Does UMBRELA Work on Other LLMs?

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

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

Abstract

We reproduce the UMBRELA LLM Judge evaluation framework across a range of large language models (LLMs) to assess its generalizability beyond the original study. Our investigation evaluates how LLM choice affects relevance assessment accuracy, focusing on leaderboard rank correlation and per-label agreement metrics. Results demonstrate that UMBRELA with DeepSeek V3 obtains very comparable performance to GPT-4o (used in original work). For LLaMA-3.3-70B we obtain slightly lower performance, which further degrades with smaller LLMs.[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. Does UMBRELA Work on Other LLMs?. 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, 9 pages. https://doi.org/10.1145/3726302.3730317

1 Introduction

Relevance labels play a critical role in training and evaluating retrieval systems. Traditionally, these labels have been assigned by human annotators, who assess the relevance of documents or passages to a query based on predefined and provided guidelines. While human annotations have been considered the gold standard, they are resource-intensive and time-consuming. As large language models (LLMs) such as GPT-4o [8] gain prominence, researchers have explored using prompting techniques to predict relevance labels. In this approach, generally a query and passage are input together to an LLM, which generates a response that is parsed and mapped to a relevance label, aiding human annotation with a scalable solution.

Recent research using this technique has proposed methods of varying complexity, ranging from direct prompts for binary relevance assessment [5, 10] to more refined methods where LLMs predict numerical relevance labels [12, 13]. Among these, the UMBRELA framework [13] stands out for its ease of use and effectiveness. UMBRELA employs a zero-shot prompting approach to produce graded relevance judgments. This framework has been widely adopted due to its high correlation with human annotations on TREC Deep Learning datasets. While it is technically compatible with any LLM, it remains an open question whether its effectiveness holds when applied with different models.

System to be reproduced: UMBRELA. The UMBRELA framework (a recursive acronym for **UM**brela is the **B**ing **REL**evance **A**ssessor) [13] is an automated relevance assessment system designed to evaluate search results using large language models (LLMs). At its core, UMBRELA employs a zero-shot DNA (Descriptive, Narrative, and Aspects) prompting [12] approach (as detailed in Figure 1) to generate graded relevance labels. By leveraging an explicitly structured prompt, the system directs the LLM to assess search relevance based on two primary aspects: (1) the alignment between the query's intent and the passage content, and (2) the trustworthiness of the passage. This approach leverages GPT-4o to provide a scalable, cost-effective, and reproducible complement to human relevance assessment.

Experimental results validate UMBRELA's effectiveness, demonstrating high correlations with human judgments on TREC Deep Learning datasets (2019–2023) [2, 3, 9]. With Kendall's τ measurements exceeding 0.8728 across all datasets, UMBRELA exhibits agreement levels akin to human expert assessors. Due to its reliability and performance, UMBRELA has been adopted as an official component of the TREC Retrieval-Augmented Generation (RAG) track in 2024, further underscoring its value in modern information retrieval evaluations.

Because UMBRELA can be used with any LLM, IR research relying on its judgments remains reproducible, even if GPT-4o becomes unavailable, by switching to models with publicly available weights. Additionally, UMBRELA's status as a public domain implementation of DNA prompting approach proposed by Thomas et al. [12], designed to work across LLMs, enhances its accessibility and utility for the IR community.

In this work, we apply UMBRELA's prompt to a range of LLM families and examine whether other LLMs can replace GPT-4o within the framework without performance degradation. While much of the existing research has focused on prompt design, the impact of the underlying LLM is often underexplored. In this work, we explore the connection between the prompt and the LLM family.

Contributions. We evaluate UMBRELA's reproducibility across various LLM families, addressing the following research questions:

- RQ1: Can we replace the LLM in UMBRELA without performance degradation?
- RQ2: To what extent do leaderboard rank correlation and per-label agreement metrics differ in their sensitivity to LLM scale?

[1] **Appendix: https://github.com/TREMA-UNH/appendix-umbrella-other-llm**

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730317

- RQ3: How consistent are UMBRELA's performance patterns across different relevance assessment datasets, and does this consistency vary with LLM scale?
- RQ4: To what extent does investing in larger language models provide meaningful improvements in relevance assessment quality beyond leaderboard rank correlation?
- RQ5: How much does the word choice of the prompt matter in comparison to the choice of LLM?

2 Methodology

2.1 Work to be Reproduced: UMBRELA's Approach

UMBRELA is an LLM-based framework designed to generate graded relevance assessments in information retrieval. It operates through a single, structured prompt, given in Figure 1, which is devised to guide the model in assigning a relevance label to a query-passage pair without requiring explicit examples, few-shot learning, or external annotations.

At the core of UMBRELA is a zero-shot prompting strategy that instructs the model to evaluate the passage's relevance along a predefined four-point scale. The prompt defines relevance categories from 0 (completely unrelated) to 3 (fully relevant), ensuring that each score corresponds to an unambiguous interpretation of relevance. To further reinforce correct label assignment, the prompt explicitly differentiates between borderline cases, instructing the model to use score 1 when the passage is related but does not answer the query, score 2 when the passage contains relevant information but is diluted by extraneous content, and score 3 when the passage is entirely dedicated to answering the query.

A critical aspect of UMBRELA's prompt design is that it implicitly encourages the model to assess relevance from multiple perspectives. Without enforcing a rigid step-by-step reasoning process, it introduces considerations such as query intent alignment. This directs the model to measure how well the passage matches the search intent. It also includes trustworthiness, which prompts the model to factor in the reliability of the passage as a source. These aspects are not explicitly separated into distinct reasoning steps but are rather only guided by the prompt, allowing the LLM to integrate them naturally while producing a final judgment.

A key design choice in UMBRELA's prompt is the requirement of a strict output format, ensuring consistency and machine-readability. The model is instructed to output only a single integer score in the form "`##final score: [score]`" without explanation or additional text. This design choice supports efficient and reliable relevance assessment across large-scale search evaluations.

By combining the state-of-the-art GPT-4o LLM with a carefully optimized prompt, UMBRELA provides a scalable, reproducible, and interpretable method for obtaining LLM-based relevance labels. Its design ensures that the model evaluates information retrieval systems consistently while maintaining flexibility across different datasets, search intents, and language models.

2.2 Reproduction Overview

In this work, we reproduce the **UMBRELA** evaluation framework [13], originally tested only on GPT-4 family models, and extend it to other large and smaller open-source LLMs. To ensure reproducibility, we:

- **Reuse**: We reuse the zero-shot UMBRELA relevance prompt (Figure 1) form the original paper without modification. We also reuse the zero-shot Basic prompt (Figure 2) from the UMBRELA repository. All prompts are sourced from the original GitHub repository.[2]
- **Re-implement**: For the re-implementation of umbrela with other LLMs, we implement our own code to load TREC DL datasets and experiments with different LLMs (both locally

Given a query and a passage, you must provide a score on an integer scale of 0 to 3 with the following meanings:
0 = represent that the passage has nothing to do with the query,
1 = represents that the passage seems related to the query but does not answer it,
2 = represents that the passage has some answer for the query, but the answer may be a bit unclear, or hidden amongst extraneous information and
3 = represents that the passage is dedicated to the query and contains the exact answer.

Important Instruction: Assign category 1 if the passage is somewhat related to the topic but not completely, category 2 if passage presents something very important related to the entire topic but also has some extra information and category 3 if the passage only and entirely refers to the topic. If none of the above satisfies give it category 0.

Query: {query}
Passage: {passage}

Split this problem into steps:
Consider the underlying intent of the search.
Measure how well the content matches a likely intent of the query (M).
Measure how trustworthy the passage is (T).
Consider the aspects above and the relative importance of each, and decide on a final score (O). Final score must be an integer value only.
Do not provide any code in result. Provide each score in the format of: ##final score: score without providing any reasoning.

Figure 1: The zero-shot Bing (UMBRELA) prompt for relevance assessment. All prompts used in this experiment are sourced from the original UMBRELA GitHub repository and presented as-is (see main text for details).

[2] https://github.com/castorini/umbrela/tree/main/src/umbrela/prompts

You are an expert judge of a content. Using your internal knowledge and simple commonsense reasoning, try to verify if the passage is relevance category to the query. Here, "0" represent that the passage has nothing to do with the query, "1" represents that the passage seems related to the query but does not answer it, "2" represents that the passage has some answer for the query, but the answer may be a bit unclear, or hidden amongst extraneous information and "3" represents that the passage is dedicated to the query and contains the exact answer.

Provide explanation for the relevance and give your answer with from one of the categories 0, 1, 2 or 3 only. One of the categorical values if compulsory in answer.

Instructions: Think about the question. After explaining your reasoning, provide your answer in terms of 0, 1, 2 or 3 category. Only provide the relevance category on the last line. Do not provide any further details on the last line.

###

Query: {query}
Passage: {passage}

Explanation:

Figure 2: The zero-shot basic prompt for relevance assessment. All prompts used in this experiment are sourced from the UMBRELA GitHub repository and presented as-is (see main text for details).

via Hugging Face [3] and API-based). We build the full inference and evaluation pipeline.

- **Copy Results**: We copy the GPT-4o results for comparison directly from Table 2 of Upadhyay et al. [13] as the baseline for comparison and keep them unchanged to maintain consistency with the original study.

This setup allows us to validate UMBRELA results and test generalization across models of varying capabilities and access levels.

2.3 Task Description

In line with the original UMBRELA paper, we adopt the same relevance assessment task defined by the TREC Deep Learning (DL) Track. Following the reproducibility goal of our study, we retain this task formulation to ensure comparability with the original results.

> The task involves assigning a relevance label to a query-passage pair using a four-point ordinal scale, as defined by human NIST assessors. The scale ranges from [0] (irrelevant) to [3] (perfectly relevant) and was originally created for the TREC Deep Learning track [3].
>
> - **[3] Perfectly Relevant**: The passage directly answers the query with exact information.
> - **[2] Highly Relevant**: The passage contains some answer to the query, but the answer may be unclear or embedded in extraneous information.
> - **[1] Related**: The passage is somewhat related to the query but does not provide a direct answer.
> - **[0] Irrelevant**: The passage has no relevance to the query.

These labels represent the ground truth annotations provided by human assessors and serve as the expected outputs for the relevance prediction task. The evaluation is based on how closely model-predicted labels match these human-annotated judgments.

While this formulation is designed for the TREC DL task, applying our approach to other evaluation settings,such as conversational relevance or multi-turn QA, requires adapting the prompt to reflect the specific criteria and label definitions of those tasks.

Table 1: Dataset Statistics

	Systems	Queries	Relevance Label Counts 0	1	2	3
DL 2019	36	43	5158	1601	1804	697
DL 2020	59	54	7780	1940	1020	646
DL 2023	35	25	2005	1233	808	377

2.4 Datasets

This framework is evaluated on the same year **TREC Deep Learning datasets** used in the original UMBRELA evaluation [13].

DL 2023 (primary dataset): Systems submitted to TREC DL 2023 with duplicates removed [9].[4]
We primarily focus our analysis on the DL 2023 dataset, as it represents the most recent public test collection and minimizes potential data leakage during LLM training [1].

DL 2020: Systems for passage retrieval from the TREC Deep Learning Track 2020 [2].

DL 2019: Systems for passage retrieval from the TREC Deep Learning Track 2019 [3].

Statistics about these datasets, including the number of submitted systems, queries, and the distribution of label statistics, are provided in Table 1.

This experimental setup allows us to compare our results with those reported in the original paper.

2.5 LLM Selection and Inference Setup

We select models from different families to study the effects of using the UMBRELA evaluation framework when GPT-4o is replaced.

[3] https://huggingface.co/

[4] Cleaned DL 2023 data available at https://llm4eval.github.io/LLMJudge-benchmark/

Table 2: Comparison of UMBRELA variants across different LLMs. Results are based on Cohen's κ scores (four-scale as well as 01-vs-34 and their correlation with ground-truth judgments. Corelation metrics Spearman's rank (ρ) and Kendall's Tau (τ) are used to compare the leaderboards. Following track guidelines, the evaluation uses nDCG@10 as the primary system evaluation metric. An asterisk (*) denotes results reported in the original paper. The DL 2023 dataset used in our experiments corresponds to the LLMJudge challenge data [9]. All approaches use Zeroshot Bing prompt (Figure 1) unless otherwise noted. The bottom provides additional results from the literature for comparison. Best results (or equivalent) are marked in bold.

	DL 2019				DL 2020				DL 2023			
	Rank Correlation		Cohen κ		Rank Correlation		Cohen κ		Rank Correlation		Cohen κ	
	ρ	τ	scale	binary	ρ	τ	scale	binary	ρ	τ	scale	binary
GPT-4o*	**0.974**	**0.893**	**0.361**	0.499	**0.992**	**0.943**	**0.351**	**0.450**	0.985	0.911	**0.308**	**0.418**
DeepSeek V3	**0.978**	**0.898**	**0.360**	**0.518**	**0.993**	**0.940**	**0.358**	0.426	**0.989**	0.929	0.262	0.371
LLaMA-3.3-70B	0.970	0.869	0.266	0.447	0.986	0.911	0.247	0.344	**0.993**	**0.946**	0.233	0.391
LLaMA-3-8B	**0.975**	**0.894**	0.108	0.244	0.973	0.870	0.115	0.187	**0.989**	0.931	0.187	0.315
FLAN-T5-large	0.964	0.862	0.050	0.292	0.880	0.717	0.036	0.231	0.971	0.868	0.062	0.209

GPT-4o (used in original study): Proprietary model offered via OpenAI's API.[5] The architectural leak reported by SemiAnalysis estimates GPT-4 as a MoE model with 16 experts, 2 active per token, each $\approx 111B$ parameters. We include it as the reference point for comparison because it was used in the original study, and based on this reproduction, we would advise users to use this model with UMBRELA.

DeepSeek V3: Recently released open-source model with reasoning enhancements distilled from DeepSeek-R1 [4], known to be trained with data generated by GPT-4o, making it a strong publicly available contender.[6] The total model size is $\approx 145B$, with MoE architecture implied; active parameter usage per token is about 2.8B.[7]

LLaMA-3.3-70B: Large-scale open-source model, included to test performance at higher capacity.[8] A dense transformer model with 70B parameters, no experts. This model is developed independently from GPT-4o, ensuring distinct training processes.

LLaMA-3-8B: Open-source model with medium-scale deployment constraints (8B), chosen for its instruction-tuning and accessibility for resource-constrained settings and because it is often compared to LLaMA-3.3-70B [11].[9]

FLAN-T5-large: Open-source, small-scale LLM with 783M parameters. We included it to test the limits of how lightweight a model can be while still supporting reliable evaluation under the UMBRELA framework.[10]

All models were prompted using the same templates to ensure consistent evaluation.

The original UMBRELA evaluation used OpenAI's latest GPT-4o model, accessed via Microsoft Azure, to assign relevance labels. For reference, we use the results from GPT-4o as provided in the original UMBRELA paper.

For smaller models, such as LLaMA-3-8B and FLAN-T5-large, we run experiments locally on an NVIDIA A40 GPU. For larger models, including DeepSeek V3 and LLaMA-3.3-70B, we use the Together API[11] for inference. For all models, during inference, we use deterministic decoding by disabling sampling (temperature = 0), ensuring consistency across runs, similar to the original study. Each input was processed individually (batch size = 1) for simplicity and consistency.

2.6 Evaluation Metrics

We use the following evaluation metrics to assess model performance, including both leaderboard rank correlation metrics and per-label agreement.

Evaluation measures for comparing the leaderboards resulting from the NDCG@10 evaluation measure under manual judgments and each of the UMBRELA variants:

Spearman's rank correlation (ρ): Measures the difference in systems' rank across the two leaderboards, taking into account each system's exact rank position.

Kendall's tau (τ): Assesses the consistency of pairwise system orderings between the two leaderboards [7].

Evaluation measures based on per-label agreement:

Cohen's κ, 4-point scale (scale): Measures the number of times human judges and UMBRELA agree on the exact label (0–3) for all passages in the judgment pool, corrected for chance agreement.

Binarized Cohen's κ (binary): The same measure, but with the graded relevance labels binarized: labels 0 and 1 are considered non-relevant, while labels 2 and 3 are considered relevant.

Significance. While significance tests do not apply to these measures, we assume that a difference of ≤ 0.005 for τ and ρ and ≤ 0.01 for Cohen's κ is not significant. The best values and equivalently good results are highlighted in bold.

[5] https://platform.openai.com/docs/models
[6] https://www.together.ai/models/deepseek-v3
[7] https://milvus.io/ai-quick-reference/what-is-the-parameter-count-of-deepseeks-r1-model
[8] https://api.together.xyz/models/meta-llama/Llama-3.3-70B-Instruct-Turbo
[9] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[10] https://huggingface.co/google/flan-t5-large
[11] https://www.together.ai/

3 Results

3.1 Comparison of Different LLMs in the UMBRELA Framework

First, we address the most pressing question: can we replace the GPT-4o LLM in the UMBRELA framework and still expect equally good results? This is formalized in the following research question:

> **RQ1: Can we replace the LLM in UMBRELA without performance degradation?** — Finding: Technically yes, but its reliability varies with model scale.

Different LLMs using the original UMBRELA prompt (Figure 1) are compared in Table 2. We find that the GPT-4o model usually performs much better than other LLMs, with DeepSeek V3 showing very similar performance, as the second-best performer. In DL 2023, DeepSeek V3 performs slightly worse than GPT-4o in Cohen's κ but outperforms it in τ and ρ.

The overall third-best system is UMBRELA with either LLaMA-3.3-70B or LLaMA-3-8B, though their per-label agreement metrics (κ) are consistently lower. However, the impact on the leaderboard ranking is small, and leaderboard rankings of both LLaMA LLMs closely resemble the original leaderboard based on human judgments.

We note that the smallest (and fastest) LLM, FLAN-T5-large, despite lower per-label agreement, still obtains strong performance in terms of leaderboard rank correlation.

3.2 Leaderboard vs. Per-label Agreement

Per-label agreement measures and leaderboard rank correlation measures reflect different evaluation goals.

> **RQ2: To what extent do leaderboard rank correlation and per-label agreement metrics differ in their sensitivity to LLM scale?** — Finding: While all LLMs produce leaderboards closely resembling those of human judges, differences are more noticeable in per-label measures.

In general, we find that all evaluation measures are in agreement on the best LLM-Judge systems (DeepSeek V3 and GPT-4o), runner-up (LLaMA-3.3-70B), and weaker systems (LLaMA-3-8B and FLAN-T5-large), as depicted in Figures 3 and 4 and Table 2.

UMBRELA on GPT-4o. On the DL 2023 collection, UMBRELA with GPT-4o achieves a Cohen's κ of 0.31 on the four-point relevance scale. This aligns with results from the LLM Judge Challenge at SIGIR 2024 [9], where UMBRELA variants reported scale-level κ scores of 0.29 and 0.28. Similarly, our observed binary κ of 0.42 is consistent with the 0.40 and 0.43 obtained by UMBRELA variants at the LLM Judge challenge.

Scale Sensitivity of Per-Label Agreement. Looking at how the UMBRELA prompt performs across models of different scales in Table 2, we observe distinct patterns in how per-label agreement (Cohen's κ) and rank correlation metrics (Spearman's ρ, Kendall's τ) respond to changes in model size. The per-label agreement metrics benefit more substantially from increased model scale than the leaderboard rank correlation measures. This is particularly evident when comparing the largest models (GPT-4o and DeepSeek V3) to smaller ones like FLAN-T5-large.

For instance, in DL 2023, the UMBRELA prompt on GPT-4o obtains a four-point scale Cohen's κ of 0.308, which drops dramatically to 0.062 for FLAN-T5-large, representing an 80% decrease. In contrast, the rank correlation metrics are remarkably robust: ρ decreases only from 0.985 to 0.971 (a 1.4% drop), and τ from 0.911 to 0.868 (a 4.7% drop). This pattern is consistent across all three datasets, DL 2019, DL 2020, and DL 2023.

Binary vs. Scale κ. Among the per-label measures, the binary version of Cohen's κ consistently yields higher values due to a greater likelihood of chance agreement. However, this gap varies with model scale—larger models like GPT-4o and DeepSeek V3 show smaller differences between their binary and scale κ scores compared to smaller models. This indicates that increased model scale particularly benefits fine-grained judgment capabilities, especially in distinguishing between non-relevant labels (0 and 1) and relevant labels (2 and 3).

These findings suggest that when accurate per-label judgments are required—such as for generating synthetic training sets or fine-grained evaluations—UMBRELA should only be used with larger models like GPT-4o and DeepSeek V3.

3.3 Generalization Across Datasets

Next, we examine whether the findings generalize across datasets.

> **RQ3: How consistent are UMBRELA's performance patterns across different relevance assessment datasets, and does this consistency vary with model scale?** — Finding: Larger models like GPT-4o and DeepSeek V3 achieve the best results across all datasets.

Examining Table 2, we observe that UMBRELA's performance exhibits meaningful patterns of consistency across the three datasets DL 2019, DL 2020, and DL 2023, though with some notable variations.

For larger models like GPT-4o and DeepSeek V3, performance remains relatively stable across datasets. GPT-4o maintains strong binary Cohen's κ scores (0.499, 0.450, 0.418) and high rank correlation metrics (ρ ranging from 0.974 to 0.992, and τ ranging from 0.893 to 0.943) across all test collections. DeepSeek V3 performs very similarly, obtaining binary κ scores of 0.518, 0.426, and 0.371 with comparable ρ ranging from 0.978 to 0.993 and τ between 0.898 and 0.940.

LLaMA-3.3-70B usually places third, with the exception of DL 2023, where it outperforms DeepSeek V3 in terms of binary κ and GPT-4o in terms of Spearman's ρ and Kendall's τ. Also, we note LLaMA-3-8B slightly outperforms LLaMA-3.3-70B in terms of rank correlation for DL 2019.

While FLAN-T5-large (the smallest model) usually places last, we find that it outperforms LLaMA-3-8B in terms of binary κ for DL 2019 and DL 2020.

Interestingly, in terms of rank correlation, models across all scales perform remarkably well, resulting in nearly indistinguishable performance. Even for the small FLAN-T5-large, we find that Spearman's rank correlation coefficient ρ on the DL 2023 and DL 2019 datasets is nearly indistinguishable from that of GPT-4o. The exception is FLAN-T5-large's reduced performance on DL 2020

Table 3: Comparison of the suggested UMBRELA prompts to the alternative "Basic Prompt" using the LLaMA-3-8B model. Same information as in Table 2.

	DL 2019				DL 2020				DL 2023 +			
	Cohen κ		Correlation		Cohen κ		Correlation		Cohen κ		Correlation	
	scale	binary	ρ	τ	scale	binary	ρ	τ	scale	binary	ρ	τ
LLaMA-3-8B (Basic)	**0.108**	0.233	0.934	0.786	**0.126**	**0.183**	**0.981**	**0.901**	0.173	0.250	**0.988**	**0.932**
LLaMA-3-8B	**0.108**	**0.244**	**0.975**	**0.894**	0.115	**0.187**	0.973	0.870	**0.187**	**0.315**	**0.989**	**0.931**

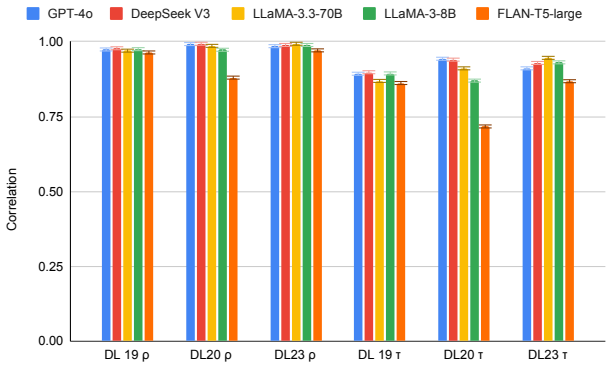

Figure 3: Effect of model scale on rank correlation metrics.

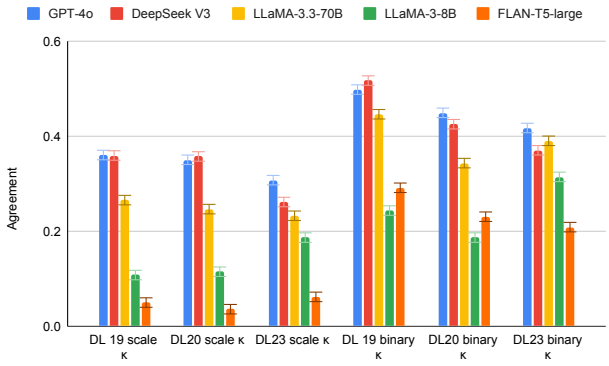

Figure 4: Effect of model scale on per-label agreement.

resulting in $\rho = 0.880$ and $\tau = 0.717$. Notably, shorter prompts [5, 10, 12] perform better for this LLM, achieving ρ of 0.94–0.97 and τ of 0.81–0.88 according to Farzi and Dietz [6].

Future research should explore whether the lower performance on small-scale models LLaMA-3-8B and FLAN-T5-large is inherent to their limited capacity, or whether the UMBRELA prompt is simply not well-suited to smaller LLMs.

3.4 Are Larger LLMs Better?

We assess the extent to which UMBRELA benefits from large-scale LLMs.

> **RQ4: To what extent does investing in larger language models provide meaningful improvements in relevance assessment quality beyond leaderboard rank correlation?** — Finding: UMBRELA performs better with larger LLMs, especially when per-label agreement is important.

Across the board, UMBRELA on any model achieves strong results above 0.85 in terms of rank correlation metrics (ρ and τ). This is encouraging, as it implies that all model variants can differentiate between top-performing and lower-performing systems. This finding even holds for UMBRELA with the small-scale FLAN-T5-large (except on DL 2020). Hence, for some evaluation tasks, smaller and more efficient models may suffice.

However, if the goal is to derive a synthetic training set, then per-label agreement measures become more important. In this case, GPT-4o and DeepSeek V3 should be preferred, as depicted in Figure 4.

Table 4: Invalid output percentages for models across datasets using UMBRELA prompt

Model	Invalid Outputs (%)
DeepSeek V3	0.00
LLaMA-3.3-70B	0.00
LLaMA-3-8B	1.59–1.87
FLAN-T5-large	0.00–0.05

3.5 Choice of Prompt

Much research has focused on tuning the LLM prompt, with less attention given to designing prompts tailored to specific LLM families.

> **RQ5: How much does the word choice of the prompt matter in comparison to the LLM?** — Finding: We find that slight differences in prompt wording do not lead to noticeable quality changes.

The UMBRELA setup also includes an alternative "Basic Prompt," shown in Figure 2, which is available in the UMBRELA GitHub repository[12] but not discussed in the original paper. The "Basic Prompt" omits instructions on query intent and trustworthiness, and hence serves as a simpler baseline. We compare prompt performance on LLaMA-3-8B, the medium-scale LLM that balances cost and performance. Results in Table 3 reveal dataset-specific differences for these two prompts. We find that both prompts yield comparable performance overall: the UMBRELA prompt performs

[12]https://github.com/castorini/umbrela/tree/main/src/umbrela/prompts

slightly better on DL 2019 and DL 2023, while the "Basic Prompt" performs better on DL 2020. This corroborates our suspicion that shorter prompts may work better with smaller LLMs (cf. discussion in Section 3.3).

We conclude that while prompt wording can yield small differences, the choice of LLM has a greater impact on UMBRELA's performance.

3.6 Output Errors and Handling

It is easy to dismiss the lower meta-evaluation scores as merely the results if a less capable model. However, based on our error analysis, we found that a significant contributing factor to performance differences was models' ability to follow the prescribed output format. Despite explicit instructions to provide relevance judgments in the format "`##final score: [score]`", smaller models like FLAN-T5-large only generated standalone numbers as the relevance label, while medium-scale models like LLaMA-3-8B frequently produced inconsistent outputs—ranging from component scores without mentioning the word "Final Score" (e.g., `M: 3, T: 3, O: 3`) to unformatted numbers. Larger models like LLaMA-3.3-70B occasionally deviated as well, generating step-by-step reasoning or alternative formatting, which necessitated more complex parsing logic and introduced extraction errors.

To extract the final UMBRELA relevance score (0–3) from LLM outputs, we use a simple, rule-based pattern-matching function that checks for specific score formats, including "`##final score: X`", "`O: X`", and a fallback standalone number. The prompt explicitly instructs the model to use the "`##final score: X`" format, and we also accommodate the "`O: X`" format as a secondary fallback. This is based on the prompt's instruction where the final score is referred to as "O", and some models may adopt this shorthand to label the relevance score. Additionally, the function can handles minor formatting variations, such as spacing between "`##`" and "`final score`". If none of the expected patterns are found, a default score of 0 is returned.

Using the UMBRELA prompt, we find 0.00% erroneous output for DeepSeek V3 and LLaMA-3.3-70B, but approximately 1.7% for LLaMA-3-8B, and up to 0.05% for FLAN-T5-large across datasets (see Table 4). See Appendix A for sample LLM outputs. LLMs that are unreliable in instruction following have the potential to negatively impact LLM-based evaluations.

3.7 Worked Example of Different LLMs

We demonstrate how relevance labels are predicted by the UMBRELA prompt with different language models in Table 5. We compare the outputs of four LLMs—LLaMA-3-8B, LLaMA-3.3-70B, DeepSeek V3, and FLAN-T5-large—on query `555530` and passage `58950` from the DL 2020 dataset. According to manually created relevance labels, this passage is "Highly Relevant", corresponding to label code 2, although one could argue that a higher rating might also be justified.

We observe that the models' output formats, and where applicable, their decomposition into sub-scores, vary. Nevertheless, all LLMs are able to identify this passage as relevant for the query. In this case, most LLMs are correctly predicting the relevance label as 2, while FLAN-T5-large predicts it as "Perfectly Relevant" (3).

Table 5: Outputs of different LLMs on the same query and passage pair (qidx=555530, docidx=58950, Relevance Label=2).

Input	**Text**
Query	what are best foods to lower cholesterol
Passage	There are two forms of fiber: soluble and insoluble. Soluble fiber attracts water and turns to gel during digestion. This slows digestion. Soluble fiber is found in oat bran, barley, nuts, seeds, beans, lentils, peas, and some fruits and vegetables. Research has shown that soluble fiber lowers cholesterol, which can help prevent heart disease. Insoluble fiber is found in foods such as wheat bran, vegetables, and whole grains.
Model	**Output**
LLaMA-3-8B	Here are the scores: M: 2 T: 3 O: 2
LLaMA-3.3-70B	##M: 2 ##T: 2 ##O: 2
DeepSeek V3	##final score: 2
FLAN-T5-large	3

4 Conclusion

Our study demonstrates that UMBRELA can operate across different LLMs, though with notable performance variations. GPT-4o and DeepSeek V3 achieve the highest per-label agreement metrics. However, even smaller models (e.g., FLAN-T5-large) produce leaderboards that closely resemble the official rankings. This highlights a key distinction: with the current prompt, leaderboard correlations remain stable across model scales, while per-label agreement metrics are highly sensitive to model scale.

Across all datasets, we find that the large-scale LLMs work better with the UMBRELA prompt. However, we also observe diminishing returns in cost-benefit tradeoffs as model scale increases, suggesting that medium-sized models, such as both LLaMA versions, may offer an optimal balance between performance and efficiency for many applications.

These findings suggest that UMBRELA implementation choices should be guided by specific evaluation goals: for system ranking, smaller models may suffice, while accurate document-level relevance labels require larger models. This nuanced understanding of LLM-based evaluation provides practical guidance for IR researchers seeking to balance computational efficiency with evaluation integrity.

Acknowledgments

We thank the authors of the original UMBRELA framework [13] for providing access to prompts and baseline results. We are also

grateful to the organizers of the TREC Deep Learning Track [2, 3] for making the test collections available, and to the organizers of the LLM Judge Challenge [9] for providing a valuable testbed for experimentation.

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.

A Appendix: Output Examples

In this appendix, we provide examples of output formats from various LLMs where the requested output format was not consistently followed. Despite using the same LLM and prompt, some models produced inconsistent outputs. The examples presented here illustrate some types of inconsistencies observed.

A.1 LLaMA-3-8B

The following examples highlight different output formats generated by LLaMA-3-8B:

- Example 1:
  ```
  M: 3
  T: 3
  O: 3
  ```
- Example 2:
  ```
  Here are the scores:
  M: 3
  T: 3
  O: 3
  ```
- Example 3:
  ```
  1: 2
  ```
- Example 4:
  ```
  0
  ```

A.2 LLaMA-3.3-70B

Examples of inconsistent output formats from LLaMA-3.3-70B include:

- Example 1:
  ```
  ## Step 1: Measure how well the content matches a likely intent of the query (M): 3
  ## Step 2: Measure how trustworthy the passage is (T): 3
  ## Step 3: Final score (O): 3
  ```
- Example 2:
  ```
  ## Step 1: Consider the underlying intent of the search
  The underlying intent of the search is to understand how a bounty hunter generates income.
  ## Step 2: Measure how well the content matches a likely intent of the query (M)
  M: 2
  ## Step 3: Measure how trustworthy the passage is (T)
  T: 2
  ## Step 4: Consider the aspects above and the relative importance of each, and decide on a final score (O) O: 2
  ##final score: 2
  ```
- Example 3:
  ```
  ## M: 3
  ## T: 2
  ## O: 3
  ## final score: 3
  ```
- Example 4:
  ```
  ## Step 1: M = 3
  ## Step 2: T = 3
  ## Step 3: O = 3
  ## final score: 3
  ```

A.3 DeepSeek V3

Although DeepSeek V3 followed the desired format most of the time ("##final score: X"), it also exhibited other output formats, such as:

- Example 1:
  ```
  ##M: 1
  ##T: 1
  ##O: 1
  ```
- Example 2:
  ```
  ##M: 0
  ##T: 1
  ##O: 0
  ##final score: 0
  ```

These examples demonstrate that even with identical prompts and models, the output format can vary significantly, highlighting a challenge in ensuring consistency when using LLMs for evaluation tasks.

References

[1] Charles L. A. Clarke and Laura Dietz. Llm-based relevance assessment still can't replace human relevance assessment, 2024.

[2] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the trec 2020 deep learning track. *arXiv preprint arXiv:2102.07662*, 2021.

[3] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*, 2020.

[4] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Jun-Mei Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shao-Ping Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen-Xuan Yu, Wentao Zhang, X. Q. Li, Xiangyu Jin, Xianzu Wang, Xiaoling Bi, Xiaodong Liu, Xiaohan Wang, Xi-Cheng Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yao Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yi-Bing Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He,

Yukun Zha, Yunfan Xiong, Yunxiang Ma, Yuting Yan, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report. *ArXiv*, abs/2412.19437, 2024.

[5] 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.

[6] 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, 2024.

[7] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

[8] OpenAI. Gpt-4o system card, 2024. arXiv:2410.21276.

[9] 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, August 2024. arXiv:2408.08896.

[10] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agent. *ArXiv*, abs/2304.09542, 2023.

[11] Rikiya Takehi, Ellen M Voorhees, Tetsuya Sakai, and Ian Soboroff. Llm-assisted relevance assessments: When should we ask llms for help? *arXiv preprint arXiv:2411.06877*, 2024.

[12] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940, 2024.

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