Criteria-Based LLM Relevance Judgments

Naghmeh Farzi
Naghmeh.Farzi@unh.edu
University of New Hampshire
Durham, NH, USA

Laura Dietz
dietz@cs.unh.edu
University of New Hampshire
Durham, NH, USA

Abstract

Relevance judgments are crucial for evaluating information retrieval systems, but traditional human-annotated labels are time-consuming and expensive. As a result, many researchers turn to automatic alternatives to accelerate method development. Among these, Large Language Models (LLMs) provide a scalable solution by generating relevance labels directly through prompting. However, prompting an LLM for a relevance label without constraints often results in not only incorrect predictions but also outputs that are difficult for humans to interpret.

We propose the Multi-Criteria framework for LLM-based relevance judgments, decomposing the notion of relevance into multiple criteria—such as exactness, coverage, topicality, and contextual fit—to improve the robustness and interpretability of retrieval evaluations compared to direct grading methods. We validate this approach on three datasets: the TREC Deep Learning tracks from 2019 and 2020, as well as LLMJudge (based on TREC DL 2023). Our results demonstrate that Multi-Criteria judgments enhance the system ranking/leaderboard performance. Moreover, we highlight the strengths and limitations of this approach relative to direct grading approaches, offering insights that can guide the development of future automatic evaluation frameworks in information retrieval.[1]

CCS Concepts

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

Keywords

large language models, LLM evaluation, relevance criteria

ACM Reference Format:
Naghmeh Farzi and Laura Dietz. 2025. Criteria-Based LLM Relevance Judgments. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '25), July 18, 2025, Padua, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3731120.3744591

1 Introduction

The evaluation of information retrieval (IR) systems relies on relevance judgments to assess the quality of retrieved passages. By determining which pieces of information are relevant to specific information needs (queries), we derive evaluation scores for different IR systems. These scores allow us to compare systems and select the most effective approaches. For a benchmark of queries and a corpus, the evaluation process depends on having relevance labels for each query and document (or passage) pair in the corpus.

While human-assigned relevance labels remain the gold standard, their acquisition at scale presents significant challenges in terms of cost, time, and consistency. Ultimately, only human judges can be the arbitrator of relevance [7, 13, 30], but we aim to provide methods to support the rapid development of new information retrieval approaches.

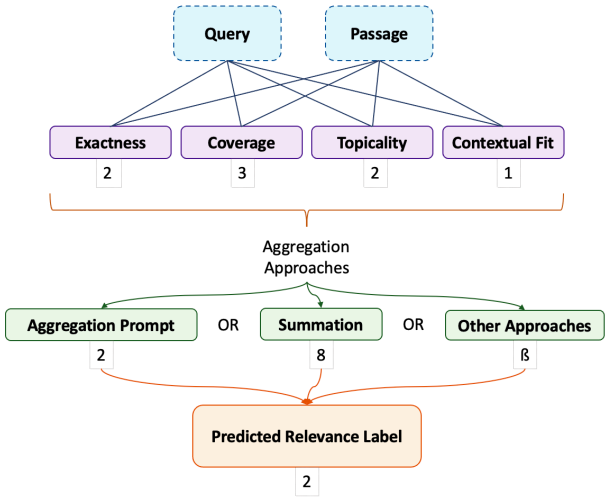

Figure 1: Overview the two phases of Multi-Criteria: grading criteria individually, then aggregating grades (see Section 3).

We study the task of automatically predicting passage-level relevance labels that can be used with standard IR evaluation tools.

Task Statement. Given a query and a passage, predict the relevance on a graded scale. The objective is to produce relevance labels that, when used to evaluate a set of retrieval systems, yield a system ranking (i.e., leaderboard) similar to the one derived from human-annotated judgments.

Example Domain. While our formulation is task-agnostic, all experiments are conducted on the TREC Deep Learning (DL) benchmark. The TREC DL uses a 0–3 relevance scale, where 3 indicates perfect relevance and 0 denotes no relevance. Retrieval systems submitted to this benchmark are ranked using gold-standard labels. Our goal is to predict relevance labels such that the resulting leaderboard closely mirrors that of the human judgments.

[1] **Online Appendix at https://github.com/TREMA-UNH/appendix-criteria-based-llm-relevance-judgments/**

ICTIR '25, Padua, Italy

ACM ISBN 979-8-4007-1861-8/2025/07
https://doi.org/10.1145/3731120.3744591

Issues with the current LLM Judge approaches. Recent advances in large language models (LLMs) have shown promising potential for automated relevance assessment [13, 18, 31, 32]. These approaches are based on the careful design of one LLM prompt, which for a given query-document pair can obtain a relevance label. Moreover, they have demonstrated strong empirical results in terms of leaderboard correlation, regularly obtaining Kendall's tau above 0.9, and an inter-annotator agreement Cohen's kappa of 0.40.

Grading with a single prompt, where the model is merely asked, "Is this document relevant?", comes with notable challenges. First, LLMs can sometimes produce unpredictable results, especially if the prompt design is not carefully controlled. Subtle variations in how a prompt is phrased may yield drastically different relevance predictions. What is worse is that this behavior changes across LLM families and versions. Second, the single-label output ("relevant" or "not relevant") obscures the rationale behind the decision, limiting the transparency of the evaluation process. In IR and many other fields, interpretability is often vital: understanding why a decision was made can help diagnose systematic biases, guide system improvements, and build confidence among stakeholders. Finally, while LLMs may have been exposed to enormous corpora, their notions of relevance could deviate from established IR conventions or human judgments if they rely on obscure patterns in their training data. Thus, while single-prompt approaches facilitate labeling large datasets, they are met with skepticism regarding their reliability, primarily due to a lack of transparency in the underlying reasoning.

Contribution. To enable a more structured assessment, we propose a framework that evaluates relevance by decomposing it into multiple pre-defined criteria. Our approach assesses each criterion independently and then combines these evaluations to derive a final relevance label for the documents. While inspired by Chain-of-Thought (CoT) prompting [35], our framework takes a distinctive approach by evaluating each criterion independently and in parallel, rather than following a sequential process. This approach allows us to investigate whether decomposing relevance into criteria can both improve evaluation quality and maintain competitiveness with existing LLM-based methods, while also making the grading process more transparent to human evaluators. We compare our method with recent frameworks, such as UMBRELA [34], which demonstrates better performance with larger LLMs [16]. Our results show that, even with smaller LLMs, our approach provides equally effective evaluations.

2 Related Work

2.1 LLM Based Relevance Judgment

Recent work has explored using LLMs for automated relevance assessment. Many approaches rely on a single, hand-crafted prompt that the LLM uses to generate a relevance label. Instances of this approach are as follows. Thomas et al. [32] demonstrate that LLMs can effectively generate relevance judgments through direct prompting, establishing a baseline for automated evaluation methods, like the UMBRELA framework [34]. Farzi and Dietz [16] show that achieving optimal results with UMBRELA requires large LLMs, such as GPT-4. Arabzadeh and Clarke [3] compare LLM-based methods (binary, graded, pairwise, and nugget-based) on TREC DL [9, 10] and ANTIQUE [17] datasets, finding pairwise preferences best align with human labels while binary and graded (UMBRELA) methods achieve higher system ranking correlations. Sun et al. [31] show zero-shot capabilities for relevance assessments. MacAvaney et al. [18] suggest one-shot learning for binary preference judgments. While Faggioli et al. [13] show that LLM-based relevance assessment can be effective, they also warn of several pitfalls, including a potential mismatch between predicted labels and human-perceived relevance.

Beyond prompting, Zhang et al. [37] introduce a utility-based framework for QA datasets, outperforming relevance judgments. Meng et al. [20] introduced QPP-GenRE based on fine-tuning LLMs to generate relevance judgments for query performance prediction on TREC DL datasets. However, Alaofi et al. [2] show LLMs misclassify irrelevant passages due to lexical overlap, and Abbasiantaeb et al. [1] find that LLM-generated relevance judgments in conversational search tasks distort rankings when mixed with partial human annotations. These biases and inconsistencies show that single, coarse-grained metrics often fail to capture nuanced judgments.

2.2 Grading Rubrics and Relevance Nuggets

Recent advancements in LLM prompting, such as Wei et al.'s [35] chain-of-thought and Zhou et al.'s [38] least-to-most prompting, highlight the benefits of task decomposition. However, these techniques have not been used for the assessment of relevance criteria. Emerging approaches focus on decomposing relevance into smaller, LLM-verifiable points that are more interpretable to humans. The RUBRIC Autograder Workbench [11, 15] creates grading rubrics for each query, using answerability as a relevance indicator. The EXAM Answerability Metric [28] and EXAM++ [14] uses multiple-choice questions and question-answering to test passage relevance. Other studies [19, 21, 22] decompose queries into topical nuggets for relevance assessment. Our approach introduces a novel method for decomposing relevance based on specific criteria.

2.3 Relevance Criteria in Classic IR

The conceptualization of relevance in information retrieval (IR) has evolved from a simple focus on topicality to a multi-criteria framework. Saracevic [29], building on Cosijn and Ingwersen [8] and Borlund [5], identified five dimensions of relevance: algorithmic, topical, cognitive, situational, and affective. Empirical studies, such as Boyce [6], further support the view that relevance includes user needs, context, and emotional goals. Barry et al. [4] and Xu et al. [36] found that users rely on multiple criteria, like accuracy, specificity, and scope, when making relevance judgments, extending beyond topicality to include factors like novelty, reliability, and understandability. As conversational AI systems emerge, Sakai et al. [27] identified a schema of twenty-one evaluation criteria. However, previous frameworks were developed mainly for user-centric IR. Our work builds on these studies using criteria tailored for topical ad hoc retrieval.

Table 1: Chat completions prompt for criterion-specific grading. Variables, denoted in blue with curly braces, are replaced with respective content.

Criterion-Specific Grading
System Message: Please assess how well the provided passage meets specific criteria in relation to the query. Use the following scoring scale (0-3) for evaluation: 0: Not relevant at all / No information provided. 1: Marginally relevant / Partially addresses the criterion. 2: Fairly relevant / Adequately addresses the criterion. 3: Highly relevant / Fully satisfies the criterion.
Prompt: Please rate how well the given passage meets the {Criterion Name} criterion in relation to the query. The output should be a single score (0-3) indicating {Criterion Description}. Query: {Query} Passage: {Passage} Score:

2.4 Research Gap

Despite the progress in LLM-based relevance judgment and multi-criteria evaluation, a gap remains in integrating these approaches. Existing LLM evaluation methods often treat relevance as a monolithic concept, while traditional criteria-based approaches have not been adapted for automated assessment. Our work addresses this gap by proposing a framework that leverages LLMs' capabilities while preserving the interpretability and theoretical foundations of criteria-based evaluation.

3 Approach

Our Multi-Criteria framework decomposes the notion of relevance into multiple fundamental criteria that together indicate whether a passage is relevant for a query.

In this work, we experiment with four criteria in particular: Exactness, Coverage, Topicality, and Contextual Fit. This decomposition addresses two key challenges in LLM-based relevance judgments: the tendency of LLMs to conflate different aspects of relevance, and the need for interpretable judgments that are explainable and auditable by a human overseeing the evaluation process. Our approach consists of two phases that systematically separate individual criterion evaluation from the final relevance label aggregation. An overview of the approach is shown in Figure 1.

Phase One: Criterion Grading: Each criterion is evaluated independently for every query-passage pair, enabling a focused assessment.

Phase Two: Relevance Label Prediction: After grading all criteria, we aggregate the individual criterion grades from Phase One to produce a relevance label that reflects the passage's overall relevance for the query. This aggregation accounts for the semantic differences among different criteria, allowing for a more structured assessment of relevance.

A key advantage of our approach is that human evaluators can verify each criterion grade and its aggregation process, providing transparent oversight of automatic grading. We demonstrate this with an example in Section 4.6. Furthermore, the criteria and the aggregation method can be tailored to specific datasets and query types, offering flexibility across different application domains.

In the following we elaborate our Multi-Criteria approach in more detail.

Table 2: Chat completions prompt used in Phase Two to aggregate criterion-level grades used into a final relevance label for each query-passage pair. Variables are replaced by the corresponding grades obtained from Phase One.

Aggregating Criterion Grades into a Relevance Label
System Message: You are a search quality rater evaluating the relevance of passages. Given a query and passage, you must provide a score on an integer scale of 0 to 3 with the following meanings: 3 = Perfectly relevant: The passage is dedicated to the query and contains the exact answer. 2 = Highly relevant: The passage has some answer for the query, but the answer may be a bit unclear, or hidden amongst extraneous information. 1 = Related: The passage seems related to the query but does not answer it. 0 = Irrelevant: The passage has nothing to do with the query. Assume that you are writing an answer to the query. If the passage seems to be related to the query but does not include any answer to the query, mark it 1. If you would use any of the information contained in the passage in such an answer, mark it 2. If the passage is primarily about the query, or contains vital information about the topic, mark it 3. Otherwise, mark it 0.
Prompt: Query: {Query} Passage: {Passage} Exactness: {Exactness Grade} Topicality: {Topicality Grade} Coverage: {Coverage Grade} Contextual Fit: {Contextual Fit Grade} Please rate how the given passage is relevant to the query based on the given scores. The output must be only a score (0-3) that indicates how relevant they are. Score:

3.1 Phase One: Criterion-Specific Grading

In the first phase, each query-passage pair is evaluated based on four relevance criteria elaborated below. These criteria align with diverse user expectations, reflecting various dimensions of relevance [29]. While these criteria are tailored towards TREC DL task, they can be adapted for different datasets and goals. We use the following criterion name and description as part of our prompt-based grading approach, as elaborated in Table 1.[2] Variables in the prompt templates (denoted in blue and curly braces), such as criterion name or query, are replaced according to the criteria definitions. To ensure grammatical compatibility within the prompt

[2]Prompts are designed for chat completion models; for models that do not support chat completion, both messages can be concatenated.

format, the question mark in the Criterion Description is removed when used in the prompt.

Criterion Name: Exactness
Criterion Description: How precisely does the passage answer the query?
Rationale: This criterion assesses whether the passage provides a specific, accurate response to the query's core information need. It prioritizes targeted information over lexical overlap, ensuring relevance for goal-oriented or question-based queries.

Criterion Name: Topicality
Criterion Description: Is the passage about the same subject as the whole query (not only a single word of it)?
Rationale: This criterion ensures the passage addresses the entire query subject, aligning with topical relevance, which is determined by the degree of "aboutness" [29], how closely the passage's subject matches the topic expressed in the query. This guards against partial matches where only isolated terms from the query appear, rather than the full thematic content.

Criterion Name: Coverage
Criterion Description: How much of the passage is dedicated to discussing the query and its related topics?
Rationale: A passage with high Coverage minimizes the user's need to seek additional information, aligning with the topical relevance through its focus on the breadth of relevant content. Additionally, it reflects a cognitive relevance [29] aspect: reducing the user's cognitive load by minimizing the need for further information seeking.

Criterion Name: Contextual Fit
Criterion Description: Does the passage provide relevant background or context?
Rationale: This criterion evaluates whether the passage offers relevant and appropriate background information that complements the overarching themes of the query. Contextual information helps resolve uncertainty and supports decision-making within a task-driven setting [29].

Each criterion is evaluated through dedicated prompts containing the criterion name and description, ensuring focused assessment of specific relevance aspects. Inspired by Chain-of-Thought prompting's principle of breaking down complex reasoning, this isolation strategy simplifies the task for LLMs by having them focus on one relevance aspect at a time rather than juggling multiple criteria simultaneously. This decomposition is particularly beneficial for smaller models, reducing task complexity and leading to more reliable criterion-specific evaluations. Additionally, the approach enhances transparency by creating clear links between individual prompts and their corresponding criterion assessments.

3.2 Phase Two: Relevance Label Prediction

Building on the criterion-based grades obtained in Phase One, this phase determines the overall relevance of the passage to the query by predicting a single relevance label. Various methods can be used for this aggregation, ranging from machine learning techniques to simple mathematical models. Here, we focus on a prompt-based aggregation approach that utilizes the criteria-specific grades.

Table 3: Mapping criterion grade sums to relevance labels used in Sumdecompose.

Criterion Grade Sum	10–12	7–9	5–6	0–4
Relevance Label	3	2	1	0

3.2.1 Multi-Criteria: Prompt-Based Aggregation. Leveraging reasoning capabilities of LLMs, we employ an aggregator prompt, as outlined in Table 2, to instruct the LLM to evaluate the passage's overall relevance. This prompt integrates the query, passage, and criterion-specific grades, enabling the LLM to generate a comprehensive relevance label. We investigate the following combinations:

All four: Using grades from all four criteria to produce a comprehensive relevance judgment.
Three criteria: Ablation study of three criteria to examine which subsets contribute most to effective relevance assessment.
Single criterion: Using each individual criterion as the relevance label to evaluate its standalone effectiveness.

Each criterion is evaluated with prompts of Table 1 which include the query passage pair under evaluation, judgment instructions, and the exact criterion name and description. By standardizing prompt structures across all criteria while preserving their unique details, we enable meaningful comparisons across combinations.

3.2.2 Sumdecompose: Summation-based Aggregation: An alternative is to sum the Phase One criterion grades and derive the final relevance label heuristically based on the total, using the mapping shown in Table 3.

To assign a relevance label, we define grade thresholds over the total score, controlling the strictness of the evaluation. The thresholds used in our experiments are tuned on the LLMJudge development set (held out from evaluation) to maximize correlation with human judgments.

4 Experimental Evaluation

We evaluate our Multi-Criteria approach to LLM-evaluation with respect to the following research questions:

RQ1: How well does Multi-Criteria perform compared to other strong methods?
RQ2: To what extent does the underlying LLM affect the results?
RQ3: How effectively do the criteria capture different aspects of relevance?
RQ4: How closely match predicted relevance labels the manual judgments?
RQ5: What examples show that Multi-Criteria behaves as expected?

4.1 Experimental Setup

4.1.1 Datasets. We conduct our experiments on test sets of three major TREC datasets to ensure robust evaluation across different time periods and query types:

LLMJudge (primary dataset) [26]: 25 queries, 4,414 query-passage pairs, using systems submitted to the TREC Deep Learning Track of 2023. We primarily focus our analysis on the LLMJudge

Table 4: Example from LLMJudge test set showing one query and two passages evaluated using Multi-Criteria grading with LLaMA-3-8B. For each passage, criterion grades (0-3) are assigned by LLaMA-3-8B in Phase-One via the Criterion-Specific prompts (see Table 1) and a predicted relevance label is obtained in Phase Two via the aggregation prompt (see Table 2). Manual relevance judgments are included from the dataset and for comparison. Rationales are written post-hoc, manually, and not generated by the LLM or drawn from the dataset. The differences in Coverage and Contextual Fit lead to different relevance predictions, illustrating nuances of the graded relevance labeling process.

Query `3100119` Do larger lobsters become tougher when cooked?			
`msmarco_passage_01_305447845`		`msmarco_passage_04_612526438`	
Be sure to not overcook the lobster tails as the lobster meat will be tough and chewy. A fully cooked lobster tail will have a bright red shell and the lobster meat will be firm and white. Serve with a squeeze of lemon, salt and pepper to taste.		Saffitz takes it a step further, explaining that the best meat comes from the lobster knuckle. The tail can be chewy and tough, she explains, and the claws are stringy and often overcooked. "The knuckle is already perfectly shredded, and so tender," she says.	
Exactness: 1	No direct answer about size-toughness relationship, only mentions tail toughness from overcooking	**Exactness: 1**	No size-toughness relationship, only describes texture differences
Coverage: 2	Mostly focuses on toughness in cooking, partly on serving	**Coverage: 1**	Content split between chef attribution and comparing textures, with less focus on lobster cooking details
Topicality: 2	Mentions cooking and toughness, misses size aspect	**Topicality: 2**	Discusses parts and toughness, misses size aspect
Contextual Fit: 2	Provides visual indicators but lacks background on how cooking methods affect texture	**Contextual Fit: 1**	Lobster's part differences don't focus on cooking and/or size
Predicted Relevance Label: 2 **Manual Judgment: 2**		**Predicted Relevance Label: 1** **Manual Judgment: 1**	

dataset, as it represents the most recent public test collection and minimizes potential data leakage with LLM training.

DL20 [9]: 43 queries, 37 systems for passage retrieval from TREC Deep Learning track of 2020.

DL19 [10]: 54 queries, 59 systems for passage retrieval from TREC Deep Learning track of 2019.

4.1.2 Large Language Models. We evaluate our approach using three different openly available LLM models:

LLaMA-3-8B (primary LLM):[3] A medium-scale public model that performs well with our approach.

LLaMA-3.3-70B:[4] A recent large-scale public model from the same family.

FLAN-T5-large:[5] A small-scale LLM used for fair comparison with the RUBRIC system [11].

Each model was used with identical prompting templates to ensure fair comparison. The temperature was set to 0.0 for deterministic outputs, and the maximum token length was set to 100 tokens. For model evaluation, we used FLAN-T5-large and LLaMA-3-8B on an NVIDIA A40 GPU, and LLaMA-3.3-70B through an API service [33]. To address RQ2 and RQ4, we compare our Multi-Criteria approach using these LLMs against a diverse set of LLMs and methods, including fine-tuned and prompt-based approaches, in the LLMJudge challenge (Table 5).

[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[4]https://api.together.xyz/models/meta-llama/Llama-3.3-70B-Instruct-Turbo
[5]https://huggingface.co/google/flan-t5-large

4.1.3 Evaluation Measures. Our main objective is to assess if an LLM-evaluation method can reliably identify which of many IR systems performs best. To accomplish this, we compare each system's ranking under automatically generated relevance labels to its official ranking based on manual assessments. We measure how similar these two leaderboards are using rank correlation metrics:

Spearman's rank correlation: Reflects how well the relative ordering of systems aligns across the two leaderboards, taking into account each system's exact rank position.

Kendall's tau: Considers the consistency of pairwise system orderings between the two leaderboards.

Both measures range from -1.0 to +1.0, where higher correlation values indicate better alignment between the automatic relevance labels and the official human-labeled assessments.

Significance analysis. As rank correlation measures do not support a typical significance analysis in terms of standard-error or paired-t-tests, we consider small differences of 0.005 as equally good and highlight only differences of at least 0.025 to the best score.

Leaderboards are produced by using predicted relevance labels with `trec_eval` using the following evaluation metrics

NDCG@10 (primary measure): Normalized cumulative gain using a cutoff at rank 10. This is the primary measure used in all TREC DL test collections.

MAP: A mean of average-precision of all relevant passages.

MRR: A mean of reciprocal rank of the first relevant passage.

Table 5: Top 15 LLMJudge challenge results from Rahmani et al. [26], evaluated using Spearman's rank correlation and Kendall's Tau, ordered by Spearman's rank. Our proposed Multi-Criteria (denoted "4prompts") method, using LLaMA-3-8B, ranks first in Spearman's correlation and second in Kendall's Tau, outperforming many large-scale LLMs. Full results are available in Rahmani et al. [26].

Submission	Spearman's rank	Kendall's Tau
TREMA-4prompts **(ours)**	**0.9919**	0.9483
prophet-setting2	0.9914	**0.9516**
NISTRetrieval-instruct0	0.9907	0.9440
NISTRetrieval-instruct1	0.9907	0.9440
NISTRetrieval-instruct2	0.9907	0.9440
RMITIR-llama70B	0.9883	0.9353
TREMA-sumdecompose **(ours)**	0.9870	0.9300
Olz-multiprompt	0.9867	0.9267
llmjudge-thomas3	0.9867	0.9181
TREMA-all	0.9863	0.9138
llmjudge-simple1	0.9863	0.9181
TREMA-questions	0.9839	0.9095
TREMA-naiveBdecompose **(ours)**	0.9838	0.9128
Olz-halfbin	0.9830	0.9085
h2oloo-zeroshot1	0.9827	0.9181
...	...	...

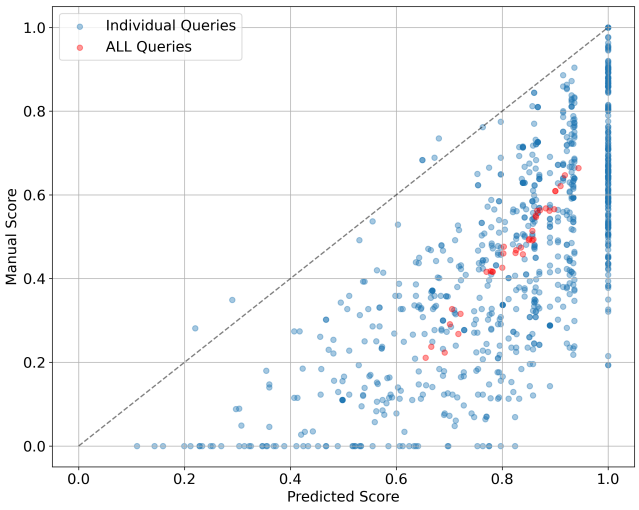

Figure 2: NDCG@10 evaluation scores under manual judgments vs. automatic relevance labels. Red dots represent the system performance, their ordering representing the leaderboard. For reference, all per-query/system evaluation scores are represented as blue dots. The red diagonal confirms that all systems are nearly in the same order under both automatic and manual methods, which results in a Spearman Rank coefficient of 0.99.

While our approach is compatible with any number of desired relevance-criteria, here we study our Multi-Criteria approach with Exactness **(E)**, Coverage **(C)**, Topicality **(T)**, and Contextual Fit **(F)** criteria, reflecting common aspects of relevance tailored to the TREC DL setting. Likewise, the aggregation step, whether prompt-based, summation-based, or classifier-based is interchangeable, enabling the use of alternative strategies as needed.

4.2 RQ1: Performance of Multi-Criteria Compared to Other Methods

LLMJudge Challenge. We evaluate our approach in the context of the First LLMJudge challenge (July 2024), comparing it against established LLM-evaluation methods such as UMBRELA [34] and RUBRIC [11]. As shown in Table 5, our Multi-Criteria approach (here denoted "4prompts") effectively reproduces the leaderboard derived from manual judgments. Accordingly, we focus on leaderboard correlation metrics Spearman's rank and Kendall's tau.

Our Multi-Criteria approach, based on the smaller LLaMA-3-8B LLM, achieved the highest Spearman's rank correlation and the second-highest Kendall's tau in the LLMJudge Challenge, outperforming all other submissions [24]. These included methods using larger LLMs (e.g., GPT-4, LLaMA-70B, T5), a variety of prompting and fine-tuning strategies, and complex ensemble techniques. This result highlights the effectiveness of our multi-criteria decomposition, demonstrating that breaking down complex information needs can yield strong performance—even with smaller models.

We also compare against UMBRELA [34] through a reproduction on LLaMA-3.3-70B [16] (results in Table 6), and find that Multi-Criteria performs at least as well. Notably, UMBRELA was originally designed for GPT-4o, which is a significantly larger model.

We also explore two variations: a sum-based aggregation of the four criteria grades ("sumdecompose") and a Naive Bayes aggregation ("naiveBdecompose"). Despite relying solely on the smaller LLaMA-3-8B model and no additional fine-tuning, Multi-Criteria methods perform remarkably well compared to other top performing systems that utilize extensive fine-tuning ("setting2") or larger LLMs ("llama70B," "thomas3," "umbrela2"). Recent work [25] shows that ensembles of large LLMs can achieve marginally higher performance.

Human vs. LLM label. Figure 2 shows why our Multi-Criteria approach achieves such a high Spearman rank correlation. Each evaluated system's performance is plotted as a red dot, with the horizontal axis showing scores derived from our predicted relevance labels, and the vertical axis showing scores from manual labels. The strictly ascending line of red dots indicates that both sets of labels yield nearly the same ranking of system quality, corresponding to a Spearman rank correlation of 0.99.

To provide a per-query perspective, each query-system score pair is shown as a blue dot. Notably, some query-system combinations receive zero points under manual assessments but a non-zero score under Multi-Criteria—visible as a horizontal line along the bottom of the figure. This pattern reflects the fact that Multi-Criteria, like many other LLM-evaluation methods, tends to be more lenient than human judgments, a trend also observed in related work [23].

Table 6: Evaluation metrics across datasets and our approaches compared to established baselines Thomas, Faggioli, and RUBRIC. Systems are evaluated with `trec_eval` metrics NDCG@10, mean-average prediction (MAP), and mean reciprocal rank (MRR); leaderboard correlation measured in Spearman's rank (S) and Kendall's Tau (T). - marks unavailable results.

	LLMJudge				dl2020				dl2019				Wins
	NDCG@10		MAP	MRR	NDCG@10		MAP	MRR	NDCG@10		MAP	MRR	
LLaMA-3-8B	S	T	S	S	S	T	S	S	S	T	S	S	
Multi-Criteria (ours)	**0.994**	**0.952**	0.968	0.874	0.972	0.871	0.617	0.626	**0.972**	0.871	0.349	0.852	3
sumdecompose (ours)	**0.990**	0.936	**0.972**	**0.960**	**0.983**	**0.905**	0.879	**0.910**	**0.972**	0.865	0.671	**0.943**	**7**
Multi-Criteria (TCF) (ours)	**0.992**	**0.950**	**0.972**	0.953	0.979	0.885	0.706	0.780	**0.979**	**0.891**	0.454	0.938	5
Thomas [32]	0.987	0.915	0.964	0.833	0.973	0.870	**0.917**	0.891	**0.971**	0.879	0.549	0.898	2
Faggioli [13]	0.975	0.892	0.963	0.943	**0.982**	**0.909**	0.902	0.866	**0.974**	0.874	**0.750**	0.909	4
UMBRELA [16]	**0.989**	0.931	-	-	0.973	0.870	-	-	**0.975**	**0.894**	-	-	
FLAN-T5-large													
Multi-Criteria (ours)	**0.988**	**0.924**	**0.971**	**0.942**	0.965	**0.867**	0.919	0.909	**0.987**	**0.923**	**0.880**	0.844	**8**
sumdecompose (ours)	0.978	0.884	0.953	0.923	0.955	0.851	0.917	0.912	**0.985**	**0.920**	**0.881**	**0.889**	4
Thomas [32]	**0.988**	**0.924**	0.748	0.643	0.936	0.810	0.828	0.751	0.960	0.833	0.341	0.755	2
Faggioli [13]	0.984	0.912	0.965	0.938	0.966	0.861	0.922	**0.940**	0.968	0.875	0.864	0.810	1
RUBRIC [11, 15]	**0.983**	0.898	0.964	0.787	**0.974**	**0.875**	**0.946**	0.845	0.961	0.848	**0.882**	0.795	5
UMBRELA [16]	0.971	0.868	-	-	0.880	0.717	-	-	0.964	0.862	-	-	
LLaMA-3.3-70B													
Multi-Criteria (ours)	**0.994**	**0.952**	**0.968**	0.911	0.988	0.927	0.919	0.921	0.967	0.856	0.762	0.881	3
sumdecompose (ours)	0.988	0.939	0.963	0.911	**0.992**	**0.938**	**0.937**	**0.965**	0.965	0.856	0.814	0.915	4
Multi-Criteria (TCF) (ours)	**0.994**	**0.953**	**0.969**	0.920	**0.991**	0.928	0.922	0.910	**0.975**	**0.879**	0.770	0.896	**6**
Thomas [32]	**0.991**	0.939	**0.970**	0.902	0.922	0.929	0.821	0.887	0.969	0.863	**0.873**	**0.925**	4
Faggioli [13]	0.985	0.912	**0.968**	**0.958**	0.980	0.901	0.857	**0.963**	0.927	0.767	0.703	0.857	2
UMBRELA [16]	**0.993**	0.946	-	-	0.986	0.911	-	-	0.970	0.869	-	-	

DL20 and DL19 datasets. Table 6 compares performance across multiple datasets, evaluation metrics, LLMs, and leaderboard correlation measures. We evaluate our methods against established prompt-based baselines, including Thomas [32], Faggioli (B) [13], RUBRIC [11, 15], and UMBRELA [16, 34].

To enable a fair comparison with UMBRELA, which was designed for the much larger GPT-4o, we use reproduction results [16] on the same LLaMA LLMs. While UMBRELA performs competitively on DL19 with LLaMA-3-8B, it struggles to maintain consistent performance across datasets when applied to FLAN-T5-large, where other methods show greater robustness.

Across 36 comparisons, the Multi-Criteria approach with prompt-based aggregation achieves the highest score more often than any other method—14 times—demonstrating particular strength on the LLMJudge dataset. On the simpler, earlier TREC Deep Learning (DL) collections, the "sumdecompose" variant is a strong contender.

In some cases, particularly on LLMJudge and DL19, the TCF variant, which uses only Topicality, Coverage, and Contextual Fit while skipping Exactness, achieves better results. Faggioli's yes or no prompt is notably effective under mean-reciprocal rank (MRR), whereas RUBRIC excels on the DL20 dataset and also performs well on LLMJudge when evaluating Spearman's rank with an NDCG@10 leaderboard.

4.3 RQ2: Impact of the Underlying LLM

A key goal of LLM-evaluation is to perform well even with smaller LLMs, such as FLAN-T5-large and LLaMA-3-8B, which can be run on more affordable GPUs and require less inference time per query. As shown in Table 6, our Multi-Criteria approach consistently delivers strong performance across a range of model sizes and datasets.

Table 7: Ablation study comparing different subsets of the four proposed criteria across three datasets and two LLMs, using prompt-based aggregation. Differences within 0.005 of the best score are considered non-significant (shown in bold), while scores that fall more than 0.025 below the best are marked in red. While some subsets occasionally outperform the full set, none show consistently better results. Overall, using all four criteria is the most reliable, falling behind in only 2 out of 12 experiments.

	LLMJudge		dl2020		dl2019	
	NDCG@10		NDCG@10		NDCG@10	
LLaMA-3-8B	S	T	S	T	S	T
All four	**0.994**	**0.952**	*0.972*	0.871	*0.972*	*0.871*
TCF (-E)	**0.992**	*0.949*	**0.979**	0.885	**0.979**	**0.891**
ECF (-T)	**0.992**	*0.942*	**0.980**	*0.891*	**0.976**	*0.883*
ETF (-C)	**0.993**	*0.949*	0.954	0.836	*0.974*	*0.875*
ETC (-F)	**0.993**	*0.945*	*0.967*	0.854	**0.977**	**0.889**
E	**0.989**	*0.932*	**0.984**	*0.905*	**0.980**	**0.891**
T	**0.994**	**0.955**	*0.967*	0.847	*0.970*	0.863
C	0.987	0.926	**0.984**	**0.909**	*0.962*	0.854
F	**0.990**	*0.935*	**0.984**	**0.911**	**0.975**	*0.875*
LLaMA-3.3-70B						
All four	**0.994**	*0.952*	*0.988*	*0.927*	*0.967*	0.856
TCF (-E)	**0.994**	**0.953**	**0.991**	*0.928*	**0.975**	*0.879*
ECF (-T)	**0.994**	**0.956**	*0.986*	0.911	*0.962*	0.844
ETF (-C)	**0.995**	**0.958**	*0.987*	0.917	*0.958*	0.841
ETC (-F)	**0.994**	*0.950*	*0.986*	0.911	*0.971*	0.871
E	*0.987*	0.922	**0.995**	**0.952**	*0.974*	*0.883*
T	**0.993**	*0.950*	*0.972*	0.892	0.947	0.819
C	**0.992**	*0.946*	**0.991**	*0.928*	**0.980**	**0.904**
F	**0.995**	**0.955**	*0.987*	0.918	*0.965*	0.856

Notably, the gains of this approach are more pronounced for smaller models, suggesting that breaking relevance into multiple criteria provides the scaffolding needed to achieve high-quality judgments, typically attainable only by larger models under simpler prompts.

For instance, Thomas' prompt [32] performs well with the larger LLaMA-3.3-70B model but struggles with LLaMA-3-8B. In contrast, our Multi-Criteria approach not only secures first place in the LLM-Judge challenge using smaller LLMs without fine-tuning, but also outperforms heavily fine-tuned methods (e.g., "setting2") and closely approaches the performance of ensemble-based strategies such as LLMBlender [25]. This highlights the effectiveness of multi-criterion prompts in closing the performance gap between small and large language models.

Our focus is on achieving strong results with small LLMs, targeting scenarios where access to powerful GPUs or commercial API services is limited. Nevertheless, we compare the runtime of our Multi-Criteria approach with UMBRELA [34] in two settings: 5 queries × 5 passages and 10 queries × 10 passages. Using LLaMA-3-8B in a chat-completion configuration, Multi-Criteria was 5.44 — 5.45× slower than UMBRELA. In contrast, results for FLAN-T5-large are comparable to UMBRELA, with only a smaller slowdown of 1.24 — 1.33×. The performance difference likely reflects several factors beyond prompt design, including model architecture, inference configuration, and processing overhead that accumulates across multiple query-passage pairs. Runtime could be further reduced through asynchronous parallelization, batching, and inference optimizations—potentially narrowing the gap while preserving output quality.

4.4 RQ3: Different Aspects of Relevance

To study whether all four criteria are necessary, we perform an ablation study in which one criterion is removed at a time, as well as experiments where only a single criterion is used. Table 7 shows these results, with top performers denoted in bold, while significantly worse methods marked in red.

We observe that, apart from two cases, the Multi-Criteria configuration with all four criteria either achieves or ties for the best performance, highlighting its robustness. Occasionally, subsets of three criteria—such as TCF on DL19 and DL20—slightly outperform the full set. Similarly, individual criteria can sometimes excel, for example Exactness with LLaMA-3-8B or Coverage with LLaMA-3.3-70B. Nevertheless, no single criterion or subset is consistently better across all experiments.

To explore how different criteria inter-relate, Figure 3 presents the correlation matrix showing the relationships between Exactness (E), Coverage (C), Topicality (T), and Contextual Fit (F), as well as their alignment with predicted relevance labels (L) and manual judgments (J).

We observe a strong correlations across criteria, their predicted labels, and true judgments, particularly for non-relevant (0) and relevant labels (2 and 3). Notably, a Topicality score of 0 consistently signals non-relevance.

A few exceptions emerge for borderline relevance labels. For instance, label 1 often pairs with a Topicality score of 1 but 0 grades in other criteria, while label 2 can coincide with 1 and 2 grades in some criteria and 2 or 3 in Topicality. This pattern suggests that

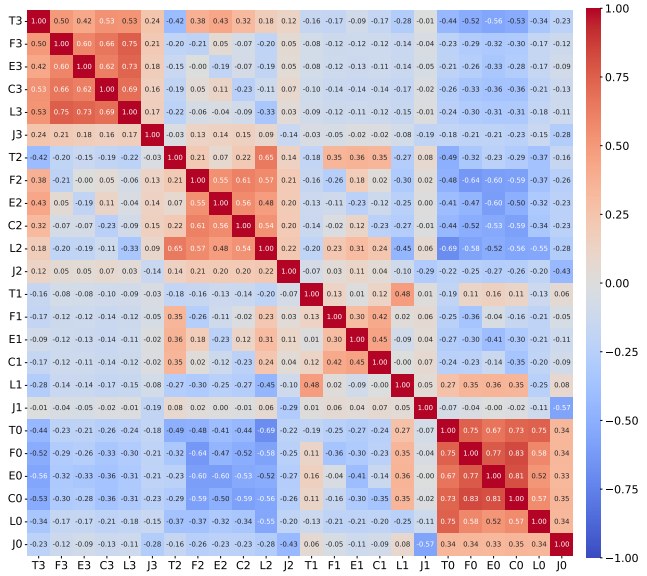

Figure 3: Abbreviations: T = Topicality, F = Contextual Fit, E = Exactness, C = Coverage, J = Ground Truth Relevance Judgment, L = Predicted Relevance Label. Numbers (0-3): indicate criteria-specific grades or relevance labels. Correlation matrix showing pairwise relationships between individual grade levels (0-3) across assessment criteria (Exactness, Coverage, Topicality, Contextual Fit) and their associations with predicted and ground truth relevance labels on LLMJudge dataset using LLaMA-3-8B.

Topicality can serve as a transitional marker between different levels of relevance.

Coverage and Exactness show strong positive correlations across all relevance levels, indicating that documents lacking exactness often fail to provide adequate coverage, yet they remain complementary for borderline relevance labels.

Topicality and Contextual Fit also exhibit meaningful correlations, especially at level 0, suggesting that off-topic documents typically fail to fit the context as well.

Since the used Pearson correlation can hide differences in population sizes, we analyse the most frequent criteria-grade combinations. As expected, the all-zero pattern is most common, reflecting the large proportion of non-relevant labels. The second most frequent pattern assigns a grade of 2 to most criteria, with a topicality score of 3, which aligns with earlier observations.

Overall, 29.6% of cases assign only high criterion grades (greater than or equal to 2), 39.4% assign only low grades (less than or equal to 1), and 31% contain a mix of high and low values (for example, a Topicality grade of 2 with a Contextual Fit grade of 1).

4.5 RQ4: Agreement With Human Judgments

Figure 3 shows that predicted relevance labels have moderate to strong correlations with manual judgments for labels 0, 2, and 3

Table 8: Inter-annotator agreement on Multi-Criteria with LLaMA-3-8B. Bold denotes highest counts per column; counts <10% per row are in light gray. 39% of labels are correct, 38% off-by-one, and Cohen's Kappa of 0 vs rest is 0.3.

Label	Judgments				Total
	3	2	1	0	
3	98	97	116	122	433
2	**243**	**596**	**682**	692	2213
1	26	72	244	409	751
0	10	43	191	**771**	1015

(approximately 0.34, 0.22, and 0.17, respectively), but are weaker for label 1.

Table 8 further indicates that our Multi-Criteria approach is generally more lenient than human annotators: when Multi-Criteria predicts a relevance label of 0, manual judgments usually concur, and when human judgments are high, Multi-Criteria also predicts high labels as well. Among 4,412 query-document pairs, there are 1,009 instances where the predicted label differs by at least two levels from the human judgment; in 92.2% of these cases, Multi-Criteria is too lenient. A closer look at zero-labeled documents predicted as label 2 suggests these are borderline on-topic responses that omit important query details (see Section 4.6 for examples). With additional fine-tuning or specialized classification, such mispredictions may be further reduced.

4.6 RQ5: Illustrative Examples

To illustrate the Multi-Criteria approach, Table 4 presents two passages retrieved for a query about lobster preparation, along with their criterion-specific scores (on a 0–3 scale), predicted relevance labels, and manual judgments. According to the output of the Multi-Criteria approach, both passages are broadly on topic (Topicality = 2), but one shows better Coverage by including more cooking-specific content and demonstrates improved Contextual Fit. This subtle difference leads to a higher predicted relevance label, matching the manual scores of 2 and 1, respectively. To clarify the reasoning behind each score, post-hoc rationales were written manually (not generated by an LLM and not part of the prompt output) for each criterion. These rationales highlight how differences in aspects like Coverage can influence overall relevance. This example shows how decomposing relevance into distinct criteria—here, Topicality, Exactness, Coverage, and Contextual Fit—supports a more transparent and human-aligned labeling process. It also facilitates iterative refinement and more nuanced system evaluation.

5 Conclusion

We present the Multi-Criteria approach to LLM-evaluation, which decomposes the notion of relevance into multiple criteria that are individually graded and subsequently aggregated into a single relevance label. In our experiments, we use the criteria Exactness, Coverage, Topicality, and Contextual Fit. This approach achieves the highest Spearman's rank correlation (0.99) in the first LLMJudge challenge of July 2024, outperforming strong contenders such as UMBRELA and RUBRIC.

In our experiments spanning three datasets, three LLMs, and several reference baselines, Multi-Criteria consistently delivers the strongest results. When examining the correlations among individual criteria, predicted labels, and judgments, we find that Topicality is essential but not by itself sufficient. Rather, the final relevance label gains robustness from the complementary nature of the different criteria. Although Multi-Criteria tends to assign higher relevance scores than human annotators, this leniency applies uniformly across all tested systems. For example, when Multi-Criteria is used to rank IR systems by quality, it produces a leaderboard that closely matches the one based on manual judgments. This is evident from the line of red dots in Figure 2, corresponding to a Spearman's rank correlation of 0.99.

One notable advantage of Multi-Criteria is that it not only remains computationally accessible—even with smaller LLMs such as LLaMA-3-8B and FLAN-T5-large—but also delivers consistently strong results. In contrast, many top contenders in the LLMJudge challenge depend on significantly larger models and extensive fine-tuning. By offering an LLM-evaluation approach that performs exceptionally well without requiring expensive hardware, we aim to broaden research access for under-resourced communities and enhance the reproducibility of IR research.

6 Limitations

As with all LLM-based evaluators, our empirical meta-evaluation has several limitations, which we outline using the framework of LLM Evaluation Tropes [12].

First, our meta-evaluation is based on IR systems submitted to the TREC Deep Learning tracks prior to 2024 (*Old System Trope*). The next generation of systems are likely to incorporate LLM-based components within retrieval or generation, potentially resulting in different empirical behavior and introduces a risk of circularity.

Second, because the training data for large language models is not publicly available, it is possible that the LLMs used in our study were exposed to content from the TREC DL test collections. This raises concerns about evaluation signal leakage, as described by the *Test Set Leak Trope*, which could inflate performance estimates.

Finally, LLMs are frequently retrained or updated, which may cause their relevance judgments to shift over time and diverge from those of human annotators. This dynamic is captured by the *LLM Evolution Trope*. We therefore recommend continued meta-evaluation of LLM-based assessors, especially when applied to new domains, model versions, or IR architectures.

Acknowledgments

We thank the organizers of the LLMJudge challenge [26] for the opportunity to participate and their support in sharing the data necessary to produce Table 5.

This material is based in part 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 authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. Can we use large language models to fill relevance judgment holes? *arXiv preprint arXiv:2405.05600*, 2024.

[2] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 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*, SIGIR-AP 2024, page 32–41, New York, NY, USA, 2024. Association for Computing Machinery.

[3] Negar Arabzadeh and Charles L. A. Clarke. Benchmarking LLM-based relevance judgment methods. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, Padua, Italy, July 2025. ACM.

[4] Carol L. Barry and Linda Schamber. Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management*, 34(2):219–236, 1998.

[5] Pia Borlund. The concept of relevance in IR. *Department of Information Studies, Royal School of Library and Information Science*, 2003.

[6] Bert Boyce. Beyond topicality: A two stage view of relevance and the retrieval process. *Information Processing & Management*, 18(3):105–109, 1982.

[7] Charles L. A. Clarke and Laura Dietz. LLM-based relevance assessment still can't replace human relevance assessment. In *Proceedings of the 10th International Workshop on Evaluating Information Access (EVIA 2025), a Satellite Workshop of the NTCIR-18 Conference*, Tokyo, Japan, June 2025. National Institute of Informatics.

[8] E Cosijn and P Ingwersen. Dimensions of Relevance. *Information Processing and Management*, 36:533–550, 2000. Series Number: 4.

[9] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the TREC 2020 deep learning track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC 2020)*, volume 1266 of *NIST Special Publication*, Online (virtual), November 2020. National Institute of Standards and Technology.

[10] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the TREC 2019 deep learning track. In *Proceedings of the Twenty-Eighth Text REtrieval Conference (TREC 2019)*, volume 1250 of *NIST Special Publication*, Gaithersburg, MD, USA, November 2019. National Institute of Standards and Technology.

[11] Laura Dietz. A workbench for autograding retrieve/generate systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1963–1972, 2024.

[12] Laura Dietz, Oleg Zendel, Peter Bailey, Charles Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. Principles and guidelines for the use of llm judges. In *Proceedings of the 11th ACM SIGIR / The 15th International Conference on Innovative Concepts and Theories in Information Retrieval*, 2025.

[13] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, N. Kando, E. Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment. *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, 2023.

[14] Naghmeh Farzi and Laura Dietz. EXAM++: LLM-based Answerability Metrics for IR Evaluation. In *Proceedings of LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval*, 2024.

[15] Naghmeh Farzi and Laura Dietz. Pencils Down! Automatic Rubric-based Evaluation of Retrieve/Generate Systems. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 175–184, Washington DC USA, August 2024. ACM.

[16] Naghmeh Farzi and Laura Dietz. Does umbrela work on other llms? In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, Padua, Italy, July 2025. ACM.

[17] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. Antique: A non-factoid question answering benchmark. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*, page 166–173, Berlin, Heidelberg, 2020. Springer-Verlag.

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

[19] Richard McCreadie and Cody Buntain. Crisisfacts: building and evaluating crisis timelines. In *Proceedings of the 20th International ISCRAM Conference*, pages 320–339, 2023.

[20] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. Query performance prediction using relevance judgments generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, Washington, DC, USA, July 2024. ACM.

[21] Virgil Pavlu, Shahzad Rajput, Peter B. Golbus, and Javed A. Aslam. IR system evaluation using nugget-based test collections. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 393–402, Seattle Washington USA, February 2012. ACM.

[22] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. Initial nugget evaluation results for the trec 2024 rag track with the autonuggetizer framework, 2024.

[23] Hossein A. Rahmani, Varsha Ramineni, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Towards understanding bias in synthetic data for evaluation, 2025.

[24] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. Judging the judges: A collection of llm-generated relevance judgements, 2025.

[25] Hossein A. Rahmani, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. Judgeblender: Ensembling judgments for automatic relevance assessment. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25)*, Sydney, NSW, Australia, April 2025. ACM.

[26] Hossein A. Rahmani, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, Paul Thomas, Charles L. A. Clarke, Mohammad Aliannejadi, Clemencia Siro, and Guglielmo Faggioli. LLMJudge: LLMs for relevance judgments. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, pages 3040–3043, Washington, DC, USA, July 2024. ACM.

[27] Tetsuya Sakai. Swan: A generic framework for auditing textual conversational systems, 2023.

[28] David P Sander and Laura Dietz. Exam: How to evaluate retrieve-and-generate systems for users who do not (yet) know what they want. In *DESIRES*, pages 136–146, 2021.

[29] Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestations of relevance. *J. Am. Soc. Inf. Sci. Technol.*, 58(13):1915–1933, November 2007.

[30] Ian Soboroff. Don't use LLMs to make relevance judgments. *Information Retrieval Research Journal*, 1(1):10.54195/irrj.19625, March 2025.

[31] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 14918–14937, Singapore, December 2023. Association for Computational Linguistics.

[32] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[33] Together AI. Together inference api, 2024. [Online Service] Available at: https://www.together.ai/.

[34] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. Umbrela: Umbrela is the (open-source reproduction of the) bing relevance assessor, 2024.

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Huai hsin Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 24824–24837. Curran Associates, 2022.

[36] Yunjie Xu and Zhiwei Chen. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science & Technology*, 57(7):961–973, May 2006. Publisher: Wiley-Blackwell.

[37] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Are large language models good at utility judgments? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, pages 1941–1951, Washington, DC, USA, July 2024. ACM.

[38] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda, May 2023. OpenReview.