When LLM Judges Inflate Scores: Exploring Overrating in Relevance Assessment

Chuting Yu
v.yu@uq.edu.au
The University of Queensland
Brisbane, Australia

Hang Li
hang.li@uq.edu.au
The University of Queensland
Brisbane, Australia

Guido Zuccon
g.zuccon@uq.edu.au
The University of Queensland
Brisbane, Australia

Joel Mackenzie
joel.mackenzie@uq.edu.au
The University of Queensland
Brisbane, Australia

Teerapong Leelanupab
t.leelanupab@uq.edu.au
The University of Queensland
Brisbane, Australia

Abstract

Human relevance assessment is time-consuming and cognitively intensive, limiting the scalability of Information Retrieval evaluation. This has led to growing interest in using large language models (LLMs) as proxies for human judges. However, it remains an open question whether LLM-based relevance judgments are reliable, stable, and rigorous enough to match humans for relevance assessment. In this work, we conduct a study of *overrating behavior* in LLM-based relevance judgments across model backbones, evaluation paradigms (pointwise and pairwise), and passage modification strategies. We show that models consistently assign inflated relevance scores — often with high confidence — to passages that do not genuinely satisfy the underlying information need, revealing a system-wide bias rather than random fluctuations in judgment. Furthermore, controlled experiments show that LLM-based relevance judgments can be highly sensitive to passage length and surface-level lexical cues. These results raise concerns about the usage of LLMs as drop-in replacements for human relevance assessors, and highlight the urgent need for careful diagnostic evaluation frameworks when applying LLMs for relevance assessments. Our code and results are publicly available.[1]

CCS Concepts

• **Information systems** → **Relevance assessment**.

Keywords

LLMs, Overrating Bias, Relevance Judgment, Relevance Cues

ACM Reference Format:
Chuting Yu, Hang Li, Guido Zuccon, Joel Mackenzie, and Teerapong Leelanupab. 2026. When LLM Judges Inflate Scores: Exploring Overrating in Relevance Assessment. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3805712.3809905

[1] https://github.com/chutingyu/Exploring-Overrating

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809905

1 Introduction and Background

Building reliable relevance annotations for evaluation has long been a central challenge in Information Retrieval (IR). Cranfield-style collections [6], such as TREC DL [7, 8], rely on trained assessors to annotate query-document pairs, costing substantial time and effort [22], while crowdsourced annotations often suffer from limited reliability [10, 14, 20]. Large language models (LLMs) offer a potential alternative by enabling relevance judgments to be generated rapidly and at scale.

Recent studies report that LLM-based relevance judgments can closely align with human assessments. Thomas et al. [25] employed real searcher feedback to select LLM–prompt configurations aligned with user preferences, demonstrating performance comparable to human assessors in identifying strong systems and difficult queries. These findings were reproduced and extended by Upadhyay et al. [26], who further released the open-source UMBRELA framework [28]. Upadhyay et al. [27] showed that LLM-generated judgments can replace traditional human labels at the *run level* in the TREC 2024 RAG Track, with limited benefit from additional human-in-the-loop intervention.

While LLMs have shown to be effective annotation tools, their potential shortcomings as relevance assessors remains insufficiently understood. LLM-based judgments raise several issues. For example, Alaofi et al. [1] show that LLMs are vulnerable to keyword stuffing, systematically overrating non-relevant passages that contain query terms by assigning inflated relevance labels. Inflated relevance labels, as compared to humans, have been reported in prior work [18, 27]. This behavior also demonstrates how strong *aggregate* performance can obscure judgment-level errors, which can directly impact downstream evaluation tasks. Moreover, using LLMs as both a retrieval method and as an evaluation (labeling) method introduces circularity, confounding true system improvements with LLM-specific preferences [5]. These issues highlight the potential risks when adopting LLMs as relevance assessors.

Since IR systems are measured over the set of relevance labels for a given set of topics, these labels serve as an upper bound on measurable effectiveness. When relevance judgments are generated by LLMs, this upper bound is implicitly shaped by the models' own judgment behaviors. LLM-based relevance assessments tend to exhibit label inflation (e.g., assigning a document a higher relevance label than that assigned by a human assessor), which may distort evaluation signals and mask genuine differences in retrieval quality.

However, it remains unclear how widespread this behavior is in different model families and evaluation settings.

In this work, we conduct a systematic analysis of overrating behavior in LLM-based relevance judgments across two TREC datasets (DL2019 and DL2020), and four open-source LLMs (Llama-3.2-3B, Gemma-3-4B, Mistral-7B, and Qwen-3-8B), on both pointwise and pairwise evaluation paradigms. We analyze overrating at both the label level, and via token-level confidence, to examine whether it reflects random fluctuations or a systematic model bias. To examine the role of semantic understanding, we further conduct controlled passage-rewriting experiments, varying passage length and syntactic voice[2] while preserving meaning, and experiment with varying the content of non-relevant passages to isolate the effects of lexical overlap and surface relevance cues.

2 Overrating Behavior in LLM Judges

We first analyze overrating in LLM-based relevance assessment in terms of both the labels, and the confidence of the assessment.

Label Overrating. Overrating in LLM-based relevance judgments occurs when LLMs assign quality valuations that exceed the ground truth established by human assessors. In pointwise evaluation (binary and graded), overrating manifests through an upward shift in the distribution of LLM-assigned labels relative to human labels. In pairwise evaluation, overrating is implicit due to the comparative nature of the task. Assuming that graded relevance labels imply a strict human preference (e.g., $2 > 0$), two failure modes emerge: *lack of discrimination*, where the model predicts a tie despite a clear human preference, and *preference hallucination*, where a lower-grade passage is preferred over a higher-grade one.

Model Confidence and Stability. Beyond label inflation, we analyze the token-level confidence on the LLM's predicted label or preference. In the pointwise setting, we determine *confidently incorrect* prediction as cases where the model assigns near-deterministic probabilities (i.e., softmax probabilities approaching 1.0) to incorrect label tokens, which are measured by the normalized logits on the generation of the label. In the pairwise setting, instability manifests where the LLM yields varying preferences under positional swapping (e.g., $A > B$ vs. $B > A$) while still assigning high confidence to the assessments.

2.1 Experimental Setup

We conduct experiments on the TREC Deep Learning (DL) Track passage ranking tasks from 2019-2020 [7, 8], using human relevance labels (0–3) as ground truth. We evaluate four open-weight LLMs: Llama-3.2-3B [13], Gemma-3-4B [24], Mistral-7B [16], and Qwen-3-8B [30] under both pointwise and pairwise settings.

In the pointwise setting, we consider both binary relevance (0 "Non-relevant", 1 "Relevant") and graded relevance (0 "Non-relevant", 1 "Related", 2 "Highly relevant", 3 "Perfectly relevant"). We apply variations of the UMBRELA prompt [28] for all judgment tasks. Each configuration is evaluated in terms of label overrating and model confidence.

[2]i.e., the grammatical relationship between a verb and its subject, determining whether the subject performs (active) or receives (passive) the action.

Pointwise Evaluation. For pointwise judgments, we measure label inflation through two metrics: (1) Overrate Ratio, the proportion of the instances where the LLM assigns a relevance label exceeding the human label; and (2) Mean Bias, the average signed difference between LLM-assigned labels and human relevance labels (LLM minus Human). We additionally report Cohen's κ to analyze agreement with human judgments. These metrics collectively quantify both the magnitude and distributional patterns of label inflation exhibited by different LLMs. To understand model confidence, we binarized both the human and LLM judgments according to the TREC standard.[3] Then, we aggregate the token-level confidence according to the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) classes when comparing the human and LLM judgment pairs. Our intuition is that models may be less confident for incorrect (FP or FN) judgments.

Pairwise Evaluation. For pairwise judgment, passages are grouped per query into three categories: *Best* (top-graded relevant), *Acceptable* (relevant, but not the highest grade), and *Unacceptable* (non-relevant) [2]. We sampled 20 pairs per query from each of the following settings: *Best vs. Acceptable*; *Acceptable vs. Unacceptable*; and *Best vs. Unacceptable*. We treat the higher-class passage as the ground-truth winner, and pairs are judged in *both* presentation orders. We force the LLM to prefer one of the two documents, rather than allowing an explicit tie option. As such, we define a *tie* as the case where a contradiction arises between the *two presentation orderings*, meaning that the judgments cannot be linearized. This indicates that the model cannot reliably distinguish the relevance difference between the passages. Therefore, we interpret such cases as the model treating the two passages as equally relevant.

Pairwise overrating behavior is measured using *Pairwise Accuracy* (Acc), the overall proportion of correctly ordered pairs (including ties), and *Pairwise Accuracy without Ties* (Acc (No Ties)), which excludes tied predictions and evaluates only non-tied comparisons. We also compute the *Tie Rate*, capturing failures to discriminate between passages of unequal quality, and the *Wrong Wins Rate*, measuring cases where the lower-grade passage is explicitly preferred over the higher-grade one in *both* presentations. We report the average token-level confidence on the set of correctly-ordered preferences and the set of incorrectly-ordered preferences.

2.2 Results: Pointwise Judgment

Table 1 demonstrates the pervasive nature of overrating across all models and datasets from our results. We consistently observe that the overrating ratio is substantially higher than the underrating ratio, indicating LLMs rarely assign labels lower than human annotators. Since the two settings use different relevance scales (Section 2.1), we binarized graded labels based on TREC standard to enable a fair comparison; the observed overrating ratio becomes lower than that obtained from direct binary evaluation. This suggests that a substantial portion of overrating errors in the graded setting occurs within the relevant or highly relevant categories. Cohen's κ is consistently lower in the graded setting than in the binary setting, which suggests that LLM judgments are less consistent with human annotations in finer-grained relevance, although agreement

[3]Not relevant: 0 and 1; Relevant: ≥ 2. See: https://trec.nist.gov/data/deep2019.html

	Model	Overrated (%)		Underrated (%)		Mean Bias		Cohen's κ		Avg Confidence			
		Bin.	Grad.	Bin.	Grad.	Bin.	Grad.	Bin.	Grad.	TP	FP	TN	FN
DL19	Gemma	37.81	61.84	3.74	8.36	0.341	0.736	0.246	0.107	0.987	0.989	0.989	0.977
	Llama	18.82	49.72	11.09	12.17	0.077	0.564	0.304	0.163	0.902	0.881	0.958	0.914
	Mistral	21.53	54.23	7.45	10.45	0.141	0.557	0.369	0.150	0.939	0.932	0.951	0.904
	Qwen	20.62	47.02	6.10	10.3	0.145	0.459	0.421	0.215	0.977	0.976	0.991	0.986
DL20	Gemma	44.78	66.51	1.66	5.36	0.431	0.895	0.163	0.088	0.983	0.979	0.986	0.981
	Llama	18.57	51.11	6.21	8.00	0.124	0.691	0.265	0.149	0.908	0.870	0.920	0.871
	Mistral	22.32	55.08	4.35	7.47	0.180	0.646	0.293	0.140	0.941	0.925	0.937	0.910
	Qwen	23.42	44.59	2.79	6.61	0.206	0.516	0.338	0.235	0.977	0.974	0.986	0.968

Table 1: Pointwise overrating and underrating metrics on DL2019 and DL2020 under binary and graded relevance labels. Average Confidence is computed under the graded relevance label setting.

is typically lower in graded scenarios due to the additional freedom afforded by having more relevance levels.

Examining the label transition matrices in the graded setting also reveals that the overrating behavior is dominated by soft errors: non-relevant passages are rarely assigned the highest relevance score, but are frequently misclassified as marginally or moderately relevant. This systematic reluctance to assign the non-relevant label inflates relevance estimates under graded evaluation.

We determine whether model confidence (token-level logits) can serve as a reliable proxy for judgment quality in pointwise settings by analyzing the average confidence scores across TP, TN, FP, and FN predictions. Ideally, a reliable model should exhibit high confidence for correct predictions and noticeably lower confidence for errors. Our results indicate a profound disconnect between correctness and confidence. Intuitively, TP decisions should carry noticeably higher confidence than FP decisions. However, our results reveal that the confidence gap is negligible or even inverted. We observe a universal tendency towards extreme overconfidence across all models and tasks. In binary scoring, every model exhibits over 95% confidence not only on TP and TN but also on FP and FN; we observe a similar pattern of high confidence in graded setting as showed in Table 1. This phenomenon reveals that, irrespective of the relevance of the passage, the model is almost always "absolutely confident" of its decision.

2.3 Results: Pairwise Judgment

Trend 1: Lack of Discrimination. As shown in Table 2, the "Tie Rate" is substantial for all models, ranging from approximately 24% to over 45%. This indicates order-induced inconsistency (preference flips under swapping), not consistent errors where the model repeats the same wrong choice. Furthermore, we investigated how judgment varies with the different graded pairs, as, intuitively, the *Best vs Unacceptable* pairs should be the easiest to discriminate. However, we observed that all four LLMs tested consistently overrate "Unacceptable" passages, treating them as qualitatively equivalent to the "Best" passages in nearly *one-third* of cases. Moreover, the tie rate increases systematically as the discrimination task becomes more difficult (e.g., *Best vs Acceptable*), reflecting the models' growing inability to distinguish between relevance grades on such higher difficulty tasks.

	Model	Pairwise Metrics (%)				Avg Confidence	
		Acc.	Acc. (No Ties)	Tie Rate	Wrong Wins	Corr.	Incorr.
DL19	Gemma	50.35	89.47	43.72	5.92	0.979	0.963
	Llama	57.59	84.65	31.95	10.44	0.981	0.978
	Mistral	56.50	87.77	35.61	7.87	0.978	0.961
	Qwen	50.58	92.71	45.44	3.97	0.974	0.946
DL20	Gemma	53.78	89.52	39.92	6.29	0.989	0.970
	Llama	60.01	85.46	23.78	10.20	0.982	0.981
	Mistral	61.36	88.19	30.42	8.21	0.984	0.964
	Qwen	54.42	93.08	41.52	4.04	0.982	0.962

Table 2: Pairwise preference-judgment results on DL2019 and DL2020. Average confidence is computed on no-tie cases only for correct and incorrect preferences.

Trend 2: Indecision vs. Hallucination. The most revealing commonality across all models is the source of error. By comparing accuracy and accuracy (no ties) in Table 2, a consistent behavioral pattern emerges: LLMs are accurate when they express a decisive preference, while overrating primarily arises from their tendency to assign equivocal judgments rather than making clear distinctions. Put simply, when ties are excluded, all models exhibit strong reliability. Qwen achieves an impressive accuracy (no ties) of 93% on DL2020, and even the lowest performing model exceeds 85%. These findings suggest that LLMs do not typically hallucinate that a bad passage is better than a good one (as demonstrated by the rather low "Wrong Wins" rate); rather, they fail to perceive the quality gap between the good and bad passages, defaulting to a Tie.

The confidence difference between correct and incorrect judgments is negligible in the pairwise scenario. On DL2019, Gemma's average confidence drops only marginally from 0.979 on correct judgments (Corr.) to 0.964 on incorrect ones (Incorr.); Llama shows an even smaller gap. In other words, LLM confidence is virtually indistinguishable, regardless of judgment correctness.

Analyzing the model confidence on the tied pairs showed that on around 75% of such cases, model confidence exceeded 0.95 on *both presentation orderings* even though the preferences were contradictory, indicating a tendency for the LLMs to express overconfidence even when their judgments are logically inconsistent. Overall, these

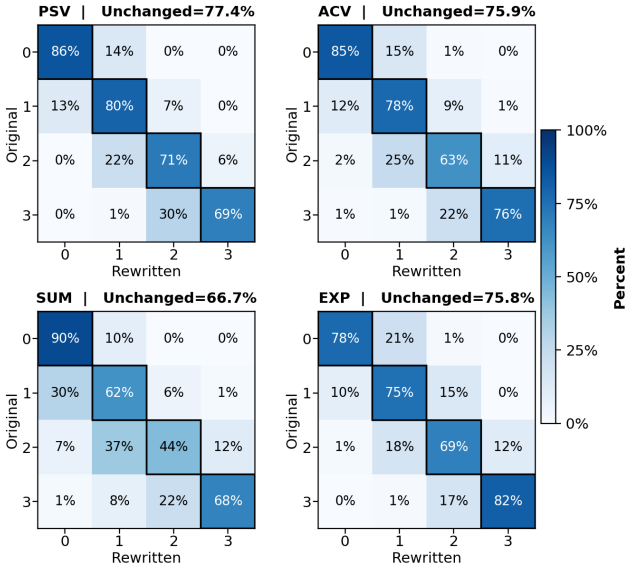

Figure 1: Label transition matrices for Qwen under different passage rewrites. Each heatmap shows normalized transitions from the LLM's original assigned relevance labels.

results demonstrate that using confidence as a signal for judgment reliability is not useful, and is highly sensitive to positional changes.

3 Exploring Relevance Cues

The overrating behaviors observed earlier raise a key question: do LLMs genuinely assess semantic relevance, or are their judgments disproportionately driven by surface-level lexical and structural cues? We evaluate robustness to semantically equivalent structural variations (syntactic and length changes), where label divergence indicates reliance on surface cues. We then introduce controlled, query-dependent sentence insertions into non-relevant passages to disentangle semantic relevance from lexical overlap.

3.1 Semantic-Preserving Structural Variations

We evaluate judgment consistency across four meaning-preserving variants of each passage to quantify sensitivity to structural form.

Syntactic Variation. LLMs may favor scientific writing styles [11], where passive constructions are more common [12]. To test whether such stylistic cues influence relevance judgments, we rewrite passages using passive (PSV) and active (ACV) voice, following expert recommendations favoring the active voice for clarity [17].

Length Variation. LLMs are known to prefer longer responses regardless of quality [21, 31], and input length shifts LLM label distributions in model-dependent ways [19]. We generate summarized (SUM) and expanded (EXP) passage variants while ensuring semantic equivalence. Comparing relevance scores across these variants reveals whether passage length is mistakenly treated as a relevance signal.

Generating Structural Variations. Findings from Balog et al. [4] indicate no general bias from LLM-judges to LLM-generated content, even within the same model family. We use Gemini-2.5-Flash

QID: 156493
QUERY: *do goldfish grow*

SEM These orange cyprinids continue to increase in size throughout their lifespan under suitable conditions.

LEX The terms *goldfish* and *grow* are shown together as an example of word grouping in text data.

QRY do goldfish grow.

Figure 2: SEM, LEX, and QRY variants for the query *"do goldfish grow"*. SEM preserves relevance without surface query terms; LEX provides the query terms such that they are not relevant to the information need; and QRY simply adds the query itself.

to generate stylistic and length-based content variations. Unlike Balog et al. [4], who assume rewriting preserves relevance (validated by human assessors [9]), we perform an explicit quality-check by prompting Gemini-2.5-Flash to verify that all rewritten passages preserve the meaning of the original passage. Across all prompts, we retain only the passages identified as "meaning preserved."

Sample Strategy. For each query, we independently sample (at maximum) 20 passages per predicted relevance label (0–3) for each LLM. For each passage, we generate four syntactic variants and assess label consistency across versions to analyze intrinsic model preferences. To avoid redundant rewriting from overlapping samples across LLMs, passages are pooled only at the rewriting stage, while each LLM is re-judged and evaluated solely on its own sampled query–rewritten passage pairs.

Results. Figure 1 shows the results as label transition matrices, capturing the change in judgment due to the four different passage re-writing treatments as compared to the original (unedited passage) LLM judgment. Across all models, passive voice rewrites and active voice rewrites produce nearly identical effects on LLM judgments, indicating no systematic preference for one form over the other. Summarization consistently reduces relevance labels across models, whereas passage expansion does not consistently inflate relevance labels. Passage expansion systematically inflates relevance labels for originally low-relevance passages – but can still degrade labels for highly relevant passages. These results indicate that even when semantic meaning is preserved, LLM-based relevance judgments exhibit a clear tendency to disfavor short passages.

3.2 Controlled Lexical–Semantic Variations

Prior work shows that query terms can cause LLMs to overrate *non-relevant* passages [1]. To disentangle lexical overlap from semantic relevance, we construct three controlled, query-dependent variants, including Semantic-only (SEM), Lexical-only (LEX), and Query-only (QRY), by inserting *a single sentence* at the beginning of passages originally labeled non-relevant by both human assessors and LLMs, where injection has the strongest influence on relevance judgments [23]. Figure 2 illustrates these three variants for the example query "do goldfish grow", where each injected sentence is prepended to non-relevant passages to re-assess their relevance.

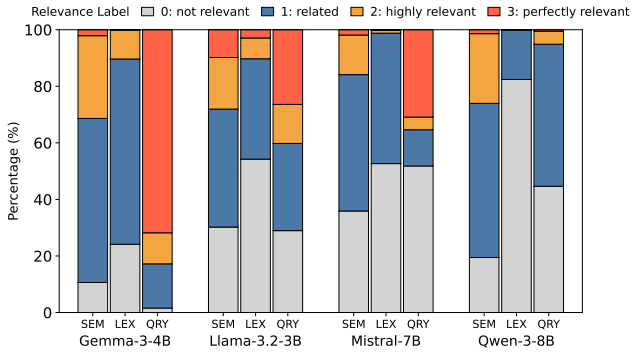

Figure 3: Predicted relevance-label distributions on non-relevant passages after three query-dependent insertions.

For the SEM variant, we prompt Gemini-2.5-Flash to generate a single sentence that directly answers the query while excluding all query terms after stopword removal. When exact synonyms are unavailable, dictionary-style paraphrases are used to preserve meaning without introducing lexical overlap. All generated sentences are manually verified for semantic correctness and constrained to minimal length to avoid confounding passage-length effects.

For the LEX variant, we construct a semantically non-relevant sentence that retains all query terms (excluding stopwords). To strictly control unigram overlap and eliminate unintended changes to the passage meaning, we use a fixed, content-neutral template: "*The terms $w_1, w_2, \ldots, w_n$ are shown together as an example of word grouping in text data.*", where w_k represent the k individual query terms.

In the QRY variant, the inserted sentence is the original query string itself [1]. This design aims to maximize full-query lexical overlap with the non-relevant passage without introducing additional semantic content beyond the query.

Results. Figure 3 reports the label distributions after inserting each variant. Under the SEM condition, the majority of the models assign 1–2 relevance labels, with *perfectly relevant* (label 3) rarely assigned, indicating systematic *under-recognition of semantic relevance* in the absence of lexical cues. Under LEX condition, unigram-level lexical overlap alone is sufficient to elevate relevance labels despite the absence of semantic alignment. This reveals a strong sensitivity to term-level lexical cues. While some models (e.g., Gemma) assign labels comparable to SEM, others (e.g., Qwen) remain largely at label 0, suggesting that query terms are insufficient to override semantic non-relevance for stronger models. The QRY condition yields severe overrating: inserting only query string frequently triggers *perfectly relevant* labels for non-relevant passages which confirms prior findings [1]. Gemma assigns *perfectly relevant* to the majority of non-relevant passages once the full query is inserted, while others (Mistral, Qwen) remain relatively robust. This demonstrates that full-query lexical anchoring can override semantic judgment in most cases. Overall, even with semantic understanding, our experiments show that LLM relevance assessment is disproportionately influenced by lexical cues, confirming strong lexical anchoring effects across models.

4 Conclusion

In this study, we investigated label overrating and model confidence in LLM-based relevance judgments across multiple open-weight LLMs and evaluation paradigms. Across all evaluation settings, we showed that overrating is a system-wide bias, particularly in graded and pairwise evaluations. Overrating is less pronounced in binary relevance judgments, suggesting that LLMs may be better suited for coarse-grained tasks that focus on distinguishing relevant from non-relevant passages. Through experiments on semantically equivalent passage variations, we demonstrate that LLM judges are sensitive to passage length even when semantic content is preserved, while showing comparatively limited sensitivity to syntactic form. Controlled query-related variant experiments indicate that, despite some semantic understanding, LLMs remain strongly influenced by lexical cues, especially direct query-string overlap. These findings underscore the need for diagnostic evaluation frameworks that explicitly probe such biases when applying LLMs for relevance assessment.

5 Future Works

The experimental setup relies only on TREC DL datasets, which primarily consist of short queries paired with relatively short passages. Future work will extend our analysis to investigate how LLM-based relevance judges behave in collections, such as TREC-8 [15] and TREC Robust04 [29], contain substantially longer, multi-topic documents. Another concern is our current study adopts the UMBRELA prompt as a representative instantiation of LLM-based relevance judges. Prior work demonstrates that variations in prompt structure can substantially influence LLM judgment behaviors. Extending beyond a single prompt is therefore essential to establish the robustness of our findings. Future work will construct analysis on multiple prompts spanning reasoning strategies, prompt structures, and persona. These extensions will strengthen the generalizability of our findings and position the proposed analysis as a framework for studying LLM judges, rather than a result tied to a single prompt instantiation.

We acknowledge the concern that modern LLMs may have been exposed to MS MARCO [3] or TREC Deep Learning [7, 8] qrels during pre-training [11, 25], which could inflate apparent agreement with ground-truth labels through memorization rather than genuine judgment capability. Our experimental design partially mitigate this risk. Specifically, the summarization and expansion rewriting generates passages that are unlikely to have appeared in pre-training, and the SEM, LEX, and QRY variants are synthetically constructed rather than directly drawn from the original collection. Hence, the bias patterns we report in this work cannot be fully explained by memorization of specific query-passage pairs. However, we cannot fully rule out that LLMs retain topical or stylistic priors from MS MARCO that bias their judgments in ways not captured by our controlled perturbations. A rigorous contamination audit remains an important future work.

Acknowledgments

We thank the anonymous referees for their insightful comments and suggestions. This work was supported by the UQ HERA initiative and by a Google Research Scholar grant.

References

[1] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2024. LLMs Can Be Fooled into Labelling a Document as Relevant: Best Café near Me; This Paper Is Perfectly Relevant. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '24)*. Tokyo, Japan, 32–41. https://doi.org/10.1145/3673791.3698431

[2] Negar Arabzadeh and Charles L. A. Clarke. 2025. 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, 3194–3204. https://doi.org/10.1145/3726302.3730305

[3] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268v3* (2018).

[4] Krisztian Balog, Donald Metzler, and Zhen Qin. 2025. Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. Padua, Italy, 3865–3875. https://doi.org/10.1145/3726302.3730348

[5] Charles LA Clarke and Laura Dietz. 2024. LLM-based relevance assessment still can't replace human relevance assessment. *arXiv preprint arXiv:2412.17156* (2024). https://doi.org/10.48550/arXiv.2412.17156

[6] Cyril Cleverdon. 1967. The Cranfield Tests on Index Language Devices. *Aslib Proceedings* 19, 6 (06 1967), 173–194. https://doi.org/10.1108/eb050097

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 Deep Learning Track. In *Proceedings of the 28th Text REtrieval Conference, TREC*.

[8] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2021. Overview of the TREC 2020 Deep Learning Track. In *Proceedings of the 29th Text REtrieval Conference, TREC*.

[9] Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024. Neural Retrievers Are Biased Towards LLM-Generated Content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. Barcelona, Spain, 526–537. https://doi.org/10.1145/3637528.3671882

[10] Tadele T. Damessie, Thao P. Nghiem, Falk Scholer, and J. Shane Culpepper. 2017. Gauging the Quality of Relevance Assessments using Inter-Rater Agreement. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Tokyo, Japan, 1089–1092. https://doi.org/10.1145/3077136.3080729

[11] Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '23)*. Taipei, Taiwan, 39–50. https://doi.org/10.1145/3578337.3605136

[12] Henry Watson Fowler. 2015. *Fowler's dictionary of modern English usage*. Oxford University Press.

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024). https://arxiv.org/abs/2407.21783

[14] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating Large Language Models: A Comprehensive Survey. *arXiv preprint arXiv:2310.19736* (2023). https://arxiv.org/abs/2310.19736

[15] David Hawking, Ellen Voorhees, Nick Craswell, and Peter Bailey. 2000. Overview of the TREC-8 Web Track. *Text Retrieval Conference (TREC)* (2000). https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=151494

[16] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023). https://doi.org/10.48550/arXiv.2310.06825

[17] M Yusri Ali Lubis, Reysha Miranti, and Yani Lubis. 2024. Passive voice and active voice in sentence structure. *Journal of Psychology, Counseling and Education* 2, 1 (2024), 59–64.

[18] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2025. Query Performance Prediction Using Relevance Judgments Generated by Large Language Models. *ACM Transactions on Information Systems* (2025), 1–35. https://doi.org/10.1145/3736402

[19] Samaneh Mohtadi, Kevin Roitero, Stefano Mizzaro, and Gianluca Demartini. 2026. The Effect of Document Summarization on LLM-Based Relevance Judgments. In *Proceedings of the 48th European Conference on Information Retrieval (ECIR '26)*. Delft, The Netherlands, 70–87. https://doi.org/10.1007/978-3-032-21300-6_5

[20] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2025. Judging the Judges: A Collection of LLM-Generated Relevance Judgements. *arXiv preprint arXiv:2502.13908* (2025). https://arxiv.org/abs/2502.13908

[21] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076* (2023). https://arxiv.org/abs/2310.10076

[22] Ian Soboroff. 2025. Don't use LLMs to make relevance judgments. *Information retrieval research journal* (2025), 10–54195. https://doi.org/10.54195/irrj.19625

[23] Manveer Singh Tamber and Jimmy Lin. 2025. Illusions of Relevance: Using Content Injection Attacks to Deceive Retrievers, Rerankers, and LLM Judges. *arXiv preprint arXiv:2501.18536* (2025). https://doi.org/10.48550/arXiv.2501.18536

[24] Gemma Team. 2025. Gemma 3. https://goo.gle/Gemma3Report

[25] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models can Accurately Predict Searcher Preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Washington DC, USA, 1930–1940. https://doi.org/10.1145/3626772.3657707

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

[27] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, and Jimmy Lin. 2025. A Large-Scale Study of Relevance Assessments with Large Language Models Using UMBRELA. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR '25)*. Padua, Italy, 358–368. https://doi.org/10.1145/3731120.3744605

[28] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. *arXiv preprint arXiv:2406.06519* (2024). https://doi.org/10.48550/arXiv.2406.06519

[29] Ellen M Voorhees. 2006. Overview of the TREC 2005 Robust Retrieval Track. *Text Retrieval Conference (TREC)* (2006). https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=150643

[30] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, ..., and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025). https://doi.org/10.48550/arXiv.2505.09388

[31] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. New Orleans, USA. https://dl.acm.org/doi/10.5555/3666122.3668142