance judgments are active learning [21] (e.g., human assessors only label documents for which an automatic relevance assessment has a low confidence) or to automatically identify potentially relevant documents that only manual runs would contribute to the pool [55]—in order to construct low-bias reusable test collections.

Instead of asking humans for relevance assessments on query–document pairs, Sander and Dietz [69] suggest to ask humans for (exam) questions related to a query / topic that should be answerable from the content of a relevant document. The more of the manually formulated questions an automatic question-answering system can answer, the more relevant a to-be-judged document is—captured by the authors' proposed EXAM answerability metric. Similar ideas have also been used successfully in other labeling tasks [29, 32, 51].

2.3 Fully Automated Test Collections

Inspired by ideas from evaluating aspect-based summarization [49] or text segmentation [6], the Wikimarks approach [30] aims to automatically create queries and judgments for a test collection. The title and subheadings of Wikipedia articles are used to formulate queries and the passage below the title / heading is assumed to be relevant for the respective query—without actual human judgments. Similar distant supervision-style approaches to acquire relevance assessments for "artificial" queries exploit other facets of human-authored (semi-structured) documents: anchor text [7], metadata of scientific articles [13], categories in the Open Directory Project [11], glosses in Freebase [26], or infoboxes [50, 57].

[2]https://chat.openai.com/chat

Table 1: A spectrum of collaborative human – machine task organization to produce relevance judgments. The △ indicates where on the spectrum each possibility falls.

Collaboration Integration	Task Organization
Human Judgment	
	Humans do all judgments manually without any kind of support.
	Humans have full control of judging but are supported by text highlighting, document clustering, etc.
AI Assistance	
	Humans judge documents while having access to LLM-generated summaries.
	Balanced competence partitioning. Humans and LLMs focus on (sub-)tasks they are good at.
Human Verification	
	Two LLMs each generate a judgment, and humans select the better one.
	An LLM produces a judgment (and an explanation) that humans can accept or reject.
$\cdot n$	LLMs are considered crowdworkers with varied specific characteristics, but supervised / controlled by humans.
Fully Automated	
	Fully automatic judgments.

Also for the task of query performance prediction (QPP) [16, 48], the goal is to estimate retrieval effectiveness (i.e., the ability to return relevant results) without having manual relevance judgments—often even without knowing the actual retrieval results. While some recent studies effectively used LLMs in QPP scenarios [4, 5, 18, 28, 34], an open question still is how well LLM-based relevance assessments agree with manual assessments. A study on the leaderboards of the TREC CAR track found a very high rank correlation [31] and some preliminary evidence seems to indicate that LLMs can replace human assessors in several NLP tasks [92]—with a high variance in the quality of the annotations, though—, but MacAvaney and Soldaini [62] found that automatic relevance judgments may correlate poorly with human assessments. Still, system leaderboards obtained from the automatic relevance judgments were comparable to those based on manual assessments.

3 SPECTRUM OF HUMAN–MACHINE COLLABORATION

To discuss potential capabilities of LLMs in the context of relevance judgments, we devise a human–machine collaboration spectrum with different levels of "labor division" between humans and LLMs (cf. Table 1 for an overview). At one end, humans manually judge without any LLM interaction, while at the other end, LLMs replace humans completely. In between, LLMs assist humans at various degrees of interdependence.

Human Judgment. On this end of the spectrum, humans manually decide what is relevant without being influenced by an LLM. In reality, of course, humans are still supported by basic features of a judgment interface. Such features might still be based on heuristics that do not require any form of automatic training / feedback. For instance, humans may define so-called scan terms to be highlighted in a text, they may limit viewing the pool of documents that have already been judged, or they may order documents by similarity so that it is easier to assign the same relevance label to similar documents. This end of the spectrum thus represents the status quo, where humans are considered the only reliable judges.

AI Assistance. More advanced assistance can come in many forms. For example, an LLM may generate a summary of a to-be-judged document so that a human assessor can more efficiently make a judgment based on the compressed representation. Another approach could be to manually define information nuggets that are relevant (e.g., exam questions / answers [69]) and to then train an LLM to automatically indicate how many test nuggets are contained in a to-be-judged document (e.g., via a QA system). This directly implies questions towards improving the human–machine collaboration: How to employ LLMs, as well as other AI tools, to aid human assessors in devising reliable judgments while enhancing the efficiency of the process? What are tasks that can be taken over by LLMs (e.g., document summarization or keyphrase extraction)?

Human Verification. For each document to judge, a first-pass judgment of an LLM is automatically produced as a suggestion along with a generated rationale. We consider this to be a *human-in-the-loop* approach: one or more LLMs provide their relevance judgment and humans verify them. In most cases, the humans might not have to intervene at all but they might still be required in challenging situations where the LLM has low confidence.

An idea could also follow the 'preference testing' paradigm [84]: two LLMs each generate a judgment, and a human will select the better one—intervening only in case of disagreement between the LLMs. Still, in the scenario of human verification, humans make the ultimate decision wherever needed. A concern then could be that some bias of the LLMs might affect the final relevance judgments, as humans might not be able to recognize all biases. Related questions that we wish to raise within the community are: What sub-tasks of the judgment process require human input (e.g., prompt engineering [78, 91]—for now) and for what tasks or judgments should human assessors *not* be replaced by machines?

Fully Automated. If LLMs were able to reliably assess relevance, they could completely replace humans in the judgment process. A fully automatic judgment system might be as good as humans in producing high-quality relevance judgments (for a specific corpus / domain) but automatic judgments might even surpass humans in terms of quality, which raises the follow-up issue of how to detect that. A question that our community ought to investigate thus is: How can and should humans be replaced entirely by LLMs in the judgment process? Indeed, one could go as far as asking whether generative LLMs can and should be used to create complete test collections by generating documents / passages, as well as queries / conversations and relevance judgments.

A central aspect to be investigated is where on this four-level human–machine collaboration spectrum we actually obtain the ideal relevance judgments at the best cost. At this point, humans perform tasks that humans are good at, while machines perform tasks that machines are good at—often referred to as competence partitioning [37, 46]: a task is assigned to either a human or a machine, depending on who is better suited. Note that in our current version of the spectrum, we still (optimistically) show *balanced* competence partitioning as part of 'AI assistance'.

4 OPEN ISSUES AND OPPORTUNITIES

In this section, we identify several issues that arise when LLMs are used during relevance judgment tasks. We discuss open questions, risks we foresee, as well as opportunities to move beyond the currently accepted retrieval evaluation paradigms.

4.1 LLM Judgment Cost and Quality

It is currently unclear what the benefits and risks of LLMs for relevance judgments are. This situation is similar to the time when crowdsourced judgments became possible. Until about 10–15 years ago, judgments typically came from (trained) in-house experts but then suddenly could be delegated to cheaper crowdworkers resulting in an increased amount of financially feasible judgments—but at a substantially decreased quality [45] so that quality-assurance methods had to be developed [27]. With LLMs, history may somewhat repeat itself. Based on current pricing models, the inference costs per LLM judgment can be much lower than for crowdsourcing (cf. the estimates in the column 'Cost' of Table 2) so that again an increase in the amount of financially feasible judgments (from LLMs) is very likely. Still, the effect with respect to judgment quality is unclear—even improvements are possible—and can only be clarified / controlled by conducting respective studies and developing LLM-specific quality estimation and assurance methods.

The pressing question is: What is the effectiveness of LLM-based judgment (support)? In Table 2, we depict our current understanding by distinguishing four assessor types (user, expert, crowdworker, and LLM) and four judgment tasks: preference (which of two documents is more relevant?), binary (is this document relevant?), graded (how relevant is this document?), and explained (justify a judgment). The table entries indicate potential substitutions in the sense that similar abilities of LLMs hint at a replaceability of respective assessors (e.g., LLMs instead of crowdworkers for binary judgments). Still, the table cannot fully clarify the role of LLMs as we are still in the early stages of development and simply do not know the eventual capabilities: $\oplus$ and $\odot$ in the 'LLM' row should thus be interpreted with these current uncertainties in mind.

Table 2: Abilities of different types of assessors to handle various types of judgments. Similar levels of ability might hint at scenarios where specific types of human assessors might be replaced by LLMs.

		Type of Judgment			
Type of Assessor	**Cost**	Preference	Binary	Graded	Explained
User	free	$\oplus$	$\oplus$	$\oplus$	$\odot$
Expert	expensive	$\oplus\oplus$	$\oplus\oplus$	$\oplus$	$\oplus$
Crowdworker	cheap	$\odot$	$\oplus$	$\oplus$	$\odot$
LLM	very cheap	$\oplus$	$\oplus$	$\odot$	$\oplus$

Legend: $\oplus\oplus$ can judge, $\oplus$ depends, $\odot$ unknown

To align their judgments with humans, LLMs could be fine-tuned by observing human relevance assessors or they might use an active learning strategy [73, 74, 89]. For instance, an LLM could start with mild suggestions to a human assessor on how relevant a document is and could then continuously learn from the actual judgments of the assessor to improve its own suggestions.

4.2 Human Verification

Using Multiple LLMs as Assessors. While hiring multiple human relevance assessors with different backgrounds usually is very easy and potentially occurring judgment disagreements are not unresolvable [35], many LLMs are trained on very similar web corpora which may yield highly correlated judgments of not yet known quality or bias. A possible solution to obtain less correlated LLMs is to train or fine-tune them on different data (e.g., subcorpora). Fine-tuning on different user types could even yield "personalized" models [53, 83, 88] that might enable automatic judgments according to specific user groups' perspectives on relevance.

Truthfulness & Misinformation. An important aspect of relevance judgments is factuality. For a question like "do lemons cure cancer?", some top-ranked document may indeed suggest lemons as a treatment for cancer. While topically matching, the content is unlikely to be factually correct and the document should therefore be judged as non-relevant. Trained human assessors may very well be able to determine the trustworthiness of a document and, at least to some extent, the truthfulness. But the ability of LLMs to do so is quite unclear and probably also depends on the characteristics of the training data that often are not disclosed. This raises at least two questions: Can we automatically assess the reliability and factuality of LLM-generated relevance judgments? Can we identify the textual training sources underlying an LLM's judgment and can we verify that they are represented accurately?

Going forward, it will also be vital to be able to distinguish human-generated from automatically generated sources, especially in contexts such as journalism where correctness is critical.

Bias. Bender et al. [12] highlight limitations of LLMs and identify bias as a severe risk. As LLMs are intrinsically biased [9, 52, 60], such bias may also be reflected in LLMs' relevance judgments. For example, an LLM might be prone to consider scientific documents as relevant, while documents written in informal language are perceived as less relevant. The IR community should focus on finding ways to evaluate LLMs in terms of judgment bias, i.e., to analyze to what extent the intrinsic bias actually affects evaluations using LLM-supported / LLM-based relevance judgments.

Faithful Reasoning. LLMs often generate text that contains inaccurate or false information (i.e., they confabulate or "hallucinate") and usually do so in an affirmative manner that makes it difficult for humans to even suspect errors. In response, the NLP community is exploring a new research direction called "faithful reasoning" [23]. This approach aims to generate text that is less opaque, also describing explicitly the step-by-step reasoning, or the "chain of thoughts" [61]. A similar idea of "reasoned" automatic relevance judgments might be an interesting IR research direction.

Explaining Relevance to LLMs. Judgment guidelines often provide a comprehensive overview of what constitutes a relevant document in what scenario—most famously, Google's search quality evaluator guidelines have more than 170 pages.[3] Still, it is open how such guidelines should be "translated" to prompt LLMs. In addition, relevance may go beyond topical relevance [71]. For instance, a certain style may be required or the desired information should allure users from certain communities or cultures with different belief systems. We do not yet know to what extent LLMs are capable of assessing such different variations of relevance so that human intervention might still play a central role when taking document aspects into account that may not yet be easily discernable by LLMs.

4.3 Fully Automated

LLM-based Evaluation of LLM-based Systems. In the fully automated scenario, a circulatory problem can arise: Why not use a good LLM-based relevance assessor as an actual approach to produce a ranking? However, in practical settings, we expect LLMs used for ranking to be much smaller (more cost effective, lower latency, etc.; e.g., via knowledge distillation) than LLMs used for judging. In addition, the judging LLMs may have additional information about relevant facts / questions / nuggets that a ranker does not know, and, as assessment latency might not be an issue, different (more complex) judging LLMs may even be combined in an ensemble.

Moving beyond Cranfield. Given limited time or monetary budgets, retrieval evaluations based on manual judgments are often only feasible due to "standard" simplifying assumptions. For example, document collections are assumed to be static, small sets of queries / topics are assumed to suffice, and a document's relevance is assumed to not change (definitely a simplification [72, 80]) and to be independent of other documents. If LLMs would produce reliable relevance judgments with little human verification effort, many of the simplifying assumptions could be relaxed. For example, in search sessions or in the TREC CAsT track [24, 25],[4] information needs are changing over the course of a session or a conversation as the user learns more about a topic. Collaborative human–machine relevance judgment might help to scale-up evaluations using such more comprehensive and thus more realistic notions of relevance.

Moving beyond Human. Finally, at one end of our proposed spectrum, machines may surpass humans in the relevance judgment task. This phenomenon has already been witnessed in a variety of NLP tasks, such as scientific abstract classification [44] or sentiment detection [82]. Humans are likely to make mistakes when judging relevance and are limited by time. It is conceivable that LLMs with sufficient monetary funds will be capable of providing a larger number of more consistent judgments. However, if we use human-annotated data as a gold standard, we will not be able to detect when LLMs surpass human judgment quality as we then will have reached the limit of measurement.

5 PRELIMINARY ASSESSMENT

To provide a preliminary assessment of today's LLMs' capability for relevance judgments, we conduct an empirical comparison between human and LLM assessors. This comparison includes two LLMs (GPT-3.5 and YouChat), two test collections (the TREC-8 ad hoc retrieval task [81] and the TREC 2021 Deep Learning track [22]), two types of judgments (binary and graded), and two tailored prompts. The experiments were conducted in January and February 2023.

5.1 Methodology

Our experiments are not meant to be exhaustive but rather to explore where LLMs (dis-)agree with manual relevance judgments.

LLMs. We selected two LLMs for our experiments: GPT-3.5, more specifically `text-davinci-003`[5] accessed via OpenAI's API,[6] and YouChat. GPT-3.5 is an established standard model for many applications and thus serves as a natural baseline, while, shortly after OpenAI's release of ChatGPT, YouChat has been one of the first LLMs to be fully integrated with a commercial search engine[7] for the task of generating a new kind of search engine result page (SERP) on which a generated text summarizes the top-k search results ($k \lesssim 5$) in a query-biased way with numbered in-text references to k results listed as "blue links" below the summary.

Test Collections. We base our experiments on (i) the ad hoc retrieval task of TREC-8 [81] and (ii) the passage retrieval task of the TREC 2021 Deep Learning track (TREC-DL 2021) [22]. Both collections have many relevance judgments but also have contrasting properties. TREC-DL 2021 comprises short documents and queries phrased as questions, while TREC-8 comprises much longer, complete documents, with detailed descriptions of information needs, explicitly stating what is (not) considered relevant. As an experimental corpus, TREC-DL 2021 provides the additional benefit that its release date (second half of 2021) falls after the time that training data was crawled for GPT-3.5 (up to June 2021) but falls before the release of GPT-3.5 itself (November 2022).[8] Hence, GPT-3.5 has not been trained on TREC-DL 2021 relevance judgments, nor has it been used as a component in any system participating in TREC-DL 2021.

[3]https://guidelines.raterhub.com/searchqualityevaluatorguidelines.pdf
[4]TREC CAsT is a shared task that aims at evaluating conversational agents and thus provides information needs in the form of multi-turn conversations, each containing several utterances that a user might pose to an agent.

[5]https://spiresdigital.com/new-gpt-3-model-text-davinci-003
[6]https://platform.openai.com/docs/api-reference/introduction
[7]https://you.com.
[8]https://platform.openai.com/docs/models/overview

```
Instruction: You are an expert assessor making TREC relevance judgments. You will be given a TREC topic and a portion of a document. If any part of the document is relevant to the topic, answer "Yes". If not, answer "No". Remember that the TREC relevance condition states that a document is relevant to a topic if it contains information that is helpful in satisfying the user's information need described by the topic. A document is judged relevant if it contains information that is on-topic and of potential value to the user.
Topic: {topic}
Document: {document}
Relevant?
```

```
Instruction: Indicate if the passage is relevant for the question.
Question: {question}
Passage: {passage}
```

Figure 2: Prompts used in our experiments on TREC-8 (top) and TREC-DL 2021 (bottom). At the placeholders `{topic}`, `{document}`, `{question}`, and `{passage}`, the actually sampled pairs are included.

Judgment Sampling. We sampled $n = 1000$ topic–document pairs each from the relevance judgments files of TREC-8 and TREC-DL 2021 but due to a limited scalability when using YouChat, for some experiments had to restrict ourselves to 100 random samples per relevance grade (binary for TREC-8, graded for TREC-DL 2021).

Prompts. We used two simple and straightforward prompts for the two collections (cf. Figure 2) but explicitly did not spend time on optimizing the prompts (so-called "prompt engineering") to keep the prompts straightforward and to the point as a first baseline. Formulating and studying better prompts is left for future work.

Answer Parsing. We recorded the models' generated answers and mapped them to binary relevance judgments. As for GPT-3.5, the prompts and setting *temperature* = 0 were sufficient to constrain the model to emit only one of the requested relevance grades. As for YouChat, the answers were more verbose but rather homogeneous. With only two exceptions, they started with "The document / passage is *relevant* […]" or with "The document / passage is *not relevant* […]" and were thus straightforward to parse.

5.2 Results

Table 3 shows the results for TREC-8. We observe a clear divide according to the relevance label. For documents judged as non-relevant by human assessors, GPT-3.5 generates the same judgment in 90% of the cases. In contrast though, for the documents judged as relevant by human assessors, this agreement drops to 47%. Likewise, YouChat has judged 74% of the non-relevant documents correctly, but this agreement drops even more to 33% for the relevant ones.

Interestingly though, the results on TREC-DL 2021 in Table 4 show an opposite trend for YouChat: the higher the relevance grade, the more YouChat is in line with the human assessors. For 96 out of 100 question–passage pairs that TREC assessors judged as highly relevant (i.e., grade 3), YouChat agreed with the assessors. In contrast, for the non-relevant question–passage pairs, the agreement seems more or less random: YouChat only agrees with the manual

Table 3: Judgment agreement on TREC-8 between TREC assessors and the LLMs; 1000 topic–document pairs for GPT-3.5 and 100 for each grade (relevant, non-relevant) for YouChat.

LLM	Prediction	TREC-8 Assessors		Cohen's κ
		Relevant	Non-relevant	
GPT-3.5	Relevant	**237**	48	0.38
	Non-relevant	263	**452**	
YouChat	Relevant	**33**	26	
	Non-relevant	67	**74**	0.07

Table 4: Judgment agreement on TREC-DL 2021 between TREC assessors and the LLMs; 100 question–passage pairs for each grade from 3 (highly relevant) to 0 (non-relevant).

LLM	Prediction	TREC-DL 2021 Assessors				
		3	2	1	0	Cohen's κ
GPT-3.5	Relevant	**89**	**65**	**48**	16	0.40
	Non-relevant	11	35	52	**84**	
YouChat	Relevant	**96**	**93**	**79**	42	0.49
	Non-relevant	4	7	21	**58**	

assessments on 42 of the 100 pairs. Similarly, on TREC-DL 2021, GPT-3.5 seems to have problems with the middle grades of 1 and 2.

We thus hypothesize that human assessors may use subtle details to distinguish 'somewhat relevant' from 'probably non-relevant documents' in the binary case that are not captured by the LLMs and similarly that also the "differences" that human assessors use to decide some difficult 1-or-0 cases on a 3–0-scale might rather still be too subtle to be recognizable for the LLMs.

6 RE-JUDGING TREC 2021 DEEP LEARNING

To complement the experiments from Section 5, we now re-evaluate submissions to the passage ranking task of the TREC 2021 Deep Learning track [22] (TREC-DL 2021) using LLM-based judgments but adhering as closely as possible to the methodology used in the track itself [22], including the use of graded judgments.

6.1 Methodology

The participants of TREC-DL 2021 submitted 63 runs, each comprising up to 1000 ranked passages for 200 questions. These runs were pooled, and the results for 53 questions were judged by assessors using a combination of methods, including active learning [1, 76]. This generated a total of 10,828 judgments on a 4-point scale: 'perfectly relevant' > 'highly relevant' > 'relevant' (named 'related' in the track) > 'non-relevant' (named 'irrelevant' in the track).

We re-judged this pool using the GPT-3.5 `text-davinci-003` language model, as accessed through Open AI's API in February 2023. Consistent with a classification task—and with our GPT-3.5 experiments reported in Section 5—, we set *temperature* = 0 and otherwise use default parameters and settings.

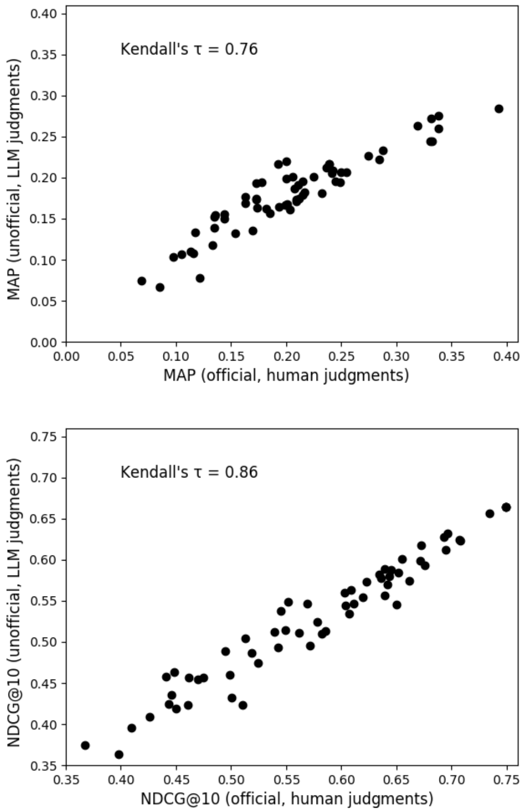

Figure 3: Scatter plots of the effectiveness of TREC-DL 2021 runs (MAP (top) and NDCG@10 (bottom)) according to the track's human judgments and our LLM-based judgments. A point represents a single run averaged over all questions.

Our relatively long prompt[9] is inspired by a prompt of Ferraretto et al. [36]: importantly—and different from the prompt in Figure 2—it leverages few-shot learning by listing multiple *examples* illustrating different levels of relevance for different questions. We provide one example each for 'perfectly relevant', 'highly relevant', and 'relevant', and we provide two examples for 'non-relevant', with one providing a judged 'non-relevant' passage, and the other providing an unrelated passage from the pool. These examples were chosen arbitrarily from the pool, based on the TREC judgments. We also used the term 'relevant' in the prompt, instead of 'related', since 'related' is a non-standard label for relevance judgments; in preliminary experiments, the LLM would sometimes return 'relevant' unprompted. Using this prompt, each judgment did cost about USD 0.01—we spent a total of USD 111.90, including a small number of duplicate requests due to failures and other issues. In comparison, Clarke et al. [19] report spending USD 0.25 per human judgment on a task of similar scope—with a single-page "prompt" and no training of assessors.

Table 5: Confusion matrices comparing all official TREC question–passage judgments with GPT-3.5 judgments on TREC-DL 2021 question–passage pairs. The upper matrix (GRADED) compares judgments on all four relevance levels. The lower matrix (BIN.) collapses the relevance labels to two levels, following the TREC-DL 2021 convention for computing binary measures.

	Prediction	**TREC-DL 2021 Assessors**			
		Perf. rel.	High. rel.	Related	Irrel.
GRADED	Perfectly relevant	**250**	248	177	87
	Highly relevant	360	**575**	628	370
	Relevant	328	880	**798**	442
	Non-relevant	148	638	1460	**3439**

	Prediction	**TREC-DL 2021 Assessors**	
		Relevant	Not relevant
BIN.	Relevant	**1433**	1262
	Non-relevant	1994	**6139**

6.2 Results

Table 5 shows the "agreement" on the full 4-point relevance scale and on a binarized relevance scale—following the TREC-DL 2021 convention, we map 'perfectly relevant' and 'highly relevant' to 'relevant', and 'relevant' and 'non-relevant' to 'non-relevant'. On the binarized judgments the Cohen's κ is 0.26, which indicates a 'fair' level of agreement. Note that on a similar experiment with two types of human judgments, Cormack et al. [21] report a Cohen's κ of 0.52 ('moderate' agreement).

Compared to the system rankings using the official judgments, using the LLM judgments to compute standard evaluation measures for the runs submitted to TREC-DL 2021 yields the correlations and Kendall's τ values shown in Figure 3. Note that the top run under the official judgments remains the top run under the LLM judgments. For comparison, Voorhees [80] report a Kendall's $\tau = .90$ for MAP on a similar experiment with two types of human judgments.

We find that measures computed under the LLM judgments are less sensitive than measures computed under human judgments. Sensitivity (or "discriminative power") measures the ability of an evaluation method to recognize a significant difference between retrieval approaches [19, 33, 67, 86]. To compute sensitivity, we take all pairs of submitted runs and compute a paired t-test between them. Here, we consider a pair with $p < 0.05$ as *distinguished* [86] and define sensitivity as $\frac{\textit{\# of distinguished pairs}}{\textit{total pairs}}$. Since we do not correct for the multiple comparisons problem, some of the distinguished pairs may not represent actual significant differences. With human judgments, 72% of the pairs are distinguished under MAP (74% under NDCG@10). In contrast, with GPT-3.5 judgments, only 65% are distinguished under MAP (69% under NDCG@10).

[9] Available at: https://plg.uwaterloo.ca/~claclark/trec2021_DL_prompt.txt

7 PERSPECTIVES FOR THE FUTURE

As this is a perspectives paper, we provide two opposing perspectives on the use of LLMs for automatic relevance judgments—for and against—and a third compromise perspective.

7.1 In Favor of Using LLMs for Judgments

In addition to providing a judgment of relevance, LLMs are able to produce a natural language explanation *why* a certain document is relevant or not to a topic [36]. Such AI-generated explanations may be used to assist human assessors in relevance judgments, particularly non-experts like crowdworkers. This setup may lead to better quality judgments as compared to the unsupported crowd. While LLM-generated labels and explanations may lead to an overreliance of human assessors, human assessors may serve as a quality control mechanism for the LLM. Furthermore, they serve as a feedback loop for the LLM to continuously improve its judgments. Our pilot experiments demonstrate that it is feasible for LLMs to indicate when a document is likely not relevant. We might therefore let human annotators assess (a) first those documents that are deemed relevant by LLMs, or (b) a subsample of documents from those considered relevant by the LLM, as an LLM can be run at scale. Thereby, we envision the use of LLMs to reduce annotation cost/time when creating high-quality IR evaluation collections.

It is noteworthy that LLMs may be better at providing fair and consistent judgments than humans. They can judge the relevance of documents without being affected by documents they have seen before, and with no boredom or tiredness effects. They are likely to assess conceptually similar documents the same way. Furthermore, they will often have seen much more information on a specific topic than most humans. Another advantage of today's LLMs is their inherent ability to process and generate text in many different languages. For multilingual corpora (which often appear in industrial settings) the assessment is typically restricted to a small subset of languages due to the limited availability of assessors. With LLMs being part of the assessment tool, this limitation no longer applies. LLMs are not just restricted to one input modality and thus conducting assessments that require the simultaneous consideration of multiple pieces of content (e.g. judging a web page based on the text but also the document's structure, visual cues, embedded video material, etc.) at the same time becomes possible. Finally, we note the cost factor—if we are able to judge hundreds of thousands of documents for a relatively small price, we can build much larger and much more complex test collections with regularly updated relevance assessments, in particular in domains that today lack meaningful test collections.

In summary, LLMs can provide explanations, scalability, consistency, and a certain level of quality when performing relevance judgments, underlining the great potential of deploying them as a complement to human assessors in certain judgments task.

7.2 Against Using LLMs for Judgments

While we have given several reasons to believe that we are close to using LLMs for automatic relevance judgment, there are also several concerns that should be addressed by the research community before deploying full-fledged automatic judgment. The primary concern is that LLMs are not people. IR measures of effectiveness are ultimately grounded in a human user's relevance judgment. Relevance is subjective, and changes over time for the same person [64]. Even if LLMs are increasingly good at mimicking human language in evaluating contents, it is a big leap of faith to fully trust the model's ability to make correct assessments without human verification. Currently, there is no proof that the judgments made by LLMs are grounded in reality. This raises an essential question: *If the output from an LLM is indistinguishable from a human-made relevance judgment, is this just a distinction without a difference?* After all, people disagree on relevance and change their opinions over time due to implicit and explicit learning effects. Usually, however, those disagreements do not have an effect on the evaluation unless there are systematic causes [8, 80]. To safely adopt LLMs to replace human annotators, the community should examine whether LLM-based relevance judgments may in fact be systematically different from those of real users. Not only do we know this affects the evaluation, but the complexity (or black-box nature) of the model precludes defining systematic bias in any useful way. There is a general concern about solely evaluating IR research with relevance assessment: Information retrieval systems are not just result-ranking machines, but are a system that is to assist a human to obtain information. Hence, only the user who consumes the results could tell which ones are useful. Another concern of applying LLMs as relevance annotators regards the "circularity" of the evaluation. Assume we are able to devise an annotation model based on LLMs. The same model could ideally also be used to retrieve and rank documents based on their expected relevance. If the model is used to judge relevance both for annotation and for retrieval, its evaluation would be overinflated, possibly with perfect performance. Vice-versa, models based on widely different rationales (such as BM25 or classical lexical approaches), might be penalized, because of how they estimate document relevance. As counter-considerations, we might hypothesize that the model used to label documents for relevance (a) is highly computationally expensive, making it almost unfeasible to use it as a retrieval system, and/or (b) has access to more information and facts than the retrieval model. The former holds as long as we do not use the automatic annotator as an expensive re-ranker capable of dealing with just a few documents. The latter, on the other hand, does not solve the problem of the automatic annotation, but simply shifts the problem: Either, the additional facts and information need to be annotated manually; then the human annotator remains essential. Or, the facts can be collected automatically; then we may assume that also a retrieval system could obtain them.

Other concerns arise if we even consider generative models as a replacement for traditional IR and search. In a plain old search engine, results for a query are ranked according to predicted relevance (ignoring sponsored results and advertising here). Each has a clear source, and each can be inspected directly as an entity separate from the search engine. Moreover, users frequently reformulate queries and try suggestions from the search engine, in a virtuous cycle wherein the users fulfill or adjust their conceptual information needs. Currently, hardly any of these is possible using LLM-generated responses: The results often are not attributed, rarely can be explored or probed, and are often completely generated. Also, best approaches for prompt engineering are not sufficiently studied, and their effect is more opaque than approaches to query reformulation. LLMs will not be usable for many information needs until they can attribute sources reliably and can be interrogated systematically. Will become available soon.

Finally, there are significant socio-technical concerns. Generative AI models can be used to generate fake photos and videos, for extortion purposes and misinformation. They are perceived as stealing the intellectual property. Furthermore, LLMs are affected by bias, stereotypical associations [9, 60], and adverse sentiments towards specific groups [52]. Critically, we cannot assess whether the LLM may have seen information that biases the relevance judgment in an unwanted way, let alone that the company owning the LLM may change it anytime without our knowledge or control. As a result, we ourselves as the authors of this perspectives paper disagree on whether, as a profession and considering the ACM's Code of Ethics, we should use generative models in deployed systems *at all* until these issues are worked out.

7.3 A Compromise: Double-checking LLMs and Human–Machine Collaboration

Our pilot study in Sections 5 and 6 finds a reasonable correlation between highly-trained human assessors and a fully automated LLM, yielding similar leaderboards. This suggests that the technology is promising and deserves further study. The experiment could be implemented to double-check LLM judgments: produce fully automated as well as human judgments on a shared judgment pool, then analyze correlations of labels and system rankings, then decide whether LLM's relevance judgments are good enough to be shared as an alternative test collection with the community. The automatic judgment paradigm should be revealed along with prompts, hyperparameters, and details for reproducibility. We also suggest to declare which judgment paradigm was chosen when releasing data resources (such as in TREC CAR). At the very least, such automatic judgments could be used to evaluate early prototypes of approaches, for initial judgments for novel tasks, and for large-scale training.

While the discussion is easily dominated by the fully automated evaluation—this is merely an extreme point on our spectrum in Section 3. The majority of authors do not believe this constitutes the best path towards credible IR research. For example, "AI Assistance" is probably the most credible path for LLMs to be incorporated during evaluation. However, it is also the least explored so far.

This calls for more research on innovative ways to use LLMs for assistance during the judgment process and how to leverage humans for verifying the LLMs' suggestions. As a community, we should explore how the performance of human assessors changes, when they are shown rationales or chain-of-thoughts that are generated by LLMs. Human assessors often struggle to see a pertinent connection when they are lacking world knowledge. An example of this issue is the task of assessing the relevance of "diabetes" for the topic "child trafficking". LLMs can generate rationales that can explain such connections. However, it requires a human to realize when such a rationale was hallucinated. Only a human can assess whether the information provided appears true and reliable.

8 CONCLUSION

In this paper, we investigated the opportunity that large language models (LLMs) now may generate relevance judgments automatically. We discussed previous attempts to automate and scale-up the relevance judgment task, and we presented experimental results showing promise in the ability to mimic human relevance assessments with LLMs. Our findings suggest that, while the path is promising and worthy of being investigated, at the time of writing several reasons prevent LLMs from being employed as fully automated annotation tools. Nevertheless, there is a spectrum of solutions to employ LLMs as support for human assessors in a human–machine collaboration. Therefore, we present our perspectives on why and why not the IR community should employ LLMs in some way in the evaluation process. Undoubtedly, more research on LLMs for relevance judgment is to be carried out in the future, for which this paper provides a starting point.

ACKNOWLEDGMENTS

This paper is based on discussions during a breakout group at the Dagstuhl Seminar 23031 on "Frontiers of Information Access Experimentation for Research and Education" [10]. We express our gratitude to the Seminar organizers, Christine Bauer, Ben Carterette, Nicola Ferro, and Norbert Fuhr.

Certain companies and software are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service, nor is it intended to imply that the software or companies identified are necessarily the best available for the purpose.

This material is based upon work supported by the National Science Foundation under Grant No. 1846017. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. 2018. A System for Efficient High-Recall Retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018.* ACM, 1317–1320. https://doi.org/10.1145/3209978.3210176

[2] Omar Alonso and Ricardo Baeza-Yates. 2011. Design and Implementation of Relevance Assessments Using Crowdsourcing. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings (Lecture Notes in Computer Science, Vol. 6611).* Springer, 153–164. https://doi.org/10.1007/978-3-642-20161-5_16

[3] Omar Alonso and Stefano Mizzaro. 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] Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. 2021. BERT-QPP: Contextualized Pre-trained Transformers for Query Performance Prediction. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021.* ACM, 2857–2861. https://doi.org/10.1145/3459637.3482063

[5] Negar Arabzadeh, Mahsa Seifikar, and Charles L. A. Clarke. 2022. Unsupervised Question Clarity Prediction through Retrieved Item Coherency. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022.* ACM, 3811–3816. https://doi.org/10.1145/3511808.3557719

[6] Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. *Trans. Assoc. Comput. Linguistics* 7 (2019), 169–184. https://doi.org/10.1162/tacl_a_00261

[7] Nima Asadi, Donald Metzler, Tamer Elsayed, and Jimmy Lin. 2011. Pseudo Test Collections for Learning Web Search Ranking Functions. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*. ACM, 1073–1082. https://doi.org/10.1145/2009916.2010058

[8] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does it Matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*. ACM, 667–674. https://doi.org/10.1145/1390334.1390447

[9] Christine Basta, Marta Ruiz Costa-jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. *CoRR* abs/1904.08783 (2019). arXiv:1904.08783 http://arxiv.org/abs/1904.08783

[10] Christine Bauer, Ben Carterette, Nicola Ferro, and Norbert Fuhr. 2023. Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education. *CoRR* abs/2305.01509 (2023). https://doi.org/10.48550/arXiv.2305.01509 arXiv:2305.01509

[11] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, and David A. Grossman. 2003. Using Titles and Category Names from Editor-Driven Taxonomies for Automatic Evaluation. In *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, 2003*. ACM, 17–23. https://doi.org/10.1145/956863.956868

[12] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*. ACM, 610–623. https://doi.org/10.1145/3442188.3445922

[13] Richard Berendsen, Manos Tsagkias, Maarten de Rijke, and Edgar Meij. 2012. Generating Pseudo Test Collections for Learning to Rank Scientific Articles. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics - Third International Conference of the CLEF Initiative, CLEF 2012, Rome, Italy, September 17-20, 2012. Proceedings (Lecture Notes in Computer Science, Vol. 7488)*. Springer, 42–53. https://doi.org/10.1007/978-3-642-33247-0_6

[14] Roi Blanco, Harry Halpin, Daniel M. Herzig, Peter Mika, Jeffrey Pound, Henry S. Thompson, and Duc Thanh Tran. 2011. Repeatable and Reliable Search System Evaluation Using Crowdsourcing. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*. ACM, 923–932. https://doi.org/10.1145/2009916.2010039

[15] Martin Braschler. 2000. CLEF 2000 - Overview of Results. In *Cross-Language Information Retrieval and Evaluation, Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers (Lecture Notes in Computer Science, Vol. 2069)*. Springer, 89–101. https://doi.org/10.1007/3-540-44645-1_9

[16] David Carmel and Elad Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval.* Morgan & Claypool Publishers. https://doi.org/10.2200/S00235ED1V01Y201004ICR015

[17] Ben Carterette, James Allan, and Ramesh K. Sitaraman. 2006. Minimal Test Collections for Retrieval Evaluation. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*. ACM, 268–275. https://doi.org/10.1145/1148170.1148219

[18] Xiaoyang Chen, Ben He, and Le Sun. 2022. Groupwise Query Performance Prediction with BERT. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13186)*. Springer, 64–74. https://doi.org/10.1007/978-3-030-99739-7_8

[19] Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. 2020. Assessing Top-k Preferences. *CoRR* abs/2007.11682 (2020). arXiv:2007.11682 https://arxiv.org/abs/2007.11682

[20] Cyril W Cleverdon. 1960. The Aslib Cranfield Research Project on the Comparative Efficiency of Indexing Systems. In *Aslib Proceedings*, Vol. 12. MCB UP Ltd, 421–431.

[21] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient Construction of Large Test Collections. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*. ACM, 282–289. https://doi.org/10.1145/290941.291009

[22] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 Deep Learning Track. *CoRR* abs/2102.07662. arXiv:2102.07662 https://arxiv.org/abs/2102.07662

[23] Antonia Creswell and Murray Shanahan. 2022. Faithful Reasoning Using Large Language Models. *CoRR* abs/2208.14271 (2022). https://doi.org/10.48550/arXiv.2208.14271 arXiv:2208.14271

[24] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview. *CoRR* abs/2003.13624 (2020). arXiv:2003.13624 https://arxiv.org/abs/2003.13624

[25] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview. *CoRR* abs/2003.13624. arXiv:2003.13624 https://arxiv.org/abs/2003.13624

[26] Bhavana Bharat Dalvi, Einat Minkov, Partha Pratim Talukdar, and William W. Cohen. 2015. Automatic Gloss Finding for a Knowledge Base using Ontological Constraints. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*. ACM, 369–378. https://doi.org/10.1145/2684822.2685288

[27] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.* 51, 1 (2018), 7:1–7:40. https://doi.org/10.1145/3148148

[28] Suchana Datta, Sean MacAvaney, Debasis Ganguly, and Derek Greene. 2022. A 'Pointwise-Query, Listwise-Document' Based Query Performance Prediction Approach. In *Proceedings of 45th international ACM SIGIR conference research development in information retrieval*. 2148—-2153. https://doi.org/10.1145/3477495.3531821

[29] Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. *Transactions of the Association for Computational Linguistics* 9 (2021), 774–789. https://doi.org/10.1162/tacl_a_00397

[30] Laura Dietz, Shubham Chatterjee, Connor Lennox, Sumanta Kashyapi, Pooja Oza, and Ben Gamari. 2022. Wikimarks: Harvesting Relevance Benchmarks from Wikipedia. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. ACM, 3003–3012. https://doi.org/10.1145/3477495.3531731

[31] Laura Dietz and Jeff Dalton. 2020. Humans Optional? Automatic Large-Scale Test Collections for Entity, Passage, and Entity-Passage Retrieval. *Datenbank-Spektrum* 20, 1 (2020), 17–28. https://doi.org/10.1007/s13222-020-00334-y

[32] Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question Answering as an Automatic Evaluation Metric for News Article Summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 3938–3948. https://doi.org/10.18653/v1/n19-1395

[33] Guglielmo Faggioli and Nicola Ferro. 2021. System Effect Estimation by Sharding: A Comparison Between ANOVA Approaches to Detect Significant Differences. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12657)*. Springer, 33–46. https://doi.org/10.1007/978-3-030-72240-1_3

[34] Guglielmo Faggioli, Nicola Ferro, Cristina Muntean, Raffaele Perego, and Nicola Tonellotto. 2023. A Geometric Framework for Query Performance Prediction in Conversational Search. In *Proceedings of 46th international ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2023 July 23–27, 2023, Taipei, Taiwan*. ACM. https://doi.org/10.1145/3539618.3591625

[35] Marco Ferrante, Nicola Ferro, and Maria Maistro. 2017. AWARE: Exploiting Evaluation Measures to Combine Multiple Assessors. *ACM Transactions on Information Systems* 36, 2 (2017), 20:1–20:38. https://doi.org/10.1145/3110217

[36] Fernando Ferraretto, Thiago Laitz, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2023. ExaRanker: Explanation-Augmented Neural Ranker. https://doi.org/10.48550/arXiv.2301.10521 arXiv:2301.10521

[37] Frank Flemisch, David Abbink, Makoto Itoh, Marie-Pierre Pacaux-Lemoine, and Gina Weßel. 2016. Shared Control is the Sharp End of Cooperation: Towards a Common Framework of Joint Action, Shared Control and Human Machine Cooperation. *IFAC-PapersOnLine* 49, 19 (2016), 72–77. https://doi.org/10.1016/j.ifacol.2016.10.464 13th IFAC Symposium on Analysis, Design, and Evaluation ofHuman-Machine Systems HMS 2016.

[38] Maik Fröbe, Christopher Akiki, Martin Potthast, and Matthias Hagen. 2022. Noise-Reduction for Automatically Transferred Relevance Judgments. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings (Lecture Notes in Computer Science, Vol. 13390)*. Springer, 48–61. https://doi.org/10.1007/978-3-031-13643-6_4

[39] Maik Fröbe, Lukas Gienapp, Martin Potthast, and Matthias Hagen. 2023. Bootstrapped nDCG Estimation in the Presence of Unjudged Documents. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13980)*. Springer, 313–329. https://doi.org/10.1007/978-3-031-28244-7_20

[40] Norbert Fuhr. 2017. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51, 3, 32–41. https://doi.org/10.1145/3190580.3190586

[41] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2019. Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work* 28, 5 (2019), 815–841. https://doi.org/10.1007/s10606-018-9336-y

[42] Debasis Ganguly, Surupendu Gangopadhyay, Mandar Mitra, and Prasenjit Majumder (Eds.). 2022. *FIRE '22: Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation* (Kolkata, India). Association for Computing Machinery, New York, NY, USA.

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

[44] Yeow Chong Goh, Xin Qing Cai, Walter Theseira, Giovanni Ko, and Khiam Aik Khor. 2020. Evaluating Human Versus Machine Learning Performance in Classifying Research Abstracts. *Scientometrics* 125, 2 (2020), 1197–1212. https://doi.org/10.1007/s11192-020-03614-2

[45] Martin Halvey, Robert Villa, and Paul D. Clough. 2015. SIGIR 2014: Workshop on Gathering Efficient Assessments of Relevance (GEAR). *SIGIR Forum* 49, 1 (2015), 16–19. https://doi.org/10.1145/2795403.2795409

[46] PA Hancock. 2013. Task partitioning effects in semi-automated human–machine system performance. *Ergonomics* 56, 9 (2013), 1387–1399. https://doi.org/10.1080/00140139.2013.816374

[47] Donna Harman. 1992. *Overview of the First Text REtrieval Conference (TREC-1)*. NIST Special Publication, Vol. 500-207. National Institute of Standards and Technology (NIST). 1–20 pages. http://trec.nist.gov/pubs/trec1/papers/01.txt

[48] Claudia Hauff. 2010. Predicting the effectiveness of queries and retrieval systems. *SIGIR Forum* 44, 1 (2010), 88. https://doi.org/10.1145/1842890.1842906

[49] Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. *Transactions of the Association for Computational Linguistics* 9 (2021), 211–225. https://doi.org/10.1162/tacl_a_00362

[50] Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A Novel Large-scale Language Understanding Task over Wikipedia. (2016). https://doi.org/10.18653/v1/p16-1145

[51] Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, 5094–5107. https://doi.org/10.18653/v1/2020.acl-main.457

[52] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

[53] Aaron Jaech and Mari Ostendorf. 2018. Personalized Language Model for Query Auto-Completion. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*. Association for Computational Linguistics, 700–705. https://doi.org/10.18653/v1/P18-2111

[54] Kalervo Järvelin. 2009. Explaining User Performance in Information Retrieval: Challenges to IR Evaluation. In *Advances in Information Retrieval Theory, Second International Conference on the Theory of Information Retrieval, ICTIR 2009, Cambridge, UK, September 10-12, 2009, Proceedings (Lecture Notes in Computer Science, Vol. 5766)*. Springer, 289–296. https://doi.org/10.1007/978-3-642-04417-5_28

[55] Gaya K. Jayasinghe, William Webber, Mark Sanderson, and J. Shane Culpepper. 2014. Improving Test Collection Pools with Machine Learning. In *Proceedings of the 2014 Australasian Document Computing Symposium, ADCS 2014, Melbourne, VIC, Australia, November 27-28, 2014*. ACM, 2. https://doi.org/10.1145/2682862.2682864

[56] Noriko Kando (Ed.). 1999. *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*. National Center for Science Information Systems (NACSIS). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings/

[57] Gjergji Kasneci, Maya Ramanath, Fabian M. Suchanek, and Gerhard Weikum. 2008. The YAGO-NAGA Approach to Knowledge Discovery. *ACM SIGMOD Record* 37, 4 (2008), 41–47. https://doi.org/10.1145/1519103.1519110

[58] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2013. An Analysis of Human Factors and Label Accuracy in Crowdsourcing Relevance Judgments. *Information Retrieval* 16, 2 (2013), 138–178. https://doi.org/10.1007/s10791-012-9205-0

[59] Mostafa Keikha, Jae Hyun Park, and W. Bruce Croft. 2014. Evaluating Answer Passages Using Summarization Measures. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*. ACM, 963–966. https://doi.org/10.1145/2600428.2609485

[60] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. *CoRR* abs/1906.07337 (2019). arXiv:1906.07337 http://arxiv.org/abs/1906.07337

[61] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful Chain-of-Thought Reasoning. *CoRR* abs/2301.13379 (2023). https://doi.org/10.48550/arXiv.2301.13379 arXiv:2301.13379

[62] Sean MacAvaney and Luca Soldaini. 2023. One-Shot Labeling for Automatic Relevance Estimation. *CoRR* abs/2302.11266 (2023). https://doi.org/10.48550/arXiv.2302.11266 arXiv:2302.11266

[63] Eddy Maddalena, Marco Basaldella, Dario De Nart, Dante Degl'Innocenti, Stefano Mizzaro, and Gianluca Demartini. 2016. Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA*. AAAI Press, 129–138. http://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14040

[64] Stefano Mizzaro. 1997. Relevance: The Whole History. *Journal of the American society for information science* 48, 9 (1997), 810–832. https://doi.org/10.1002/(SICI)1097-4571(199709)48:9%3C810::AID-ASI6%3E3.0.CO;2-U

[65] Zahra Nouri, Henning Wachsmuth, and Gregor Engels. 2020. Mining Crowdsourcing Problems from Discussion Forums of Workers. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6264–6276. https://doi.org/10.18653/v1/2020.coling-main.551

[66] Virgiliu Pavlu, Shahzad Rajput, Peter B. Golbus, and Javed A. Aslam. 2012. IR System Evaluation Using Nugget-Based Test Collections. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012*. ACM, 393–402. https://doi.org/10.1145/2124295.2124343

[67] Tetsuya Sakai. 2006. Evaluating Evaluation Metrics Based on the Bootstrap. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*. ACM, 525–532. https://doi.org/10.1145/1148170.1148261

[68] Tetsuya Sakai. 2020. On Fuhr's Guideline for IR Evaluation. *SIGIR Forum* 54, 1, 12:1–12:8. https://doi.org/10.1145/3451964.3451976

[69] David P. Sander and Laura Dietz. 2021. EXAM: How to Evaluate Retrieve-and-Generate Systems for Users Who Do Not (Yet) Know What They Want. In *Proceedings of the Second International Conference on Design of Experimental Search & Information REtrieval Systems, Padova, Italy, September 15-18, 2021 (CEUR Workshop Proceedings, Vol. 2950)*. CEUR-WS.org, 136–146. http://ceur-ws.org/Vol-2950/paper-16.pdf

[70] Tefko Saracevic. 1995. Evaluation of Evaluation in Information Retrieval. In *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995 (Special Issue of the SIGIR Forum)*. ACM Press, 138–146. https://doi.org/10.1145/215206.215351

[71] Tefko Saracevic. 1996. Relevance Reconsidered. In *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)*. 201–218.

[72] Linda Schamber. 1994. Relevance and Information Behavior. *Annual review of information science and technology (ARIST)* 29 (1994), 3–48.

[73] Seungmin Seo, Donghyun Kim, Youbin Ahn, and Kyong-Ho Lee. 2022. Active Learning on Pre-trained Language Model with Task-Independent Triplet Loss. In *Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 11276–11284. https://ojs.aaai.org/index.php/AAAI/article/view/21378

[74] Akanksha Rai Sharma and Pranav Kaushik. 2017. Literature Survey of Statistical, Deep and Reinforcement Learning in Natural Language Processing. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*. 350–354. https://doi.org/10.1109/CCAA.2017.8229841

[75] Aashish Sheshadri and Matthew Lease. 2013. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2013, November 7-9, 2013, Palm Springs, CA, USA*. AAAI. http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7550

[76] Ian Soboroff. 2021. Overview of TREC 2021. In *30th Text REtrieval Conference*. Gaithersburg, Maryland. https://trec.nist.gov/pubs/trec30/papers/Overview-2021.pdf

[77] Ian Soboroff, Charles K. Nicholas, and Patrick Cahan. 2001. Ranking Retrieval Systems without Relevance Judgments. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*. ACM, 66–73. https://doi.org/10.1145/383952.383961

[78] Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Association for Computational Linguistics, 819–862. https://doi.org/10.18653/v1/2022.acl-long.60

[79] Lynda Tamine and Cecile Chouquet. 2017. On the Impact of Domain Expertise on Query Formulation, Relevance Assessment and Retrieval Performance in Clinical Settings. *Information Processing & Management* 53, 2 (2017), 332–350. https://doi.org/10.1016/j.ipm.2016.11.004

[80] Ellen M. Voorhees. 2000. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Inf. Process. Manag.* 36, 5 (2000), 697–716. https://doi.org/10.1016/S0306-4573(00)00010-8

[81] Ellen M. Voorhees and Donna Harman. 1999. Overview of the Eighth Text REtrieval Conference (TREC-8). In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999 (NIST Special Publication, Vol. 500-246)*. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec8/papers/overview_8.ps

[82] Christian Weismayer, Ilona Pezenka, and Christopher Han-Kie Gan. 2018. Aspect-Based Sentiment Detection: Comparing Human Versus Automated Classifications of TripAdvisor Reviews. In *Information and Communication Technologies in Tourism 2018, ENTER 2018, Proceedings of the International Conference in Jönköping, Sweden, January 24-26, 2018*. Springer, 365–380. https://doi.org/10.1007/978-3-319-72923-7_28

[83] Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2022. Leveraging Similar Users for Personalized Language Modeling with Limited Data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Association for Computational Linguistics, 1742–1752. https://doi.org/10.18653/v1/2022.acl-long.122

[84] Jennifer Windsor, Laura M Piché, and Peggy A Locke. 1994. Preference Testing: A Comparison of Two Presentation Methods. *Research in developmental disabilities* 15, 6 (1994), 439–455. https://doi.org/10.1016/0891-4222(94)90028-0

[85] Jiechen Xu, Lei Han, Shazia Sadiq, and Gianluca Demartini. 2023. On The Role of Human and Machine Metadata in Relevance Judgment Tasks. *Information Processing & Management* 60, 2 (2023), 103177. https://doi.org/10.1016/j.ipm.2022.103177

[86] Ziying Yang, Alistair Moffat, and Andrew Turpin. 2018. Pairwise Crowd Judgments: Preference, Absolute, and Ratio. In *Proceedings of the 23rd Australasian Document Computing Symposium, ADCS 2018, Dunedin, New Zealand, December 11-12, 2018*. ACM, 3:1–3:8. https://doi.org/10.1145/3291992.3291995

[87] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. 2008. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*. ACM, 603–610. https://doi.org/10.1145/1390334.1390437

[88] Seunghyun Yoon, Hyeongu Yun, Yuna Kim, Gyu-tae Park, and Kyomin Jung. 2017. Efficient Transfer Learning Schemes for Personalized Language Modeling using Recurrent Neural Network. In *The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California, USA (AAAI Technical Report, Vol. WS-17)*. AAAI Press. http://aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15144

[89] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, Jae Sung Park, Ximing Lu, Prithviraj Ammanabrolu, Rowan Zellers, Ronan Le Bras, Gunhee Kim, and Yejin Choi. 2022. Multimodal Knowledge Alignment with Reinforcement Learning. *CoRR* abs/2205.12630 (2022). https://doi.org/10.48550/arXiv.2205.12630 arXiv:2205.12630

[90] 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

[91] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large Language Models Are Human-Level Prompt Engineers. *CoRR* abs/2211.01910 (2022). https://doi.org/10.48550/arXiv.2211.01910 arXiv:2211.01910

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