

LLM-Driven Usefulness Labeling for IR Evaluation

Mouly Dewan
University of Washington
Seattle, WA, United States
mdewan@uw.edu

Jiqun Liu
The University of Oklahoma
Norman, OK, United States
jiquunliu@ou.edu

Chirag Shah
University of Washington
Seattle, WA, United States
chirags@uw.edu

Abstract

In the information retrieval (IR) domain, evaluation plays a crucial role in optimizing search experiences and supporting diverse user intents. In the recent LLM era, research has been conducted to automate document relevance labels. These labels have traditionally been assigned by crowd-sourced workers, a process that is both time consuming and costly. This study focuses on LLM-generated *usefulness* labels, a crucial evaluation metric that considers the user's search intents and task objectives, an aspect where relevance falls short. Our experiment utilizes task-level, query-level, and document-level features along with user search behavior signals, which are essential in defining the usefulness of a document. Our research finds that (i) pre-trained LLMs can generate moderate usefulness labels by understanding the comprehensive search task session, and (ii) pre-trained LLMs perform better judgment in short search sessions when provided with search session contexts. Furthermore, we investigate whether LLMs can capture the unique divergence between relevance and usefulness, along with conducting an ablation study to identify the most critical metrics for accurate usefulness label generation. In conclusion, this work explores LLM-generated usefulness labels by evaluating critical metrics and optimizing for practicality in real-world settings.

CCS Concepts

- Information systems → Language models.

Keywords

Information Retrieval, Usefulness Judgment, LLM Evaluation

ACM Reference Format:

Mouly Dewan, Jiqun Liu, and Chirag Shah. 2025. LLM-Driven Usefulness Labeling for IR 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, 5 pages. <https://doi.org/10.1145/3726302.3730223>

1 Introduction

Assessing relevance is not enough for many IR scenarios, as in many cases it cannot fully capture the extent to which a document is actually *useful* for addressing the motivating tasks. Although many scholars have advocated this for decades [10, 19], mainstream IR has remained focused on the seemingly objective assessment of relevance, following the traditional Cranfield evaluation paradigm [11]. Relevance is often used as a proxy for other constructs due

to its strong correlation with user centric evaluation labels like usefulness and satisfaction [22]. However, usefulness is tied to both task context and users' context dependent preferences, making it a richer but more challenging construct to capture in real time. For web search evaluations, relevance based metrics alone cannot capture the user's search goal and satisfaction, especially during complex task driven sessions [4, 18]. Therefore, incorporating usefulness as an evaluation metric provides a more holistic assessment, capturing not only document relevance, but also user search intent and satisfaction throughout sessions [20, 24].

To mitigate time and cost constraints of traditional evaluation challenges [28], IR researchers have been exploring the possibility and techniques of automated search evaluation with LLMs [3, 7, 13]. LLMs for automated evaluation (LLM4eval) have recently attracted growing attention in IR community because of their flexibility and increasing task-solving capabilities [25]. In LLM automated evaluations, researchers have primarily focused on relevance labels, while user centric usefulness metrics remain largely unexplored. As usefulness is a measure that directly reflects a *document's contribution to achieving user goals*, in this paper we particularly focus on usefulness as the key metric for evaluating web search along with relevance. To achieve this, we generate usefulness labels using LLMs and employ search session, task-level, query-level and document-level user metrics, aiming to better capture how users perceive the quality of search systems. We also seek to explore whether LLMs can capture the diverging relationship between relevance and usefulness. In some cases, a retrieved document may be relevant to the query but not necessarily useful to the user, or it may be irrelevant to the query yet highly valuable for completing the task. Being able to identify and address such situations can open up new possibilities for enhancing IR applications, especially in complex, open ended tasks.

Existing ML-based approaches for predicting usefulness often overlook the broader task and user contexts. This is the place where LLMs can step in with their pre-trained existing knowledge to better understand user, task, and query context through appropriate prompting. The purpose of generating usefulness labels is to gauge how well and accurately these LLMs can perform in automated usefulness judgment when given all relevant features (both implicit and explicit) [8, 14, 21, 31]. Specifically, our research seeks to answer the following three research questions (RQs):

RQ1: To what extent can large language models (LLMs) accurately label document *usefulness* in search sessions?

RQ2: To what extent LLMs understand the comprehensive search goal of a user and navigate the divergence between relevance and usefulness of the retrieved documents?

RQ3: What are the most effective search metrics in improving LLMs' document usefulness judgments?



This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '25, July 13–18, 2025, Padua, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1592-1/2025/07
<https://doi.org/10.1145/3726302.3730223>

2 Related Works

In LLM enabled search engine evaluation, most of the research has solely focused on automating relevance labels [1, 13, 26, 29], where else user centric usefulness metrics have not been explored yet. Faggioli et al. [13] were among the first to examine how LLMs can be used for relevance judgment, offering unique insights into human-LLM collaboration. Their work explored direct prompting strategies, highlighting both the potential and limitations of using LLMs to generate relevance labels. Another prominent work in LLM automated relevance assessment by Thomas et al. [28] asserted that LLMs can accurately predict searcher’s preferences and perform relevance labeling as accurately as human labelers. Studies conclude that LLMs are more positive in assessing relevance and thus should be used to assess all documents, not just relevance gaps [1] and, to some extent LLMs are studied for utility judgment using retrieval-augmented generation (RAG) [27, 32]. To address some limitations of Thomas et al. [28], Upadhyay et al. [29] conducted large scale relevance assessment on DL TREC tracks using UMBRELLA [30], an LLM based relevance assessor for academic TREC-style evaluation. Most LLM automated IR evaluations have focused solely on relevance labeling, but relevance alone is insufficient, as usefulness directly reflects user goal achievement [4, 10]. In this work, we explore usefulness as a key metric for web search evaluation alongside relevance. We generate usefulness labels using LLMs, incorporating implicit user search behavior signals like CTR and dwell time [5, 9, 16]. Additionally, we examine whether LLMs can capture contextual relationships in search sessions by integrating relevance and satisfaction labels to assess search system quality while exploring the contrast between relevance and usefulness.

3 Methodology

Our main goal is to evaluate the extent to which LLMs can generate usefulness labels when given user search behavior signals along with task-level, query-level and document-level information. In real-world scenarios, when evaluation is conducted by crowd-sourced workers they perform individual search sessions for each task and assign usefulness labels for specific query and task within the context of the entire search session. In our experiment, we follow the same methodology where LLMs are provided vital information about each user’s search session and it generates a usefulness label for each task and for all the clicked documents.

We conduct the experiments using our zero-shot prompt¹ that has been created taking inspiration from the structure of the widely used DNA prompt template by Thomas et al. [28] that uses LLMs to generate relevance labels. We conduct experiments on 6 different types of pre-trained LLMs such as GPT-4omini, GPT-3.5turbo [2], Llama 3.1 8B-Instruct, Llama 3.2 3B-Instruct, Llama 3.3 70B-Instruct [12] and DeepSeek-R1-Distill-Qwen-14B [17] to evaluate the ability to generate usefulness labels. Additionally, we explore LLM’s ability to process the divergence cases of *relevance-usefulness* to capture the nuances of search session contexts. Finally, our research includes an ablation study of various search behavior metrics to identify the most effective ones for input into an LLM for optimizing accuracy while considering real-world constraints, such as token limitations, model sizes and associated costs.

¹<https://github.com/moulydewan/LLMUsefulnessLabels>

Table 1: Statistics of Behavior Logs

Dataset	#Tasks	#Participants	#Sessions	#Queries	#Clicks
THUIR-KDD’19	9	50	447	735	1,431
QRef	—	42	2,024	4,809	7,126

3.1 Datasets

For this experiment, we used two datasets: THUIR KDD’19 [23] and TianGong-QRef [6], with behavioral log statistics shown in Table 1. Both datasets include query satisfaction, session satisfaction, and usefulness labels while only THUIR-KDD’19 provides relevance labels from external assessors. They capture user search interactions (CLICK, SCROLL, HOVER, MOVE) and session timestamps. From these, we extracted CTR, URL dwell time, query dwell time, and task dwell time per user. From the statistics, we can see that THUIR-KDD’19 contains longer search sessions with 9 tasks, whereas QRef features shorter task-free sessions. For this reason, we perform task-level and query-level analysis for THUIR-KDD’19 and for QRef we perform session-level and query-level analysis (further shown in Table 2). The variety in these datasets provided a diverse foundation for our study, allowing us to explore real-world user-centric search behaviors across both structured task-driven sessions and more open ended task-free search interactions.

3.2 Baseline Data and Session Data

After curating all the additional metrics, we create our LLM dataset by treating each data point as an individual batch regardless of sessions and call them baseline data. Each data point represents a single user click on a web document with the associated metrics irrespective of user sessions. All corresponding queries, along with the metrics for each clicked document within a user’s task, are grouped into a single batch, which we refer to as session data. The reason for aggregating all these information together is that in practical scenario, the overall session history matters when labeling usefulness for each clicked documents. For a particular task, users typically either find the required information from the first query and the corresponding clicked documents, successfully completing their search, or they need to reformulate their query to refine the search results and then locate the desired information. If a user finds document X useful, a similar relevant document Y seen later may feel redundant and not useful. Yet without session context, the LLM might still rate both X and Y as equally useful. For this reason, to preserve the contextual information of a search session for a task, we formulate one batch as one task session for one user and provide it to the LLMs as a part of session data. Formulating each search session for each user in Table 1, we get around 447 sessions from THUIR-KDD’19 dataset and 2,024 sessions from QRef dataset.

4 Experimental Setup

We conduct our primary experiment on two datasets: (1) baseline data without session context and (2) session data with previous search context. Both include document-level metrics (e.g., URL, title, summary, SERP snippets, rank) and behavioral metrics (e.g., CTR, URL-query-task dwell time), along with relevance and satisfaction labels for a comprehensive view of user interactions. Using our zero-shot prompting, we asked LLMs to label clicked document

Table 2: Comparison of LLM performances on THUIR-KDD'19 and QRef datasets using Spearman's Rank (ρ).

Model	THUIR-KDD'19 Dataset						QRef Dataset					
	Baseline			Session			Baseline			Session		
	Overall	Task	Query	Overall	Task	Query	Overall	Session	Query	Overall	Session	Query
GPT-4omini	0.35	0.30	0.28	0.36*	0.30	0.25	0.29	0.35	0.44+	0.37	0.41	0.48*
GPT-3.5turbo	0.37+	0.31	0.26	0.33	0.28	0.26	0.21	0.34	0.42	0.37	0.45	0.47*
Llama 3.1 8B-Instruct	0.31	0.26	0.20	0.31	0.27	0.27	0.25	0.29	0.39	0.27	0.32	0.39
Llama 3.2 3B-Instruct	0.19	0.12	0.06	0.22	0.17	0.21	0.13	0.23	0.32	0.30	0.37	0.46
Llama 3.3 70B-Instruct	0.38+	0.34	0.30	0.37*	0.32	0.28	0.33	0.38	0.44+	0.37	0.41	0.42
DeepSeek-R1-Distill-Qwen-14B	0.33	0.27	0.22	-	-	-	0.15	0.13	0.11	-	-	-

usefulness: 3 = Very Useful, 2 = Fairly Useful, 1 = Somewhat Useful, 0 = Not Useful at All. We evaluate OpenAI's GPT-3.5 Turbo, GPT-4omini, and Meta's latest open-source LLMs (Llama 3.3 70B-Instruct, Llama 3.2 3B-Instruct, Llama 3.1 8B-Instruct) [12] for usefulness judgment. For LLM coverage, we included the reasoning model DeepSeek-R1-Distill-Qwen-14B along with other LLMs. Experiments were conducted via AWS Bedrock for Llama models and OpenAI APIs for GPT models, using a temperature of 0 and `top_p` of 1. DeepSeek-R1-Distill-Qwen-14B was hosted locally via Ollama with the same temperature parameter.

5 Results

Our experiments address the three RQs for baseline and session data and evaluate LLM-generated usefulness at the overall level across all user interactions, while task and query scores are averaged separately to capture different granularity of ranking quality.

5.1 RQ1: Baseline and Session Analysis

We prompt the LLM models with our prompt and generate usefulness labels for each of the clicked documents. Both datasets already contain the usefulness labels annotated by humans that we use as our ground truth. To address our **RQ1**, we evaluate how well the LLM-generated usefulness performs with respect to our ground truth. As the usefulness labels are ordinal, starting from 0 (low) to 3 (high), we conduct Spearman's rank correlation (ρ) to measure association between ranked variables. The detailed analysis in Table 2 shows a moderate positive correlation between human and LLM-generated usefulness labels. The best performing models in both datasets are denoted with + for baseline and * for session. Considering **Baseline** analysis, for THUIR-KDD'19 dataset for overall, Llama 3.3 70B-Instruct has the highest correlation ($\rho = 0.38$), followed by GPT-3.5turbo ($\rho = 0.37$). The weakest correlation is observed in Llama 3.2 3B-Instruct ($\rho = 0.19$), which is likely due to its relatively small parameter size (also shown in Figure 1). For the QRef dataset, the overall trend is similar with Llama 3.3 70B-Instruct performing moderately well ($\rho = 0.33$). Task-level and query-level findings for THUIR-KDD'19 align closely with overall trends, reinforcing consistency in our findings across all LLM models. In QRef's baseline dataset for query level, both GPT-4omini ($\rho = 0.44$) and Llama 3.3 70B-Instruct ($\rho = 0.44$) perform the best. For **Session** analysis, performance gradually improves compared to baseline, indicating that LLMs benefit from session-context search data. In QRef, improvement is observed in session compared to baseline in overall, task-level and query-level session data. This suggests that

Table 3: Comparison of LLM performances across 4 cases in baseline (B) and session (S) data.

Model	HR-HU		HR-LU		LR-HU		LR-LU	
	B	S	B	S	B	S	B	S
GPT-3.5turbo	0.10	0.12	0.02	0.09	0.17	0.04	0.07	0.05
GPT-4omini	0.12	0.13	-0.13	0.11	0.10	0.05	0.05	0.13
Llama 3.1 8B-Instruct	0.09	0.11	0.05	0.03	0.12	0.04	0.04	0.11
Llama 3.2 3B-Instruct	0.06	0.06	0.04	0.05	0.19	0.05	0.03	0.03
Llama 3.3 70B-Instruct	0.16	0.14	0.13	0.08	-0.02	0.06	0.11	0.11
DeepSeek-R1-Distill-Qwen-14B	0.13	-	0.12	-	0.06	-	0.12	-

LLMs perform better in shorter (QRef) rather than longer (THUIR-KDD'19) search sessions. Notably, in QRef session data at the query level ($\rho = 0.48$ and 0.47), GPT models perform the best among all. This further supports the observation that session-level context enhances LLM performance, particularly in query-level analysis. Despite DeepSeek's proficient reasoning ability, it has struggled in this task. During usefulness label generation, even after direct prompting instead of providing direct usefulness labels it often generated reasoning-based responses, requiring additional filtering and extraction of the labels from the generated responses. From our experiments, Llama 3.3 70B-Instruct remains the best performing model across different levels of analysis. Its 70B parameter count likely contributes to its superior ability to process and interpret nuanced contextual information. GPT-4omini follows next, performing well in both baseline and session contexts. Smaller models like Llama 3.2 3B-Instruct struggle to generate usefulness labels but improve performance with session data particularly in QRef's session data at both session ($\rho = 0.37$) and query level ($\rho = 0.46$).

5.2 RQ2: Relevance vs Usefulness Divergence

We analyze the correlation between relevance and usefulness across four cases: HR-HU (High Relevance, High Usefulness), HR-LU (High Relevance, Low Usefulness), LR-HU (Low Relevance, High Usefulness) and LR-LU (Low Relevance, Low Usefulness). As only THUIR-KDD'19 dataset has relevance labels we choose this dataset to conduct this part of experiment. We divide the dataset into four cases of usefulness-relevance combinations and tested the performance of LLMs. Both the relevance and usefulness labels are labeled from 0 to 3 where we consider 0 & 1 as Low and 3 & 4 as High. From our initial observation, we report the Spearman rank correlation (ρ) of the baseline and session data for 5 selected LLMs in Table 3. To address our **RQ2**, we find that LLMs face challenges in capturing the contrasting relationship between relevance and usefulness in

Table 4: Ablation study of features in GPT-4omini and LaMA 3.3 70B-Instruct with session data.

Features	GPT-4omini	LaMA 3.3 70B-Instruct
U_{R+S+U}	0.36	0.37
U_{R+S}	0.35	0.35
U_{R+U}	0.34	0.33
U_{S+U}	0.34	0.31
U_R	0.32	0.33
U_S	0.27	0.29
U_U	0.27	0.28

both baseline and session analysis. According to Spearman’s rank correlation (ρ), all five LLM models are showing weak positive correlation and few of them have weak negative correlation. For the diverging cases HR-LU and LR-HU, the models face challenges in generating accurate usefulness labels. This may be due to the nature of Spearman’s rank correlation, which relies on the overall rank order of data points. When the dataset is divided into four cases, the reduced range of rank variation and smaller sample sizes can weaken the correlation, making it harder to capture the relationship between human and LLM-generated usefulness labels within each subgroup. Furthermore, it shows that LLMs generalize measures such as high relevance will lead to high usefulness which is not the case always in real settings. The DNA structured prompt is solely not enough to capture the divergent cases, as such cases should be explicitly mentioned in the prompts for better clarification if expanded for future research.

5.3 RQ3: Ablation Study

To address our **RQ3**, we conduct an ablation study of metrics provided to the LLM. We categorize the features into five levels: query features (Q), document features (D), relevance metrics (R), satisfaction metrics (S), and user behavior features (U) as shown in Table 4. For the ablation study, we keep Q and D constant while changing R, S and U. This approach aims to identify the most critical features required for LLMs to effectively generate usefulness labels. In real-world applications using all possible inputs in the LLM prompt is token-intensive and cost-prohibitive that’s why it is important to optimize feature selection. For conducting the ablation study, we chose Llama 3.3 70B-Instruct and GPT-4omini on the THUIR-KDD’19 session data as these two consistently outperformed others in usefulness judgment. From the ablation study shown in Table 4, observing the Spearman’s correlation (ρ) models seem to perform the best when all three features such as relevance (R), satisfaction (S) and user features (U) are provided along with query (Q) and document features (D) with a moderate positive correlation ($\rho = 0.36$ and 0.37). The next is followed by relevance (R) and satisfaction features (S) that also has a moderate positive correlation ($\rho = 0.35$). This shows that LLMs can infer usefulness if both relevance and satisfaction levels are present supporting prior research on usefulness [23]. Notably, when satisfaction (S) and user features (U) are used individually it shows the lowest correlation ($\rho = 0.27$ and 0.28) indicating that both features are associated with usefulness should be considered for improved performance. The findings of

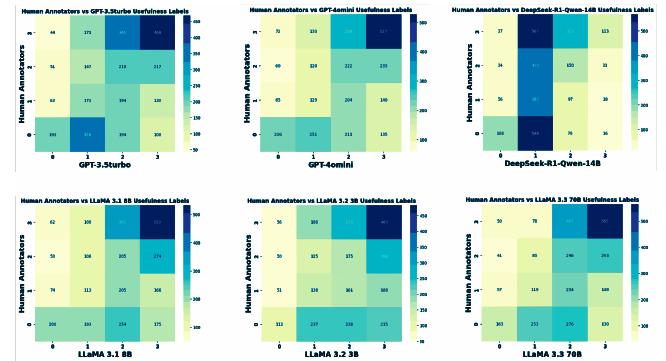


Figure 1: Human annotators vs. LLM usefulness labels.

the ablation study highlights relevance and satisfaction as key in determining usefulness and both combined can improve LLM’s ability to capture nuanced contextual information. However, when features are provided independently, LLM struggles to effectively evaluate usefulness making it challenging to perform an accurate usefulness evaluation. This highlights the importance of combining features and focusing on relevance to enable the LLM to interpret and evaluate the usefulness of search sessions effectively.

6 Discussion and Conclusion

In this study, we explored the potential of LLMs in automated usefulness judgment tasks and assessed the performance of existing pre-trained models across different datasets and search contexts. Our findings indicate that while LLMs exhibit low to moderate performance in usefulness labeling, their effectiveness improves in shorter search sessions where they achieve higher correlation with human judgments. Additionally, larger models show a better grasp of contextual search relationships, consistently outperforming baseline approaches. Differing from traditional TREC-style relevance assessments, which primarily rely on query-document pair semantic understanding, usefulness judgment requires a more comprehensive understanding of user intents, tasks, and search contexts, making it inherently more complex and user-centered. A key limitation we encountered was the lack of a uniform dataset for usefulness evaluation, making it difficult to establish consistent benchmarks. Additionally, performance of pre-trained LLMs in usefulness labeling proved moderate, highlighting the need for more targeted fine-tuned adaptations in user-centered IR evaluation. This study opens a new research and application direction under *LLMs-as-Judges* research [15] by examining how LLMs can analyze search behavior logs and predict usefulness, and sheds light on both their potential and challenges. Built upon the findings, our future research will develop fine-tuning techniques and specialized models to enhance LLM performance in document usefulness labeling. Furthermore, we aim to implement models for automated usefulness judgment in text and conversational retrieval evaluation, real-time *usefulness-based* re-ranking, and search path recommendations.

Acknowledgments

This work was partially supported by cloud computing credits provided through an Amazon Research Award to Dr. Chirag Shah.

References

- [1] Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. 2024. Can We Use Large Language Models to Fill Relevance Judgment Holes? <http://arxiv.org/abs/2405.05600> arXiv:2405.05600.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyam Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [3] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2024. LLMs can be Fooled into Labelling a Document as Relevant: best café near me; this paper is perfectly relevant. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 32–41.
- [4] Nicholas J Belkin, Michael Cole, and Ralf Bierig. 2008. Is relevance the right criterion for evaluating interactive information retrieval. In *Proceedings of the ACM SIGIR 2008 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments*. <http://research.microsoft.com/~pauben/bbr-workshop>. Citeseer.
- [5] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 621–630.
- [6] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the web conference 2021*. 743–755.
- [7] Nuo Chen, Jiqun Liu, Xiaoyu Dong, Qijiong Liu, Tetsuya Sakai, and Xiao-Ming Wu. 2024. AI Can Be Cognitively Biased: An Exploratory Study on Threshold Priming in LLM-Based Batch Relevance Assessment. <https://doi.org/10.48550/arXiv.2409.16022> arXiv:2409.16022 [cs].
- [8] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 15–24.
- [9] Aleksandr Chuklin, Pavel Serdyukov, and Maarten De Rijke. 2013. Click model-based information retrieval metrics. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 493–502.
- [10] Michael Cole, Jingjing Liu, Nicholas J Belkin, Ralf Bierig, Jacek Gwizdka, Chang Liu, Jin Zhang, and Xiangmin Zhang. 2009. Usefulness as the criterion for evaluation of interactive information retrieval. In *Proceedings of the third workshop on human-computer interaction and information retrieval Cambridge*. HCIR, 1–4.
- [11] Gordon V Cormack, Christopher R Palmer, and Charles LA Clarke. 1998. Efficient construction of large test collections. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 282–289.
- [12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [13] Guglielmo Faggioli, Laura Dietz, Charles 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. <http://arxiv.org/abs/2304.09161> arXiv:2304.09161.
- [14] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)* 23, 2 (2005), 147–168.
- [15] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghang Shen, Shengjie Ma, Honghai Liu, et al. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594* (2024).
- [16] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and predicting graded search satisfaction. In *Proceedings of the eighth ACM international conference on web search and data mining*. 57–66.
- [17] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Cheng-gang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
- [18] Jiqun Liu. 2021. Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management* 58, 3 (2021), 102522.
- [19] Jiqun Liu. 2022. Toward Cranfield-inspired reusability assessment in interactive information retrieval evaluation. *Information Processing & Management* 59, 5 (Sept. 2022), 103007. <https://doi.org/10.1016/j.ipm.2022.103007>
- [20] Jiqun Liu, Shawon Sarkar, and Chirag Shah. 2020. Identifying and predicting the states of complex search tasks. In *Proceedings of the 2020 conference on human information interaction and retrieval*. 193–202.
- [21] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. " Satisfaction with Failure" or " Unsatisfied Success" Investigating the Relationship between Search Success and User Satisfaction. In *Proceedings of the 2018 world wide web conference*. 1533–1542.
- [22] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does Relevance Mean Usefulness and User Satisfaction in Web Search? In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, Pisa Italy, 463–472. <https://doi.org/10.1145/2911451.2911507>
- [23] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 463–472.
- [24] Matthew Mitsui, Jiqun Liu, Nicholas J Belkin, and Chirag Shah. 2017. Predicting information seeking intentions from search behaviors. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. 1121–1124.
- [25] Hossein A Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. LLM4Eval: Large Language Model for Evaluation in IR. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Washington DC USA, 3040–3043. <https://doi.org/10.1145/3626772.3657992>
- [26] Hossein A Rahmani, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, Paul Thomas, Charles LA Clarke, Mohammad Aliannejadi, Clemencia Siro, and Guglielmo Faggioli. 2024. Llmjudge: Llms for relevance judgments. *arXiv preprint arXiv:2408.08896* (2024).
- [27] Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2395–2400.
- [28] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. <http://arxiv.org/abs/2309.10621> arXiv:2309.10621
- [29] 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. <https://doi.org/10.48550/arXiv.2411.08275> arXiv:2411.08275 [cs].
- [30] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: UMBrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. <https://doi.org/10.48550/arXiv.2406.06519> arXiv:2406.06519 [cs].
- [31] Fan Zhang, Ke Zhou, Yunqiu Shao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. How well do offline and online evaluation metrics measure user satisfaction in web image search?. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 615–624.
- [32] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are Large Language Models Good at Utility Judgments? <https://doi.org/10.48550/arXiv.2403.19216> arXiv:2403.19216 [cs].