The LLM-as-a-Judge Co-adaptation Spiral

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

Abstract

The same LLM judge is now commonly used both as a system component in RAG pipelines and as an evaluator on leaderboards. As judge-style system components proliferate, systems drift toward outputs the LLM judge recognizes, and meta-evaluation of LLM judges against a fixed human reference stops being meaningful. Trustworthy LLM-as-Judge evaluation therefore requires continuous human involvement in each evaluation cycle, anchored by manual evaluation artifacts. These artifacts allow re-verification of the current population of systems and LLM judges.

Keywords

LLM-as-judge, RAG evaluation, meta-evaluation, generative IR

1 Introduction

A retrieval metric is meaningful only insofar as it tracks human judgment on the systems being scored. Pooled human relevance judgments satisfy this requirement by construction. LLM judges are now increasingly used as substitutes, and reported agreement with human assessors is strong enough that some work argues they can replace humans entirely [27]. This paper argues that claims that "LLM can replace humans" overlook a distinction between two roles an LLM judge can play in an IR pipeline.

1. **LLM-Judge for Evaluation:** It is used to predict quality scores for each system response to obtain a leaderboard of systems.
2. **LLM-Judge as a System Component (RAGE):** The same idea is used as a reranker, a relevance feature, a training-data filter, a source of self-labeled training data, or when training with user simulation.

Claims about evaluator reliability rest on meta-evaluation: the agreement between the LLM judge and human assessors on a validation population S_{val}, together with the assumption that this agreement transfers to the judge's deployment population S_{dep}. Our central claim is that widespread adoption of LLM-judge components undermines that transfer.

2 When LLM-as-Judge Works

Recent work on LLM judges identifies operational conditions under which evaluator use produces reliable numbers [4, 8–10]. Four matter here.

(1) Independence between the LLM judge and the system being scored. When the evaluated system shares model family, prompts, or training signals with the judge, the score reflects shared signal rather than relevance. Independence is the precondition for non-circular meta-evaluation [14].

(2) Decomposition over holistic judgment. Rubric-based methods such as EXAM and Criteria-Based judgments decompose relevance into externally inspectable sub-questions, each answered by the LLM judge as a constrained verification rather than a global score [3, 5, 20]. The rubric is human-inspectable, the judge's task is narrower, and failure modes are more localized.

(3) Evaluation artifacts that an LLM cannot guess. When the gold answer or relevance signal can be inferred by the LLM judge from the query alone, the evaluation reduces to checking that two LLM approaches agree, even when these are referred to as system and evaluator. The safeguard is human evaluation artifacts that an evaluation is based on, such as nugget banks. It is important that an LLM cannot regenerate these artifacts [1, 7, 23]. We acknowledge that it is increasingly difficult for humans to provide evaluation artifacts that "outsmart" the LLM.

(4) Continuous human-anchored audit. Reported correlations expire as system populations shift. A safeguard is to continuously supply fresh systems and human evaluation artifacts in each cycle, rather than a one-time validation that may leak into LLM training data [12, 25].

Conditions 2–4 are evaluator-side design choices. Condition 1, by contrast, depends on the system population the LLM judge is asked to evaluate. The remainder of this paper argues that field-wide adoption of LLM-judge components will erode condition 1 on the set of systems S_{dep} that will be evaluated, but are not available during meta-evaluation.

3 Why RAGE Adoption Limits Evaluator Use

3.1 Issue: Judging the Judges on Old Systems

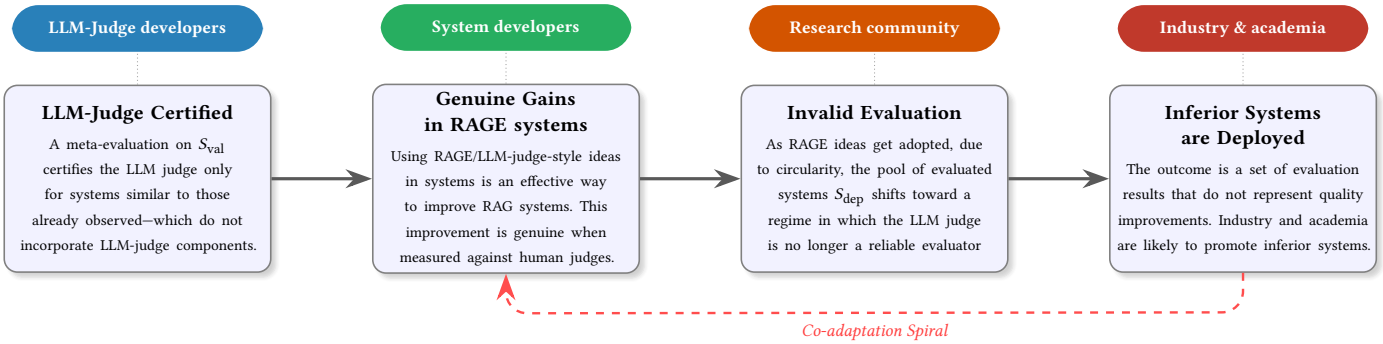
LLM judges as evaluators are validated positively whenever an LLM judge produces the same leaderboard or relevance judgments as humans. The human judgments are collected on a judgment pool generated by a limited set of systems S_{val}, such as those submitted to an open challenge.

The general assumption is that a positive meta-evaluation suggests that an LLM judge will also be effective on innovative future systems S_{dep}. Hence positive meta-evaluations of the LLM-judge paradigm (e.g. UMBRELA) lead to widespread adoption for cheap and fast evaluation. But we need human evaluation to verify this assumption.

3.2 Adoption of LLM-Judge in Systems

Several studies demonstrate the usefulness of an LLM judge as a system component against human assessment. These are genuine improvements measured with human assessments—not an artifact of circularity in evaluation.

For example, reranking TREC RAG 2024 submissions with UMBRELA improves NDCG@10 under manual judgment [4] for several IR systems. Farzi and Dietz [11] show the same for a multi-criterion judge. RankVicuna [18] and RankZephyr [19] use the same strategy



as an LLM judge in order to rank a pool of documents. In fact, this is also referred to as "LLM-judge as a Ranker" [8].

For retrieval-augmented generation systems, CRUCIBLE incorporates ideas from Nugget-based LLM judges to drive extraction and generation [6].

Using LLM-as-a-Judge as a component in IR systems is successful, because optimizing with LLM-judge-signals often improves outputs that humans also prefer.

3.3 The Co-Adaptation Spiral

Evaluator use becomes unreliable when the LLM judge evaluates a system that uses LLM judge ideas as components. This effect is also referred to as *circularity* and has been documented empirically [7, 16, 17, 21]. The resulting evaluation score partly reflects the system's access to the judge's truth signal rather than underlying relevance as judged by humans.

To offer a tongue-in-cheek analogy: if we were to use the top 20 of a BM25 ranking to define relevant documents, then obviously BM25 would obtain a perfect evaluation score. But this does not imply that BM25 is the perfect ranking system.

3.4 The LLM Knowledge Ceiling

The LLM judge can only distinguish systems along dimensions its own parametric knowledge recognizes as relevant [26]. This could lead to the systematic leniency of LLM judges, which compresses observable differences [28].

As IR systems improve at producing outputs that satisfy what the judge would check, the LLM judge's ability to detect differences between the best systems is diminished—not because the systems are equally good, but because the judge has reached its discriminative ceiling. This limit is independent of architectural overlap.

This claim can be tested by studying the ability of an LLM judge to differentiate among better and best systems when these are improved with LLM-as-a-Judge components.

3.5 Safeguards and Human Judgment

Adoption of LLM-judge ideas in systems changes the system pool S_{dep} out of the regime where the LLM judge remains a valid evaluator [8]. This shift is hard to mitigate without also giving up what makes RAGE-style components useful.

Subversion probes [7] may help to detect such issues. Spiliopoulou et al. [24] provide a statistical framework for isolating self- and family-bias on a given judge–system pair.

While many safeguards are discussed, such as ensembles, hidden labels, and prompt variation, all of these interventions can be incorporated in the system as well—rendering them ineffective.

Human judgment differs in a structural way: it cannot be embedded as a reusable system component in the same way.

Human judgment is different: it cannot be embedded as a reusable system component. TREC Auto-Judge and TIRA provide useful infrastructure for controlled meta-evaluation, but not for one-time certification.

LLM judges remain useful as system components when their effect is validated by humans.

3.6 Constructive Outlook

A credible evaluation pipeline re-applies the conditions in every cycle: renewed meta-evaluation of the LLM judge against fresh human judgments on current submissions, and probes targeting the next likely adoption patterns. TREC Auto-Judge supports this recurring process; meta-evaluation frameworks [13, 15, 22] and human-in-the-loop tooling [2] provide complementary infrastructure. Leaderboard scores from the judge are valid only for the certification context that produced them: a particular cycle, a particular human reference set, and a particular class of systems for which condition 1 was verified to hold. Outside that context, they are extrapolations.

This is the practical limit of LLM-as-Judge evaluation. The judge reduces evaluation cost when applied within a current, condition-verified certification context. Without that context, a leaderboard built on the judge measures systems' access to the judge's signal more than progress on the IR task.

4 Conclusion

We discussed why LLM-as-Judge produces trustworthy numbers only when its operating conditions are re-established in every evaluation cycle. The same widespread adoption that makes LLM judges effective as components in IR systems is what makes them unreliable as evaluators against a fixed reference. The constructive path is not to reject LLM-as-Judge, but to anchor every leaderboard cycle to fresh human judgment on current submissions.

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 2024). 32–41.

[2] Zahra Ashktorab, Elizabeth M. Daly, Erik Miehling, Werner Geyer, Martin Santillan Cooper, Tejaswini Pedapati, Michael Desmond, Qian Pan, and Hyo Jin Do. 2025. EvalAssist: A Human-Centered Tool for LLM-as-a-Judge. *CoRR* abs/2507.02186 (2025). doi:10.48550/ARXIV.2507.02186 arXiv:2507.02186

[3] Yinzhu Chen, Abdine Maiga, Hossein A. Rahmani, and Emine Yilmaz. 2026. Automated Rubrics for Reliable Evaluation of Medical Dialogue Systems. *arXiv preprint arXiv:2601.15161* (2026).

[4] 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, NTCIR-18 Conference*.

[5] Kaustubh D. Dhole and Eugene Agichtein. 2026. RubricRAG: Towards Interpretable and Reliable LLM Evaluation via Domain Knowledge Retrieval for Rubric Generation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2026)*. arXiv:2603.20882.

[6] Laura Dietz, Bryan Li, Gabrielle Liu, Jia-Huei Ju, Eugene Yang, Dawn Lawrie, William Walden, and James Mayfield. 2026. Incorporating Q&A Nuggets into Retrieval-Augmented Generation. In *Proceedings of the 48th European Conference on Information Retrieval (ECIR 2026)*.

[7] Laura Dietz, Bryan Li, Eugene Yang, Dawn Lawrie, William Walden, and James Mayfield. 2026. Insider Knowledge: How Much Can RAG Systems Gain from Evaluation Secrets?. In *Proceedings of the 48th European Conference on Information Retrieval (ECIR 2026)*. arXiv:2601.13227.

[8] 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 '25)*. doi:10.1145/3731120.3744588

[9] Naghmeh Farzi and Laura Dietz. 2024. Pencils Down! Automatic Rubric-based Evaluation of Retrieve/Generate Systems. In *Proceedings of the 2024 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '24)*. doi:10.1145/3664190.3672511

[10] Naghmeh Farzi and Laura Dietz. 2025. Criteria-Based LLM Relevance Judgments. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR '25)*. doi:10.1145/3731120.3744591

[11] Naghmeh Farzi and Laura Dietz. 2026. Learning to Rank with Multi-Criteria LLM-Judge Annotations. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*. ACM, Melbourne, VIC, Australia. doi:10.1145/3805712.3809580

[12] Naghmeh Farzi, Tim Hagen, Eugene Yang, Maik Fröbe, Ronak Pradeep, Hossein A. Rahmani, Xi Wang, Oleg Zendel, Martin Potthast, and Laura Dietz. 2026. Auto-Judge: A Cross-Task Benchmark for Comparing LLM Judges for Citation-Grounded RAG Systems. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)* (July 20–24, 2026). ACM, Melbourne, VIC, Australia. doi:10.1145/3805712.3808601

[13] Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, and Asaf Yehudai. 2025. JustRank: Benchmarking LLM Judges for System Ranking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[14] Dongryeol Lee, Yerin Hwang, Taegwan Kang, Minwoo Lee, Younhyung Chae, and Kyomin Jung. 2026. Judging Against the Reference: Uncovering Knowledge-Driven Failures in LLM-Judges on QA Evaluation. *arXiv preprint arXiv:2601.07506* (2026).

[15] Xiaochuan Li, Ke Wang, Girija Gouda, Shubham Choudhary, Yaqun Wang, Linwei Hu, Joel Vaughan, and Freddy Lecue. 2025. Who Judges the Judge? LLM Jury-on-Demand: Building Trustworthy LLM Evaluation Systems. *arXiv preprint arXiv:2512.01786* (2025).

[16] 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. doi:10.18653/v1/2024.findings-acl.753

[17] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*. arXiv:2404.13076.

[18] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankVicuna: Zero-Shot Listwise Document Reranking with Open-Source Large Language Models. *arXiv preprint arXiv:2309.15088* (2023). https://arxiv.org/abs/2309.15088

[19] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze! *arXiv preprint arXiv:2312.02724* (2023). https://arxiv.org/abs/2312.02724

[20] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2025. The Great Nugget Recall: Automating Fact Extraction and RAG Evaluation with Large Language Models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 180–190.

[21] Hossein A. Rahmani, Varsha Ramineni, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. 2025. Towards Understanding Bias in Synthetic Data for Evaluation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 5166–5170.

[22] 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).

[23] 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. doi:10.18653/v1/2024.emnlp-main.427

[24] Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. 2025. Play Favorites: A Statistical Method to Measure Self-Bias in LLM-as-a-Judge. *arXiv preprint arXiv:2508.06709* (2025).

[25] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*. 404–430.

[26] Minzhu Tu, Shiyu Ni, and Keping Bi. 2026. How Long Reasoning Chains Influence LLMs' Judgment of Answer Factuality. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*. arXiv:2604.06756.

[27] 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).

[28] C. Yu, H. Li, G. Zuccon, J. Mackenzie, and T. 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 2026)*.