Principles and Guidelines for the Use of LLM Judges

Laura Dietz
University of New Hampshire
Durham, NH, USA

Oleg Zendel
RMIT University
Melbourne, Australia

Peter Bailey
Canva
Sydney, Australia

Charles L. A. Clarke
University of Waterloo
Waterloo, Canada

Ellese Cotterill
Canva
Sydney, Australia

Jeff Dalton
University of Edinburgh
Edinburgh, United Kingdom

Faegheh Hasibi
Radboud University
Nijmegen, Netherlands

Mark Sanderson
RMIT University
Melbourne, Australia

Nick Craswell
Microsoft
Seattle, WA, USA

Abstract

Relevance judgments for information retrieval (IR) evaluation, once the domain of human assessors, are now often produced by Large Language Models (LLMs). While some studies report alignment between LLM and human judgments, claims that LLMs can replace human judges raise concerns about reliability, validity, and long-term impact. As IR systems increasingly rely on LLM-generated signals, evaluation risks becoming self-reinforcing, leading to potentially misleading conclusions.

This paper examines scenarios where LLM evaluators may falsely indicate success, particularly when LLM-based judgments influence both system development and evaluation. We highlight key risks, including bias reinforcement, reproducibility challenges, and inconsistencies in assessment methodologies. To address these concerns, we propose tests to quantify adverse effects, guardrails, and a collaborative framework for constructing reusable test collections that integrate LLM judgments responsibly. By providing perspectives from academia and industry, this work aims to establish best practices for the principled use of LLMs in IR evaluation.

CCS Concepts

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

Keywords

LLM-Based Evaluation, Validity of Experimentation, LLM Tropes

ACM Reference Format:
Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges. 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, 12 pages. https://doi.org/10.1145/3731120.3744588

ICTIR '25, Padua, Italy

ACM ISBN 979-8-4007-1861-8/2025/07
https://doi.org/10.1145/3731120.3744588

1 Introduction

Large Language Models (LLMs) are increasingly used in the evaluation of information retrieval (IR) systems, generating relevance judgments that were traditionally the domain of human assessors. Given an information need (or topic) and a set of documents, assessors determine the relevance of each document to the topic. A process that forms the foundation of retrieval evaluation. However, due to the vast number of topic-document pairs, traditional assessment relies on pooling methods to identify a subset of documents for judgment. LLMs offer an alternative approach, with the potential to scale relevance assessments far beyond the limits of human annotation. However, Soboroff [75] writes: "Letting the LLM write your truth data handicaps the evaluation by setting that LLM as a ceiling on performance." In this paper we aim for a middle ground by discussing 14 ways in which LLMs can negatively impact the evaluation and how to avoid this adverse effect.

Pro. One of the most impactful advantages of LLM-based evaluation is speed. Unlike human assessors, who require coordination, training, and extensive annotation time, LLMs can generate relevance labels almost instantly. This dramatically lowers the cost of evaluation, making it possible to assess larger datasets, cover a wider range of retrieval tasks, and conduct evaluations more frequently. These benefits have led to the rapid adoption of LLMs in large-scale evaluation pipelines. Microsoft, for example, now uses OpenAI's GPT models for relevance assessment in Bing [79]. More recently, Upadhyay et al. [84] introduced Umbrela, an open-source toolkit based on a similar prompt that uses proprietary LLMs to label unjudged documents. Its application in a recent TREC task, further reinforces the notion that human assessors could be replaced [82]. Beyond labeling, LLMs have been proposed for fully synthetic test collections, where they replace human users in both query generation and relevance judgment [60].

Con. Although empirical studies demonstrate the effectiveness of LLM-based judgments, concerns remain regarding their reliability, validity, and long-term implications for IR evaluation [29]. The increasing reliance on LLMs for test collection creation raises fundamental questions about reproducibility. At the same time traditional pooled judgment methodologies are becoming impractical for assessing generative and multi-modal systems. Furthermore, there is no consensus on how to mitigate the risks associated with these models, including biases, inconsistencies, and potential vulnerabilities such as susceptibility to query stuffing [3]. Without a shared principled framework for responsible adoption, the unchecked use of LLMs in evaluation studies may lead to misleading conclusions, where prior work is cited without appropriate consideration of its limitations.

Below we outline our contributions by specifying questions that are addressed in this paper.

1.1 Can I Use LLM Judgments in My Next Research Paper?

Our short answer:

- **Yes,** if the focus is on improving runtime efficiency.
- **It depends,** if the goal is to improve result quality.

Focusing on quality evaluation, we highlight key validity challenges in using LLMs as evaluators for IR. We introduce a set of LLM Evaluation Tropes in Section 2, a concept that emerged from discussions at the fourth Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025) [80], that capture common failure modes, including circularity and overfitting. We frame these tropes within a broader evaluation paradigm – both with and without human involvement – and consider their downstream effects on systems optimized for LLM-based feedback. We believe that maintaining scientific rigor requires identifying and recognizing, quantifying, and safeguarding against these risks. Hence, we propose best practices, including a set of guardrails for research experimentation,[1] to ensure that LLM-based evaluation remains reliable and meaningful for IR research.

This paper is not intended as grounds for rejecting research that uses LLM-based evaluation. Rather, our goal is to support the community by offering a shared terminology and a set of principles for navigating emerging risks. By making common pitfalls explicit, we aim to foster transparency and shared expectations among authors, reviewers, and organizers. Delaying this conversation increases the risk of inconsistent evaluation practices and unintended consequences as LLM judgments become more widely adopted.

1.2 What is a Valid Experiment?

We take *valid* LLM evaluators to mean evaluators whose measurements align with human intuition and quantify the utility for real people. We follow the guidance from Spärck Jones and van Rijsbergen [76]:

> "It is apparent in particular that it is most important that the ideal collection(s) should be a means of relating **valid abstract studies** of information retrieval and those of operational systems and user behaviour."

In industry, the effectiveness of systems is typically validated through A/B testing and manual assessments. In academia, evaluation helps to determine which approaches constitute research advances and should be submitted to conferences and shared tasks. In both industry and academia, it is essential to obtain quantitative quality measures that credibly reflect relevance while mitigating unintentional biases, such as test data leakage, which may compromise the validity of drawn conclusions.

Figure 1: LLM-evaluation tropes that can lead to invalid conclusions about evaluation, systems, and the efficacy of human judges that oversee the process. Overarching patterns are circularity ↻, Goodhart's law ♡, and loss of variety ≅.

Evaluation Tropes:

↻ **#1 Circularity**: Leaking the evaluation signal into the IR system.

↻ **#2 LLM Evaluator as a Ranker**: Using the same approach in the system and the evaluation.

♡ **#3 LLM Narcissism**: LLMs prefer text from their own model.

≅ **#4 Loss of Variety of Opinion**: When all judges think alike.

Meta-Evaluation Tropes:

♡ **#5 Ignored Label Correlation**: When human and LLM judges disagree on relevance labels.

≅ **#6 Old Systems**: Evaluators need to identify the best systems of the future.

#7 LLM Evolution: LLMs are not static; they can improve or degrade over time.

System Tropes:

♡ **#8 Test Set Leak**: LLMs trained on test collections create the illusion of quality.

↻ **#9 Self-Training Collapse**: Concept drift from training LLMs on LLM output.

♡ **#10 Goodhart-style Overfitting**: Strategically gaming an automatic LLM-based metric.

♡ **#11 Adversarial Threats**: Bad actors want to manipulate the systems and evaluation.

Judge Tropes:

≅ **#12 Rubber-Stamp Effect**: Lack of critical oversight when humans blindly trust LLM labels.

#13 Black-box Labeling: When relevance is complex, labels may be difficult to interpret.

↻ **#14 Predictable Secrets**: When human data can be guessed by an LLM.

In this paper, we discuss conditions under which the use of LLM evaluators may (inadvertently) threaten the validity of the experiment.

1.3 Why Does the Old Evaluation Paradigm No Longer Apply?

The traditional IR evaluation paradigm – rooted in the Cranfield framework [18] and widely used in TREC,[2] CLEF,[3] FIRE,[4] and NTCIR,[5] – assumes that effective systems retrieve a similar set of relevant documents. To approximate a comprehensive relevance set, pooling methods select top-ranked documents from multiple

[1] We hope our recommendations offer a balanced compromise that satisfies both authors and reviewers.

[2] https://trec.nist.gov/

[3] https://www.clef-initiative.eu/

[4] https://fire.irsi.org.in

[5] https://research.nii.ac.jp/ntcir/index-en.html

systems for manual assessment [85]. While this remains foundational, it departs from Cranfield's original setup, which involved judging all documents. The scale of modern web collections makes such exhaustive assessment infeasible. More recently, generative systems have begun to challenge core assumptions. Rather than returning fixed documents, they produce paraphrased or alternative responses, many of which may be equally valid but differ in surface form. A minor change in wording can shift relevance, and manually judging all variations is not practical. Even in non-generative settings, newer retrieval paradigms such as dense retrieval, neural re-ranking, and query reformulation often return unjudged documents that fall outside traditional pools, reducing evaluation completeness [64]. To address these challenges, the IR community has developed approximation strategies and refined its methodologies [33, 50, 53, 54, 66, 68]. Our work extends this prior work by contributing guidance for developing guardrails and validation methods in the context of LLM-based evaluation.

1.4 Outline

We start by exploring different ways in which the use of LLMs in evaluation can inadvertently negatively impact the validity of the evaluation. Section 3 supports this with case studies from two companies who identified and overcame validity issues in their experimentation. We recap a study that demonstrates of how circularity can arise with data from a recent TREC task in Section 4 and suggest an annual effort to cooperatively build test collections with recent systems, evaluators, and content modifiers in Section 5, before concluding the paper.

2 LLM Evaluator Tropes and Guardrails

During system development, various forms of Goodhart's Law [37] and test data leakage may compromise evaluation integrity. A critical hazard is *circularity*, a feedback loop in which evaluator assumptions and system design decisions reinforce one another, distorting evaluation outcomes. As a result, evaluation metrics may no longer reflect human preferences under realistic conditions, thereby invalidating the evaluation paradigm.

We present a taxonomy of recurring tropes observed in LLM-based evaluations (see Figure 1 for an overview). These tropes can negatively affect different stages of the evaluation process, and we discuss several types that pose particular challenges to evaluation validity:

(1) **Eval**: Tropes that lead to misleading or incorrect evaluation measurements of systems.
(2) **Meta-Eval**: Tropes that give the false impression of high evaluation quality or reliability of LLM evaluation approaches.
(3) **System**: Tropes that cause an IR system to perform poorly or unreliably in real-world scenarios.
(4) **Judge**: Tropes that inadvertently undermine the effectiveness of human judgment in the evaluation process.

Below, we examine these common trope patterns, highlighting their pitfalls, and propose guardrails to mitigate their shortcomings.

2.1 Evaluation Tropes

We begin by describing a set of common evaluation tropes that can undermine the validity of LLM-based evaluation systems. These issues arise when the design or application of the evaluation process produces misleading metrics, circular validation, or inflated estimates of system performance.

Eval Trope #1: Circularity

– Leaking the evaluation signal into the IR system. –

Circularity arises when the very LLM judge embedded inside an IR system (for instance as a late-stage re-ranker) is also used as the official evaluator of that system. Because the optimization and assessment objectives are identical, such evaluations can become self-reinforcing rather than genuinely informative [2, 9, 16, 88].

Even when system developers and LLM evaluators work independently, unintentional contamination can occur [29]. A system may unknowingly integrate aspects of an LLM evaluator's methodology, either through training data, algorithmic choices, or shared heuristics. This accidental feedback loop can result in inflated performance in LLM-based evaluations without corresponding gains for real-world applications [9, 65, 67, 73].

Repeated circularity, in which systems are trained and evaluated with signals from the same LLM, raises concerns about model collapse, as captured by the *Self-training Collapse trope* [71].

Quantify effect. The impact of this trope can be assessed by comparing against a manual evaluation paradigm and measuring divergence in leaderboard rankings [10], particularly for systems that may have been influenced by evaluation signal leakage. We present one such study in Section 4.

Guardrail. At a minimum, multiple LLM evaluator paradigms should be included to reduce the risk of accidental circular evaluation. Incorporating fresh human judgments provides an independent check against self-reinforcing feedback loops. However, care must also be taken to avoid the *Predictable Secret* trope.

Eval Trope #2: LLM Evaluator as a Ranker

– Using the same approach in the system and the evaluation. –

This trope describes a specific form of circularity that arises when the same LLM evaluator is used both within the ranking system and as the evaluation metric. It affects any form of self-refinement procedure [92]. In information retrieval, this occurs when the same method is used to generate both the ranking scores and the evaluation scores. The result is a superficial alignment that can produce artificially high evaluation scores – even for systems that perform poorly under human judgment [9, 16, 35].

A similar failure mode can be illustrated with BM25: if the top ten documents retrieved by BM25 were assumed to define the ground-truth relevance, then a BM25 ranker would trivially achieve perfect P@10. Clearly, no one would accept such a circular and invalid evaluation paradigm for measuring the relevance of IR systems.

We recognize that system developers will want to include notions of LLM evaluation in their system [63]. In Section 4 we examine this scenario in the context of the TREC RAG 2024 track, which employed the UMBRELA LLM evaluator. Figure 2 demonstrates that reranking submitted runs using UMBRELA improves performance under manual assessment. This is a valid and actionable finding that supports system-side use of LLM evaluators.

However, evaluating such reranked systems using the same Umbrela metric introduces circularity and **leads to invalid evaluation outcomes**. In our analysis, this reuse of the evaluator causes significant divergence from human assessments: human and LLM evaluators disagree on the relative quality of 18% of system pairs–more than twice as many as found on original systems. Moreover, while twelve systems score above 0.95 in Umbrela-NDCG, their manual NDCG scores range from only 0.68 to 0.72.

Quantify effect. This effect can be quantified by repeating the analysis in Section 4, directly comparing system rankings under LLM and human evaluation.

Guardrail. Avoid using an LLM evaluation procedure if the same (or closely related) procedure may be embedded within the system under evaluation.

Eval Trope #3: LLM Narcissism

– LLMs prefer text from their own model. –

Being language models, LLM evaluators tend to assign higher scores to text that aligns closely with their own generation patterns, as they effectively equate textual quality with per-token likelihood. This leads to a preference for outputs produced by the same model family. For instance, GPT-4 may systematically favor responses generated by GPT-4-based systems, even when human assessors detect no meaningful quality difference [22, 48, 49, 58, 92]. This results in distorted system rankings and compromises the validity of the evaluation outcomes.

Models may be optimized to align with the LLM's biases rather than real-world relevance assessments, undermining the credibility of the evaluation.

Quantify effect. The experimental protocol from Liu et al. [48] can quantify this effect. It involves recording the LLM versions used in both systems and evaluators, and analyzing how often evaluators favor systems built with the same underlying model.

Guardrail. One mitigation strategy is to reserve a specific LLM (or family of LLMs) exclusively for evaluation purposes, ensuring it is not used in any system under test. However, due to overlapping training corpora across models, this bias may still persist. A more robust alternative is to involve multiple LLMs in the evaluation and aggregate relevance judgments using majority voting, while omitting the vote of any evaluator that shares lineage with the system under consideration.

Eval Trope #4: Loss of Variety of Opinion

– When all judges think alike. –

LLM-based evaluations risk homogenizing judgment. Prior work has shown that LLMs can exhibit gender and cultural biases, often reinforcing dominant perspectives while penalizing creative, diverse, or unconventional—yet valid—outputs [4, 12, 57]. In contrast, human assessors are better equipped to recognize nuance, novelty, and contextual diversity, which LLMs frequently overlook [72].

More fundamentally, when LLMs define what is relevant across the board, they implicitly set a ceiling for what systems can achieve. This can penalize systems that offer innovative or non-standard responses that fall outside the LLM's implicit norms [4, 75].

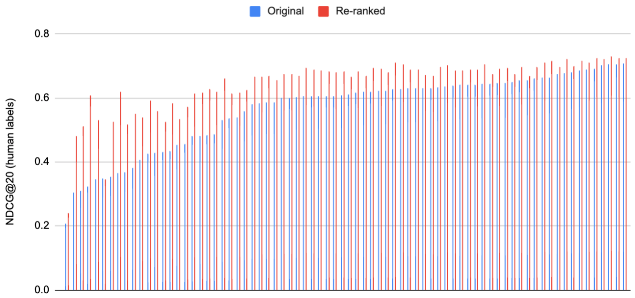

Figure 2: Reranking with an LLM evaluator (Umbrela) improves performance under human relevance labels. This plot compares the original and reranked versions of all TREC RAG 24 systems based on manual assessment.

Quantify effect. This trope's impact can only be assessed through independent evaluations involving human judges from diverse socio-cultural backgrounds.

Guardrail. While human annotation workflows can be designed to ensure a variety of perspectives, achieving this with LLMs is far more difficult. Persona-based prompting strategies [81] have been proposed as a mitigation, but emerging evidence highlights their limitations [15, 28, 42, 43]. We recommend rigorous quantification of this effect before relying on such methods in high-stakes evaluation.

2.2 Meta-Evaluation Tropes

The quality of different LLM-judge approaches is often validated through meta-evaluation – a paradigm that measures how well LLM judgments reproduce either manually created relevance labels or leaderboard rankings under an official evaluation metric [29, 98]. However, such meta-evaluations can foster a false sense of reliability or progress, masking deeper issues in metrics, methodology, or evaluator behavior [16, 99].

Meta-Eval Trope #5: Ignored Label Correlation

– When human and LLM judges disagree on relevance labels. –

Meta-evaluations of LLM-based judges often measure the correlation between system rankings or per-query document rankings derived from evaluation metrics such as NDCG, using either manual or LLM-based relevance labels [56, 84]. However, such high-level agreement can obscure important differences at the level of individual judgments.

For example, in the context of conversational systems, Mehri and Eskenazi [52] demonstrate that even when system-level Spearman correlation is perfect (i.e., $\rho = 1$), agreement on individual relevance labels can vary widely – from as low as 0.12 to 0.61 – depending on the underlying metric. This highlights the risk of relying solely on system-level comparisons.

To establish that LLM-generated judgments are reliable, agreement should be assessed directly at the label level; that is, for each query and document pair. This helps ensure that the evaluation does not limit the measurable performance of newer systems. If the judgments fail to capture certain relevance signals, they may prevent improved systems from being properly recognized as such.

Quantify effect. To assess the reliability of LLM-generated judgments, measure their agreement with human relevance labels directly at the label level. Rahmani et al. [61] propose using Bland-Altman plots [5] to quantify leniency and other biases in label agreement.

Guardrail. Incorporating label-level agreement analysis alongside system-level metrics ensures that inconsistencies or biases in LLM evaluations are not overlooked, providing a more complete view of evaluator reliability.

Meta-Eval Trope #6: Old Systems

– Evaluators need to identify the best systems of the future. –

The primary goal of evaluation is to identify the next generation of state-of-the-art systems. Accordingly, meta-evaluations of LLM-based judges aim to show that these evaluators can correctly identify the best-performing systems. However, this claim is often tested on *legacy systems*, those that were state-of-the-art at the time the test collection was created.

As new IR paradigms emerge, they are rarely reflected in existing test collections. As a result, a meta-evaluation on a dataset can only confirm whether the LLM evaluator recognizes high-performing systems that era.

Yet such studies are frequently used to argue that LLM evaluators will *also* be effective for future systems. This assumption remains untested for future systems, which are expected to differ significantly. Such systems are likely to employ LLMs more extensively, integrate higher-quality models, or adopt innovations that differ significantly from past approaches. There is a real danger that LLM evaluators – especially those themselves were evaluated on outdated LLMs – may fail to recognize these future breakthroughs.

Quantify effect. A simple change in community practice is to collect implementations of the recent IR systems and release an expanded judgment pools (such as suggested in Section 5). By repeating meta-evaluations on these new systems using test collection artifacts, one can assess whether the LLM evaluator still identifies best performing systems.

Guardrail. Older TREC collections remain relevant, because of available manual runs [86]. In addition, the community should regularly re-run meta-evaluations with updated systems to detect and mitigate the effects of this trope.

Meta-Eval Trope #7: LLM Evolution

– LLMs are not static; they can improve or degrade over time. –

Meta-evaluations of LLM-based judges often rely on a single prompt or a single LLM family, despite the wide variety of models available. Crucially, LLMs are not static – model behavior evolves over time as new versions are released [14]. Future iterations of an LLM may judge relevance differently than earlier ones, introducing inconsistencies in longitudinal evaluations. This drift becomes especially problematic when newer models are trained on data that includes outputs from earlier versions, potentially leading to feedback loops and self-training collapse [71].

These issues are compounded by the fact that LLM providers may seamlessly retire older versions or update models without notice.[6] This makes it difficult – or even impossible – to reproduce prior evaluation findings using the same version of the evaluator.

Quantify effect. To track the impact of model evolution, meta-evaluations should be periodically repeated using updated LLM versions. Key indicators of behavioral drift include changes in the ranking of top systems, inconsistencies in relevance labels compared to human judgments, and increased variability in the labeling of previously unjudged documents [1, 6].

Guardrail. Because access to specific model versions cannot be guaranteed over time, LLM-based evaluation methods must be continually re-validated. In Section 5, we recommend that the community adopts a recurring meta-evaluation protocol to ensure the ongoing reliability and relevance of LLM-based evaluators.

2.3 System Tropes

Next, we examine tropes that degrade IR system performance as a result of reliance on artifacts such as LLM-generated relevance labels. While synthetic data and automated evaluators can improve scalability, their improper use can introduce systemic biases and encourage overfitting to unreliable or unstable evaluation signals.

System Trope #8: Test Set Leak

– LLMs trained on test collections create the illusion of quality. –

Some LLMs are trained on publicly available test collections used in IR evaluation [24]. This contaminates evaluation outcomes by inflating the performance of systems that incorporate such LLMs, creating the illusion of high accuracy that fails to generalize to real-world scenarios [91, 97].

There are, however, legitimate reasons to train an LLM on relevance labels, for example, when developing an LLM evaluator specifically designed to support assessment. Prior work, such as the AutoTAR evaluation framework [19, 20], demonstrates that targeted training can yield valid and scalable evaluation systems.

Nevertheless, if such test collections are also used in meta-evaluation, training-induced memorization can create a misleading appearance of alignment between LLM and human judgments, which would undermine the credibility of the evaluation approach [23, 90].

Quantify effect. After collecting fresh manual relevance labels on new topics for the task, a drop in performance on fresh topics would signal potential overfitting or memorization. The protocol of Bordt et al. [13], developed in the context of table learning, provides a useful template for quantifying this effect.

Guardrail. Avoid conducting IR research on test collections likely to have been included in LLM training data, as this risks measuring memorization rather than generalization. Regularly collect fresh human relevance judgments on new topics to track performance drift. A trusted entity, such as a leaderboard system [34], should retain a secret subset of the test collection, to be exposed only via metrics in order to safeguard against future test set leakage.

[6] https://openai.com/index/gpt-4-api-general-availability/

System Trope #9: Self-Training Collapse

– Concept drift from training LLMs on LLM output. –

The increasing use of LLM-generated content as training data for other LLMs raises serious concerns about concept drift and long-term quality degradation [93]. Rather than fostering diversity or nuance, repeated training on synthetic outputs may entrench biases and amplify systematic errors [27, 36, 44, 70].

In the context of IR, this phenomenon arises when LLM-based evaluators are used to generate synthetic training data for IR systems. The problem compounds when the outputs of these synthetically trained systems are then used to fine-tune the next generation of LLM evaluators, forming a recursive feedback loop. This recursive co-training process can amplify subtle biases and lead to concept drift – and, ultimately, model collapse [71]. This is a concrete manifestation of unintended circularity during system development, wherein models achieve high training or evaluation scores but fail to generalize in real-world scenarios.

Quantify effect. This effect can be quantified [70] by tracking evaluation performance on a fixed set of held-out, human-labeled data across multiple rounds of recursive training. A consistent decline in agreement with manual judgments would signal the onset of model collapse.

Guardrail. One should adopt guardrails from Reinforcement Learning from AI Feedback [27] and generative AI [70] to detect when systems are degrading.

To avoid inadvertently exercising this trope, training data should be released with proper documentation of how training data was obtained, and to which extent LLMs were used in the generation.

System Trope #10: Goodhart-style Overfitting

– Strategically gaming an automatic LLM-based metric. –

Goodhart stated that when a metric becomes a target, it ceases to be a reliable measure of success [37]. Goodhart-style overfitting arises when developers iteratively probe or train against synthetic evaluation signals produced by LLM evaluators [10]. Systems that imitate the evaluation labels often record dramatic gains on the chosen metric (e.g., NDCG or an Umbrela score), while stalling – or even regressing – on user-centered outcomes such as click satisfaction, dwell time, and spam resilience [78, 96]. Analogous reward-hacking failures have been documented in RLHF explainability, where models are tuned to produce persuasive yet unfaithful rationales [31]. Not even ensembles of LLM judges are immune: clever systems can overfit to shared blind spots across the LLM evaluator ensemble [26].

Quantify effect. A practical diagnostic is to measure cross-judge volatility: each run is scored with a diverse suite of judges consisting of different models, prompts, and synthetic collections. Large volatility, as in the protocol of Siska et al. [74], signals that a system is overfitting to the evaluation benchmark rather than capturing genuine relevance [10].

Guardrail. Robust evaluations should include a set of human-labeled relevance judgments that remains hidden from system developers and is used periodically to test IR systems (along with LLM evaluators). This setup helps detect cases where systems are narrowly optimized for a specific LLM-derived signal while failing to generalize across other critical evaluation dimensions [94, 95].

System Trope #11: Adversarial Threats

– Bad actors want to manipulate the systems and evaluation. –

Adversarial behavior is an increasing concern [11, 41] as LLMs become central to both retrieval and evaluation pipelines. These models are susceptible to manipulation, particularly via the *LLM Narcisissm Trope*, which can be exploited for search engine optimization (SEO) [55].

Recent studies demonstrate that LLMs can be guided to rewrite content to improve evaluation scores [10, 87]. Such techniques can distort rankings, spread misinformation, or amplify propaganda. System developers, who target known evaluation setups, may optimize their outputs to align with known evaluator biases – effectively training to pass the test [31, 62] while following their own agenda. This risk increases when evaluation models and prompts are publicly disclosed, enabling targeted reverse-engineering.

LLMs can also be deceived into labeling irrelevant documents as relevant using simple adversarial attacks [3, 59, 69]. These vulnerabilities threaten the integrity of evaluation pipelines and call into question the trustworthiness and reliability of LLM-based assessments.

Quantify effect. This effect can be measured by analyzing performance changes when outputs are explicitly optimized for a specific LLM evaluator, prompt, or configuration [89]. Comparative studies help estimate to which extent evaluation scores are inflated by evaluation-aware tuning.

Guardrail. We advocate developing adversarial test inputs, e.g., targeted content rewrites, to assess the resilience of evaluation metrics under manipulation (cf. Section 5).

To reduce vulnerability, evaluation campaigns (e.g., TREC) should avoid exposing evaluator identities and prompt designs during the submission phase [10]. Blind evaluation setups, where system developers are unaware of the specific LLM and prompt, can reduce gaming. Rotating or ensembling multiple evaluators and using different LLM families adds further robustness. Where feasible, human judgments should remain part of the evaluation loop to validate and audit automated assessments [95].

2.4 Judge Tropes

A common solution to many evaluation and system tropes is to incorporate human judges into the evaluation process. Rather than relying solely on pristine manual judgments, many current approaches involve a collaboration between human assessors and LLMs to generate relevance labels. However, this hybrid setup introduces new risks: subtle forms of bias or priming that can arise during human verification. We refer to these as "judge tropes".

While human involvement is often viewed as the gold standard, misalignment between task design, instructions, or expectations can inadvertently render human judgments ineffective – or even misleading. These issues can compromise the integrity of relevance assessments and, in severe cases, invalidate experimental findings.

Biases may stem from overreliance on LLM outputs, cognitive fatigue, or inadequate oversight, emphasizing the need for robust guardrails and diverse, well-calibrated evaluation protocols.

Judge Trope #12: Rubber-Stamp Effect

– Lack of critical oversight when humans blindly trust LLM labels. –

Experimental studies show that when human assessors are shown LLM-generated answers before making their own judgments, they are significantly more likely to conform to the model's assessment, even when it is demonstrably incorrect [8, 32, 40, 77]. In psychology this is known as Ash conformity experiments [7, 38]. Moreover, as assessor fatigue and task repetition set in, human verification of LLM-generated labels often turns into passive agreement, driven more by trust than by critical scrutiny. This creates a feedback loop: despite involving human judges, evaluations increasingly mirror LLM outputs – even when those outputs diverge from human intuition or real-world utility.

Quantify effect. This effect can be measured by comparing outcomes under fully manual relevance labels versus human-verified LLM labels. Divergence in label quality or ranking decisions will help quantify the degree of automation bias introduced.

Guardrail. To counteract this effect, we draw inspiration from vigilance protocols in security contexts [17]. We propose incorporating vigilance tests in the annotation workflow, by rewarding annotators for identifying errors in LLM outputs. Randomly flipped or adversarial labels can be inserted to test whether annotators are critically engaged. If assessors fail to flag introduced errors, this signals a breakdown in oversight and provides a measurable indicator of rubber-stamping behavior.

Judge Trope #13: Black-box Labeling

– When relevance is complex, labels may be difficult to interpret. –

Relevance labels are often used to represent complex judgments in a simplified form. Whether assigned by humans or LLMs, it can be difficult to determine why a particular passage received a given label – especially when the decision is based on multiple, opaque criteria. This challenge is exacerbated when LLMs provide relevance judgments without clear or trustworthy rationales, increasing the risk of uncritical acceptance by human verifiers [39, 45].

Lack of transparency in LLM-generated relevance labels is a concern [80, Section 8]. Although LLM evaluators can generate explanations alongside labels, these rationales may themselves be flawed and must be critically assessed [71] while avoiding the *Rubber-Stamp trope*.

Quantify effect. Variability between independent manual relevance labels and human-verified LLM labels can reveal the extent of black-box behavior.

Guardrail. To mitigate this issue, complex labeling tasks should be broken into smaller steps, each with explicit reasoning guidelines [30, 51]. Both LLMs and human judges should articulate their reasoning at multiple stages, which makes decisions more interpretable and auditable. Stepwise reasoning, inspired by chain-of-thought prompting in GPT models, can increase transparency and robustness in evaluation [25].

Judge Trope #14: Predictable Secrets

– When human data can be guessed by an LLM. –

Many evaluation paradigms incorporate *secrets*, information known only to human judges and withheld from the system, to prevent evaluation leakage. These include human-generated relevance labels, grading rubrics [30], or nugget annotations [46]. Such mechanisms are designed to guard against the negative effects of LLM-based evaluation.

However, these guardrails become ineffective when an LLM can reliably infer the secret. This introduces inadvertent circularity and undermines the purpose of human oversight [23]. Predictable secrets typically signal that test points are too simplistic or follow an obvious pattern. Evaluation labels generated or structured by LLMs may exhibit consistent patterns that make them predictable and leakable. This allows systems to infer and exploit the evaluation signal, even in good-faith settings [44, 75].

If an IR system can use an LLM to anticipate the secret and incorporate it into its output, it may achieve inflated scores, despite the apparent use of human judgment in the evaluation pipeline.

Quantify effect. The guessability of a secret can be measured by having an LLM predict secrets directly or by computing its per-token likelihood. Downstream effects can be evaluated by replacing the manually created secret with the predicted secret and observing its impact on system rankings.

Guardrails. When validity of the evaluation relies on secret information known only to human judges, it is essential to ensure that the secret is complex and varied enough to resist LLM inference. Designing tasks where secrets require true contextual understanding or subjective reasoning can help maintain this barrier.

3 Case Studies

To demonstrate that our listed LLM evaluation tropes are in fact real issues, we explore two case studies of LLM evaluation methods used in industry and which guardrails were implemented to combat the risks.

3.1 Canva

In this case study (detailed in Cotterill [21]), the LLM evaluator is used in a known-item or re-finding task. It is particularly valuable in a private or enterprise search environment, in which queries and documents are not readily available, let alone viewable by system developers due to privacy restrictions. Rather than providing relevance labels over a corpus of items, instead the LLM is used to synthesize a known-item according to some desired properties. Any number of additional items are also generated, both ones that are "distant" from the target item and ones that are "near" to the target item. It generates one or more queries that represent a user trying to re-find the target item. The characteristics of items and queries are derived from anonymized aggregated statistics over the real user data, thereby grounding the LLM evaluator, avoiding *circularity*. In this way, we generated a conventional test collection, but with the properties that the relevance judgments are known at inception, rather than requiring subsequent human annotation.

The goals of this setup are three-fold:

- Eliminate privacy challenges allowing conventional eyes-on analysis and debugging of search systems.
- Directly construct retrieval and ranking challenges that match specific areas for product improvement e.g., spell correction.
- Ensure repeatability of evaluation, through archiving of generated test collections. Changes to the search system are efficiently and deterministically evaluated offline, accelerating rejection of bad improvements before testing with people.

Although this offline evaluation was then succeeded by online interleaving and A/B experiments, we demonstrated that, provided the improvements we observe in the entirely synthetic LLM evaluation framework are directionally aligned to these later-stage human-centered evaluations, we have no need to also involve humans in the first stage. Both the known-item task (exactly one right answer) and involvement of humans at later-stage evaluations derisk *LLM Narcissism* and *Circularity*; we had no LLM involved in the search system either.

3.2 Valence

This case study examines AI-assisted enterprise coaching, where sessions address complex workplace challenges. The system integrates multiple large commercial LLMs with specialized dialogue components for domain expertise, personalized memory, and user work profiles.

An LLM-based evaluator assesses dialogue quality at both conversational and turn levels, incorporating Client Satisfaction (CSAT) scores and proprietary coaching effectiveness measures. The evaluation methodology follows a rubric-based framework, similar to Lin et al. [47] but manually adapted to coaching tasks by subject matter experts. The evaluation framework serves key purposes including:

- Privacy-Preserving Evaluation: Assesses dialogue quality without exposing sensitive conversations to human reviewers.
- LLM-Based User Simulation: Tests alternative prompts, system configurations, and model components through synthetic interactions.
- LLM as an Autonomous Judge: Enables offline optimization of coaching effectiveness across key conversational dimensions.

A key challenge is *LLM Narcissism*, leading to inflated effectiveness estimates compared to human assessments. To mitigate this, we use separate LLMs for generation and scoring, calibrated against human-labeled datasets.

As our system evolves, enhancing LLM-based evaluation is critical for scalability, privacy, and expert-level assessment quality. Two emerging challenges stand out: 1) LLM as a Reward Model for RL Optimization, while promising, risks reward hacking, where the model optimizes for evaluation heuristics rather than genuine coaching effectiveness. 2) There is a shift to real-time evaluation where evaluation moves from offline to online assessment at the turn-level to enable active intervention when dialogue quality drops, but could further amplify *Circularity* biases potentially creating negative feedback loops such as *Self-training Collapse* and *Overfitting*.

4 Quantifying Circularity

We summarize findings from Clarke and Dietz [16], which quantify circularity due to the *LLM Evaluator as a Ranker trope*, and how it affects the evaluation of retrieval systems using data from the TREC Retrieval-augmented Generation Track (TREC RAG 2024) [83]. The study examines 60 top-performing systems and evaluates how the UMBRELA LLM-based evaluator interacts with systems that either do or do not incorporate LLM-based re-ranking.

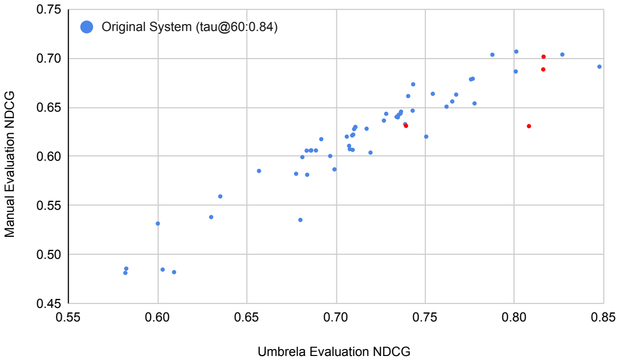

Figure 3: Agreement between the UMBRELA LLM evaluator and manual judgments on the top 60 original TREC RAG 2024 systems. A high Kendall's tau (0.84) confirms the positive findings for UMBRELA when applied to systems that do not incorporate LLM-based ranking strategies. Red dots indicate systems known to include such strategies [16].

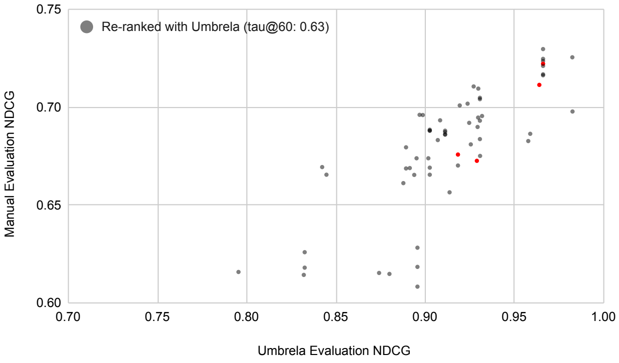

Figure 4: Evaluation of systems re-ranked with UMBRELA, showing divergence between manual and LLM-based NDCG scores. Kendall's tau drops to 0.63, revealing circularity effects when the same LLM is used for both ranking and evaluation.

Figure 3 shows that UMBRELA's system rankings closely match human judgments (Kendall's tau = 0.84) when it is used only as an evaluator on systems without LLM-based components, corroborating findings by UMBRELA developers [83].

However, circularity becomes a concern when UMBRELA is used in two roles: first to re-rank retrieval results, and then to evaluate them. As shown in Figure 4, agreement with manual judgments deteriorates substantially: Kendall's tau drops to 0.63 for the top 60 systems, and to 0.44 for the top 20. These results correspond to a sharp increase in discordant system pairs and signal a breakdown in evaluation validity.

The study observes that NDCG scores produced by Umbrela-based evaluation are often inflated, which can mislead researchers into thinking that systems perform better than indicated by human judgments.

By using the study's experimental setup, we can quantify the effects of circular evaluation that arise when the same LLM is used for both system optimization and performance measurement.

5 Suggested Experimentation Infrastructure

Comparing LLM-based and traditional evaluation metrics on the same set of recent IR systems is essential to quantify effects of LLM evaluators and to reliably identify the best LLM Judge paradigm.

A range of tropes – *Old System*, *LLM Evolution*, *Test Set Leak*, *Self-Training Collapse*, and *Adversarial Threats* – can be systematically studied through continuous experimentation. We propose a TREC-style "Coopetition",[7] built around a shared task with predefined topics. Participants submit in three categories:

1. IR Systems that attempt to solve the task using retrieval-based, generative, or mixed-modality approaches.
2. LLM Evaluators that assess system outputs, either by ranking systems by quality or generating relevance judgments.
3. Content Modification Strategies to deliberately alter documents with the goal of testing system and evaluator robustness, and also help quantify the effects of *Adversarial Threats*.

The outcome is a test collection with human-verified relevance labels and one or more strong LLM-based evaluators. A public leaderboard system could track evaluation results over time, ensuring stability against the *LLM Evolution trope*.

Modified content introduces adversarial challenges that stress-test both retrieval systems and evaluators. This helps identify vulnerabilities and develop more robust methods, contributing to the study of *Adversarial Threats*.

Beyond identifying top systems, the Coopetition supports long-term benchmarking. A portion of labels could remain hidden for blind validation, mitigating *Test Set Leak* and indicating *Self-Training Collapse*.

Collecting implementations. Some evaluation platforms, such as TIRA [34], support the collection of system implementations to facilitate reproducibility.

A shared repository of systems, evaluators, and content modifiers enables ongoing meta-evaluation, supports the detection of failures due to *LLM Evolution*, and provides a reusable experimental test bed. The ability to add systems over time helps prevent *Old System* effects in future meta-evaluations. Simulations allow researchers to study circularity effects arising from the *LLM Evaluator as a Ranker trope*. Ensembles of LLM evaluators may help mitigate *LLM Narcissism*. Continuous submissions also make it possible to introduce fresh judgments to combat *Test Set Leak*, while keeping labels hidden to support reproducible research. Finally, this setup ensures that the best-performing LLM evaluators remain accessible to the broader research community.

Continuous efforts. We envision the Coopetition as an annual effort with evolving tasks, topics, adversarial content, and manual judgments. If a better LLM evaluator emerges, it replaces the current one, updating the test collection. This provides a test bed for studying best practices, failure modes, and evaluation guardrails.

Between iterations, researchers are encouraged to use the current best LLM evaluator for experiments, development, and publications.

6 Conclusions

With this paper we aim to codify the best practices for ensuring that LLM-based evaluation remains a valid experimentation approach for IR research. Maintaining scientific rigor requires identifying and recognizing the risks associated with synthetic training data and LLM-based ranking, while ensuring they are cross-validated with human-verified benchmarks. Automatic evaluation methodologies must be adopted cautiously, treating them as validation tools rather than definitive measures of system performance.

To address these challenges, we propose a new form of TREC-style Coopetition which annually identifies the best LLM evaluation approaches measuring state-of-the-art IR systems on fresh test collections. This would ensure we are using (1) the best LLM evaluators, and (2) continuously confirm the evaluation validity with manual judgments.

Can I use LLM-based judgments in my next conference paper? LLM-based judgments can be used for system evaluation, but only under conditions that safeguard the validity and integrity of the evaluation:

- The LLM-based metrics should have been recently validated against human or user judgments, and used in combination with diverse, complementary metrics to reduce the risk of overfitting or bias.
- The evaluation setup must ensure that LLM-based judgments are not influencing system development in a way that introduces circularity or test signal leakage; such risks should be demonstrably mitigated.
- Known failure modes and evaluation tropes associated with the LLM evaluator should be acknowledged, quantified, and addressed through appropriate guardrails.

Following this framework helps ensure that LLM-based evaluations remain trustworthy, reproducible, and scientifically sound. This reduces the risk of producing IR systems that offer limited value to human users.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under Grant No. 1846017 and the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005). It was also supported by the Engineering and Physical Sciences Research Council grant EP/V025708/1. 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 funding agencies.

[7] Coopetition refers to cooperative competition, where research groups collaboratively compete to identify the most effective LLM-based evaluator, grounded in manual assessments.

References

[1] Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. 2024. Can We Use Large Language Models to Fill Relevance Judgment Holes?. In *EMTCIR '24: The First Workshop on Evaluation Methodologies, Testbeds and Community for Information Access Research.*

[2] Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. 2024. Generative information retrieval evaluation. In *Information Access in the Era of Generative AI*. Springer, 135–159.

[3] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2024. LLMs can be Fooled into Labelling a Document as Relevant: best café near me; this paper is perfectly relevant. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2024)*. 32–41.

[4] Thales Sales Almeida, Giovana Kerche Bonás, João Guilherme Alves Santos, Hugo Abonizio, and Rodrigo Nogueira. 2025. TiEBe: A Benchmark for Assessing the Current Knowledge of Large Language Models. *arXiv preprint arXiv:2501.07482* (2025).

[5] D. G. Altman and J. M. Bland. 1983. Measurement in Medicine: The Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society. Series D (The Statistician)* 32, 3 (1983), 307–317. http://www.jstor.org/stable/2987937

[6] Negar Arabzadeh and Charles LA Clarke. 2025. A Human-AI Comparative Analysis of Prompt Sensitivity in LLM-Based Relevance Judgment. *arXiv preprint arXiv:2504.12408* (2025).

[7] Solomon Asch. 1958. Effects of group pressure on the modification and distortion. *Readings in social psychology. New York: Holt, Rinehart and Winston* (1958).

[8] Authors. 2023. The Importance of Distrust in AI. *arXiv preprint arXiv:2307.13601* (2023). https://arxiv.org/abs/2307.13601

[9] Krisztian Balog, Donald Metzler, and Zhen Qin. 2025. Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. *arXiv preprint arXiv:2503.19092* (2025).

[10] Niv Bardas, Tommy Mordo, Oren Kurland, Moshe Tennenholtz, and Gal Zur. 2025. Prompt-Based Document Modifications in Ranking Competitions. *arXiv preprint arXiv:2502.07315* (2025).

[11] Ran Ben Basat, Moshe Tennenholtz, and Oren Kurland. 2015. The Probability Ranking Principle is Not Optimal in Adversarial Retrieval Settings. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. 51–60.

[12] Shaily Bhatt and Fernando Diaz. 2024. Extrinsic Evaluation of Cultural Competence in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 12345–12356.

[13] Sebastian Bordt, Harsha Nori, Vanessa Rodrigues, Besmira Nushi, and Rich Caruana. 2024. Elephants Never Forget: Memorization and Learning of Tabular Data in Large Language Models. In *Proceedings of the First Conference on Language Modeling (COLM 2024)*. Philadelphia, USA. https://arxiv.org/abs/2404.06209 Camera-ready version; supersedes arXiv:2404.06209.

[14] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009* (2023).

[15] Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 10853–10875.

[16] Charles L. A. Clarke and Laura Dietz. 2025. LLM-based relevance assessment still can't replace human relevance assessment. In *EVIA 2025: Proceedings of the Tenth International Workshop on Evaluating Information Access (EVIA 2025), a Satellite Workshop of the NTCIR-18 Conference, June 10-13, 2025, Tokyo, Japan*. 1–5. doi:10.20736/0002002105

[17] Victoria L Claypoole, Daryn A Dever, Kody L Denues, and James L Szalma. 2019. The effects of event rate on a cognitive vigilance task. *Human factors* 61, 3 (2019), 440–450.

[18] Cyril Cleverdon. 1967. The Cranfield Tests on Index Language Devices. *Aslib Proceedings* 19, 6 (1967), 173–194.

[19] Gordon V Cormack and Maura R Grossman. 2015. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv preprint arXiv:1504.06868* (2015).

[20] Gordon V Cormack and Maura R Grossman. 2016. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 1039–1048.

[21] Ellese Cotterill. 2024. How to improve search without looking at queries or results. https://www.canva.dev/blog/engineering/how-to-improve-search-without-looking-at-queries-or-results/

[22] 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*. 526–537.

[23] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating Data Contamination in Modern Benchmarks for Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 8698–8711.

[24] Dario Di Palma, Felice Antonio Merra, Maurizio Sfilio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. Do LLMs memorize recommendation datasets? a preliminary study on MovieLens-1m. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025)*.

[25] Laura Dietz and Naghmeh Farzi. 2025. Criteria-Based LLM Relevance Judgments. In *Proceedings of the 11th ACM SIGIR / The 15th International Conference on Innovative Concepts and Theories in Information Retrieval*.

[26] Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M Khapra. 2024. Finding Blind Spots in Evaluator LLMs with Interpretable Checklists. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 16279–16309. doi:10.18653/v1/2024.emnlp-main.911

[27] Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. 2024. Model Collapse Demystified: The Case of Regression. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. Vancouver, Canada. Main-conference track; replaces arXiv:2402.07712.

[28] Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a Personalized Judge?. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10126–10141. doi:10.18653/v1/2024.findings-emnlp.592

[29] Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*. 39–50.

[30] Naghmeh Farzi and Laura Dietz. 2024. Exam++: Llm-based answerability metrics for ir evaluation. In *Proceedings of LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval*.

[31] Pedro Ferreira, Wilker Aziz, and Ivan Titov. 2025. Truthful or Fabricated? Using Causal Attribution to Mitigate Reward Hacking in Explanations. *arXiv preprint arXiv:2504.05294* (2025).

[32] Raymond Fok and Daniel S Weld. 2024. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine* 45, 3 (2024), 317–332.

[33] Maik Fröbe, Lukas Gienapp, Martin Potthast, and Matthias Hagen. 2023. Bootstrapped nDCG Estimation in the Presence of Unjudged Documents. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*. 313–329.

[34] Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, Niklas Deckers, Simon Reich, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. The Information Retrieval Experiment Platform. In *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2826–2836. doi:10.1145/3539618.3591888

[35] Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Zijian Zhang, Wanyu Wang, Yuyang Ye, Shanru Lin, Huifeng Guo, and Ruiming Tang. 2024. LLM-enhanced Reranking in Recommender Systems. *arXiv preprint arXiv:2406.12433* (2024).

[36] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. 2024. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. *arXiv preprint arXiv:2404.01413* (2024).

[37] Charles Goodhart. 1975. Problems of monetary management: the UK experience in papers in monetary economics. *Monetary Economics* 1 (1975).

[38] Lewis D Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly T Mai, Maria Vau, Matthew Caldwell, and Augustine Marvor-Parker. 2023. Susceptibility to influence of large language models. *arXiv preprint arXiv:2303.06074* (2023).

[39] Jiazhou Ji, Ruizhe Li, Shujun Li, Jie Guo, Weidong Qiu, Zheng Huang, Chiyu Chen, Xiaoyu Jiang, and Xinru Lu. 2024. Detecting Machine-Generated Texts: Not Just "AI vs Humans" and Explainability is Complicated. *arXiv preprint arXiv:2406.18259* (2024).

[40] Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2021. Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions. *arXiv preprint arXiv:2106.16122* (2021).

[41] Oren Kurland and Moshe Tennenholtz. 2022. Competitive Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. 2838–2849.

[42] David La Barbera, Riccardo Lunardi, Mengdie Zhuang, and Kevin Roitero. 2025. Impersonating the Crowd: Evaluating LLMs' Ability to Replicate Human Judgment in Misinformation Assessment. In *Proceedings of Innovative Concepts and Theories in Information Retrieval (ICTIR)*.

[43] Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. LLM Generated Persona is a Promise with a Catch. *arXiv preprint arXiv:2503.16527*

(2025).

[44] Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025. Preference Leakage: A Contamination Problem in LLM-as-a-judge. *arXiv preprint arXiv:2502.01534* (2025).

[45] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv:2306.01941* (2023).

[46] Jimmy Lin and Dina Demner-Fushman. 2007. Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 561–568.

[47] Ying-Chun Lin, Jennifer Neville, Jack Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, Deepak Gupta, Sujay Kumar Jauhar, Xia Song, Georg Buscher, Saurabh Tiwary, Brent Hecht, and Jaime Teevan. 2024. Interpretable User Satisfaction Estimation for Conversational Systems with Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 11100–11115.

[48] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2511–2522. doi:10.18653/v1/2023.emnlp-main.153

[49] Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2024. LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores. In *Findings of the Association for Computational Linguistics ACL 2024*. 12688–12701.

[50] Xiaolu Lu, Alistair Moffat, and J Shane Culpepper. 2017. Can Deep Effectiveness Metrics Be Evaluated Using Shallow Judgment Pools?. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 35–44.

[51] James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, et al. 2024. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1904–1915.

[52] Shikib Mehri and Maxine Eskenazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 681–707.

[53] Alistair Moffat, Falk Scholer, Paul Thomas, and Peter Bailey. 2015. Pooled Evaluation Over Query Variations: Users Are as Diverse as Systems. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 1759–1762.

[54] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27 (2008).

[55] Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. 2025. Adversarial Search Engine Optimization for Large Language Models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net. https://openreview.net/forum?id=hkdqxN3c7t

[56] Harrie Oosterhuis, Rolf Jagerman, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2024. Reliable Confidence Intervals for Information Retrieval Evaluation Using Generative A.I.. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. 2307–2317.

[57] Vishakh Padmakumar and He He. 2023. Does Writing with Language Models Reduce Content Diversity? *arXiv preprint arXiv:2309.05196* (2023).

[58] Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems* 37 (2024), 68772–68802.

[59] Andrew Parry, Maik Fröbe, Sean MacAvaney, Martin Potthast, and Matthias Hagen. 2024. Analyzing Adversarial Attacks on Sequence-to-Sequence Relevance Models. In *Advances in Information Retrieval. 46th European Conference on IR Research (ECIR 2024) (Lecture Notes in Computer Science)*. Springer, Berlin Heidelberg New York.

[60] Hossein A. Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. 2024. Synthetic Test Collections for Retrieval Evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[61] Hossein A. Rahmani, Varsha Ramineni, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. 2025. Towards Understanding Bias in Synthetic Data for Evaluation. arXiv:2506.10301 [cs.IR] https://arxiv.org/abs/2506.10301

[62] Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 7499–7517.

[63] Kun Ran, Shuoqi Sun, Khoi Nguyen Dinh Anh, Damiano Spina, and Oleg Zendel. 2025. RMIT-ADM+S at the SIGIR 2025 LiveRAG Challenge – GRAG: Generation-Retrieval-Augmented Generation. In *LiveRAG Challenge at SIGIR 2025*. 9 pages.

[64] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2020. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association* 27 (2020), 1431–1436.

[65] Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2023. Data Contamination Through the Lens of Time. *arXiv preprint arXiv:2310.10628* (2023).

[66] Kevin Roitero, Alessandro Checco, Stefano Mizzaro, and Gianluca Demartini. 2022. Preferences on a Budget: Prioritizing Document Pairs When Crowdsourcing Relevance Judgments. In *Proceedings of the ACM Web Conference 2022*. 319–327.

[67] Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for Each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 10776–10787. doi:10.18653/v1/2023.findings-emnlp.722 Replaces arXiv:2310.18018.

[68] Tetsuya Sakai, Sijie Tao, and Zhaohao Zeng. 2021. WWW3E8: 259,000 Relevance Labels for Studying the Effect of Document Presentation Order for Relevance Assessors. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2376–2382.

[69] Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2024. Optimization-based prompt injection attack to llm-as-a-judge. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 660–674.

[70] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. *arXiv preprint arXiv:2305.17493* (2023). https://arxiv.org/abs/2305.17493

[71] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature* 631, 8022 (2024), 755–759.

[72] Chenglei Si et al. 2024. Evaluating Large Language Model Biases in Persona-Steered Generation. In *Findings of the Association for Computational Linguistics: ACL 2024*. 6789–6800.

[73] Aaditya K Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, David Esiobu, Maria Lomeli, and Gergely Szilvasy. 2024. Evaluation Data Contamination in LLMs: How Do We Measure It and (When) Does It Matter? *arXiv preprint arXiv:2411.03923* (2024).

[74] Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. 2024. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10406–10421.

[75] Ian Soboroff. 2025. Don't use LLMs to make relevance judgments. *Information retrieval research journal* 1, 1 (2025), 10–54195.

[76] Karen Spärck Jones and C. J. van Rijsbergen. 1975. Report on the need for and provision of an 'ideal' information retrieval test collection. *Computer Laboratory* (1975).

[77] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence* (2025), 1–11.

[78] Paul Thomas, Gabriella Kazai, Nick Craswell, and Seth Spielman. 2024. What matters in a measure? A perspective from large-scale search evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 282–292.

[79] 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*.

[80] Johanne R. Trippas and J. Shane Culpepper. 2025. Report from the Fourth Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025). *ACM SIGIR Forum* 59, 1 (June 2025), 68 pages.

[81] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 16612–16631.

[82] 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).

[83] 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).

[84] 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).

[85] Ellen M. Voorhees. 2019. The Evolution of Cranfield. In *Information Retrieval Evaluation in a Changing World*, Nicola Ferro and Carol Peters (Eds.). Vol. 41. Springer International Publishing, 45–69.

[86] Ellen M Voorhees, Ian Soboroff, and Jimmy Lin. 2022. Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models? *arXiv preprint*

arXiv:2201.11086 (2022).

[87] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv preprint arXiv:2212.03533* (2024).

[88] Fangyun Wei, Xi Chen, and Lin Luo. 2024. Rethinking generative large language model evaluation for semantic comprehension. *arXiv preprint arXiv:2403.07872* (2024).

[89] Yuchen Wen, Keping Bi, Wei Chen, Jiafeng Guo, and Xueqi Cheng. 2024. Evaluating Implicit Bias in Large Language Models by Attacking From a Psychometric Perspective. *CoRR* (2024).

[90] Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244* (2024).

[91] Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking Benchmark Leakage in Large Language Models. arXiv:2404.18824 [cs.CL] https://arxiv.org/abs/2404.18824

[92] Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15474–15492.

[93] Wei Jie Yeo, Teddy Ferdinan, Przemyslaw Kazienko, Ranjan Satapathy, and Erik Cambria. 2024. Self-training Large Language Models through Knowledge Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 12345–12356.

[94] Fan Zhang et al. 2020. Towards a Better Understanding of Evaluation Metrics. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1439–1442.

[95] Fan Zhang et al. 2023. Constructing and Meta-Evaluating State-Aware Evaluation Metrics for Information Retrieval. *Information Retrieval Journal* (2023).

[96] Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval*. 379–388.

[97] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't Make Your LLM an Evaluation Benchmark Cheater. *arXiv preprint arXiv:2311.01964* (2023).

[98] Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. 2025. Evaluating Judges as Evaluators: The JETTS Benchmark of LLM-as-Judges as Test-Time Scaling Evaluators. *arXiv preprint arXiv:2504.15253* (2025).

[99] Justin Zobel. 2023. When Measurement Misleads: The Limits of Batch Assessment of Retrieval Systems. *SIGIR Forum* 56 (Jan. 2023), 20 pages.