LLM-as-a-Judge: Rapid Evaluation of Legal Document Recommendation via Retrieval-Augmented Generation

Anu Pradhan, Alexandra Ortan, Apurv Verma, Madhavan Seshadri

Bloomberg

New York, NY, USA

{apradhan11,aortan,averma239,mseshadri}@bloomberg.net

Abstract

The evaluation bottleneck in recommendation systems has become particularly acute with the rise of Generative AI, where traditional metrics fall short of capturing nuanced quality dimensions that matter in specialized domains like legal research. Can we trust Large Language Models to serve as reliable judges of their own kind? This paper investigates LLM-as-a-Judge as a principled approach to evaluating Retrieval-Augmented Generation systems in legal contexts, where the stakes of recommendation quality are exceptionally high.

We tackle two fundamental questions that determine practical viability: which inter-rater reliability metrics best capture the alignment between LLM and human assessments, and how do we conduct statistically sound comparisons between competing systems? Through systematic experimentation, we discover that traditional agreement metrics like Krippendorff's alpha can be misleading in the skewed distributions typical of AI system evaluations. Instead, Gwet's AC2 and rank correlation coefficients emerge as more robust indicators for judge selection, while the Wilcoxon Signed-Rank Test with Benjamini-Hochberg corrections provides the statistical rigor needed for reliable system comparisons.

Our findings suggest a path toward scalable, cost-effective evaluation that maintains the precision demanded by legal applications—transforming what was once a human-intensive bottleneck into an automated, yet statistically principled, evaluation framework.

Keywords

LLM-as-a-Judge, Large Language Models, RAG, Evaluation

ACM Reference Format:

Anu Pradhan, Alexandra Ortan, Apurv Verma, Madhavan Seshadri. 2025. LLM-as-a-Judge: Rapid Evaluation of Legal Document Recommendation via Retrieval-Augmented Generation. In *EARL '25 at RecSys '25, Sept 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Recommendation systems play an essential role across various sectors, including legal research, where the accuracy and quality of recommended documents can directly impact professional decision-making. The advent of LLMs has revolutionized the field of natural language processing, offering unprecedented capabilities in understanding and generating human-like text. These models, such as GPT-4 and others, have shown promise not only in generating content but also in evaluating and assessing - a concept we refer to as "LLM-as-a-Judge". Leveraging LLMs in evaluation tasks holds significant potential for increasing efficiency and consistency in areas like search relevancy and answer quality assessments. However, using LLMs as evaluators introduces challenges related to the reliability and variability of their judgments compared to human raters.

We view the LLM-as-a-Judge as an artificial rater, and explore the use of Inter-Rater Reliability (IRR) metrics to measure the reliability with human raters. IRR is the degree of agreement among different raters assessing the same set of items. High IRR indicates consistent and reliable ratings, which are essential for tasks like tuning search algorithms or training machine learning models. In this paper, we explore the application of LLM-as-a-Judge to evaluate different generative AI (Gen AI) solutions across different Bloomberg Law products. In particular, we study how multiple raters—both human and LLM-based—evaluate items on ordinal scales assessing relevance, safety, hallucinations, correctness, and overall quality. We formulate two research questions (RQs) to guide our investigation.

RQ1: How can we effectively evaluate and select LLM judges for legal RAG systems using a comprehensive set of inter-rater reliability metrics? To address the question of how to effectively evaluate and select LLM judges for legal RAG systems, we emphasize the need to consider a comprehensive set of inter-rater reliability metrics rather than relying on a single measure. This multimetric approach is crucial due to the complex nature of legal language and the varied challenges in AI evaluation. We examine traditional IRR metrics like Krippendorff's Alpha (K-Alpha), which, while widely used, may misrepresent agreement levels in skewed data distributions common in Gen AI system evaluations. To address these limitations, we also explore more recent techniques such as Gwet's AC2, which offers improved robustness in these scenarios. Additionally, we consider correlation metrics such as Spearman rank correlation and Kendall tau (τ), particularly suited for ordinal ratings and measuring agreement on relative rankings [11, 14, 29]. By evaluating this diverse set of metrics based on their ability to measure LLM-human agreement, handle skewed distributions, and reliably rank Gen AI systems, our aim is to provide a nuanced and comprehensive framework for selecting LLM judges.

RQ2: Which statistical methods are most effective for comparing legal RAG systems evaluated by LLM judges, and how do different multiple hypothesis testing corrections impact these comparisons? In evaluating LLM judges for legal RAG systems, we carefully selected the Wilcoxon Signed-Rank Test (WSRT) as our primary statistical method due to its nonparametric nature, which

IRR Metric	Ordinal Scale Support	Distribution Robustness	Rank Sensitivity	Clear Interpretation	Missing Data Tolerance	Computational Efficiency	Multi-Rater Capability
Cohen's Kappa	✗	✗	✗	✓	✗	✓	✗
Percent Agreement	✗	✗	✗	✓	✗	✓	✓
Kendall's Tau	✓	✓	✓	✓	✗	✗	✗
Spearman's Rank	✓	✓	✓	✓	✗	✓	✗
Krippendorff's Alpha	✓	✗	✓	✓	✓	✓	✓
Gwet's AC2	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of IRR Metrics for Legal RAG System Evaluation. Note: ✓ indicates the metric supports the attribute; ✗ indicates it does not.

is ideal for ordinal data from LLM evaluations that may not follow normal distributions [32]. To address multiple testing issues, we evaluated three correction methods: Bonferroni, Benjamini-Hochberg (B-H), and Holm-Bonferroni. Given the exploratory nature of our study and the multiple performance aspects, we selected the B-H method to balance error control with detection sensitivity [6]. This approach allows us to effectively identify significant differences in various legal comparisons of the RAG system while maintaining statistical rigor.

2 Related Work

Retrieval Augmented Generation (RAG) effectively enhances the capabilities of LLMs by integrating external knowledge sources, making it valuable for recommendation tasks in specialized domains [10, 19]. Traditional automated evaluation metrics such as ROUGE and BLEU depend on reference responses, which limits their effectiveness in complex open-ended recommendation scenarios [20, 25, 34]. While human evaluations are typically considered the gold standard due to their accuracy, they are impractical at large scales given the significant time and expertise required [8]. Recent advances in LLMs have sparked interest in their potential as automated evaluators, particularly in specialized domains, such as law. This section examines three key areas of relevant literature: the use of LLMs as judges, challenges in ensuring reliable evaluations, and statistical methods for analyzing LLM-based assessments.

LLM-as-a-Judge: Using LLMs as automated judges has emerged as a scalable alternative to human evaluation [21, 35]. Although GPT-4 has shown promise in achieving human-level agreement on certain tasks [27], recent studies have identified key challenges, including cognitive biases [17], self-preference [24], and systematic errors in evaluation [31]. To address these limitations, researchers have explored committee-based approaches using multiple LLM [5, 30] and specialized training of smaller expert judges [16, 36]. However, these methods often lack rigorous guarantees of reliability and agreement with human preferences.

Reliable Evaluation: While LLMs can scale evaluation across large datasets, ensuring reliability as judges presents significant challenges in terms of cognitive biases [17], self-preference [24] and systematic errors in evaluation [31]. To address these challenges, researchers have explored various approaches. For instance, Chan et al. (2023) proposed a multi-agent debate framework to mitigate individual model biases [5]. Sottana et al. (2023) demonstrated that GPT-4 can achieve human-level agreement on certain tasks, but noted persistent challenges in complex evaluations [27]. These studies collectively underscore the importance of developing robust methods for LLM-based evaluations.

Statistical Analysis: Recent studies have emphasized the critical need for robust statistical methods to evaluate language models, highlighting issues in experimental design, variance quantification and uncertainty estimation. Miller (2024) introduced a framework for adding error bars to LLM evaluations, proposing the use of paired differences at the question level for statistical inference [22]. Oosterhuis et al. (2024) developed methods to construct reliable confidence intervals for IR evaluation metrics using LLM-generated annotations [23]. Card et al. (2020) revealed the prevalence of underpowered experiments in NLP, particularly in popular benchmarks, highlighting the importance of power analysis in experimental design [4]. Together, these works offer valuable information for designing statistically sound experiments and reliable evaluation procedures for LLMs.

3 Experimental Design

We evaluated several metrics, including Cohen's Kappa [7], Krippendorff's Alpha [18], Spearman's rank correlation [28], Kendall's Tau [15], percent agreement and two extensions of Gwet's AC2 with linear weighting schemes (Gwet AC2-L) and quadratic (Gwet AC2-Q) [12]. While we did not evaluate every available IRR metric, such as Fleiss's Kappa [9] and the Brennan-Prediger coefficient [3], our aim was to emphasize the importance of utilizing a diverse set of metrics rather than conducting an exhaustive analysis of all available IRR metrics.e evaluated several metrics, including Cohen's Kappa [7], Krippendorff's Alpha [18], Spearman's rank correlation, Kendall's Tau, percent agreement and two extensions of Gwet's AC2 with linear weighting schemes (Gwet AC2-L) and quadratic (Gwet AC2-Q). While we did not evaluate every available IRR metric, such as Fleiss's Kappa and the Brennan-Prediger coefficient, our aim was to emphasize the importance of utilizing a diverse set of metrics rather than conducting an exhaustive analysis of all available IRR metrics. As a leading provider of legal information services, Bloomberg Law faces unique challenges in evaluating RAG systems. Legal content requires high accuracy standards, as even minor errors could impact

critical legal decisions. Additionally, the proprietary nature of our content and systems, combined with client confidentiality requirements, creates constraints on sharing detailed system specifications or complete evaluation results. Our evaluation was conducted on two legal RAG systems operating over a comprehensive legal corpus reflecting real production challenges at Bloomberg Law where we evaluate thousands of legal query-document pairs monthly across multiple products. Each RAG system consists of two critical components: a retrieval component that identifies relevant legal documents, and an answer generation component that synthesizes the retrieved information into coherent responses.

We compared two distinct legal RAG systems. System A utilizes traditional BM25 retrieval combined with an open-source LLM summarizer applied to the top 5 retrieved documents. System B incorporates improvements in the retrieval system and employs the proprietary GPT-4 model by OpenAI as the summarizer. This comparison reflects realistic industry scenarios and evaluates significant technological enhancements and their practical impacts on recommendation quality. Our evaluation framework specifically targeted both these components: the retrieval effectiveness through search relevancy assessment, and the generation quality through answer evaluation. For search relevancy, we evaluated passage-query pair relevance on a scale of 1 to 4, while the answer quality assessment examined multiple dimensions including relevance, conciseness, readability, completeness, and extrinsic hallucination using the same scale of 1 to 4, while the answer quality assessment examined multiple dimensions including relevance, conciseness, readability, completeness, and extrinsic hallucination using the same scale [26].

The evaluation dataset consisted of 117 anonymized legal user queries, carefully selected to represent actual user interactions, including legal research queries from practicing attorneys and librarians. Although generating large-scale, expert-curated queries poses practical challenges due to the significant domain expertise and effort required, our dataset size is realistic and practically representative of typical industry evaluations. Additionally, the selected statistical methods, particularly non-parametric tests such as the WSRT with BH corrections, are robust and specifically suitable for datasets of this scale, ensuring the statistical validity and reliability of our findings. For each query, both systems generated answers in the form of summaries derived from the top-k retrieved documents, accompanied by supporting references. The answers were structured to provide concise legal analyses while maintaining traceability to source documents.

3.1 IRR Metric Analysis

In this study, we investigated and calculated various IRR metrics in human and LLM evaluation data sets (see Table 1). Our goal is to discover a metric or combination of metrics that can guide us in selecting the most suitable LLM model for specific evaluation tasks. Currently, numerous proprietary and open-source LLM models are available for such tasks. Evaluation tasks often involve different measurement levels, including nominal (e.g., categories), ordinal (e.g., rankings), and, to a lesser extent, interval and ratio data. Most evaluations focus on nominal and ordinal data, underscoring the importance of identifying a metric set tailored to these tasks. In addition, IRR metrics typically assume specific distributions of categories and ratings to calculate the percentage of chance agreement. When this assumption is violated, the metrics may not provide the expected insight. For instance, in situations with skewed ratings distributions, relying solely on Krippendorff's Alpha [18] might lead to an underestimation of raters' agreement due to inflated chance agreement calculations. Gwet's AC2 offers a solution by providing a more stable estimation that is less affected by category prevalence, thus giving a truer reflection of IRR.

We evaluated several metrics, including Cohen's Kappa, Krippendorff's Alpha, Spearman's rank correlation, Kendall's Tau, percent agreement and two extensions of Gwet's AC2 with linear weighting schemes (Gwet AC2-L) and quadratic (Gwet AC2-Q). While we did not evaluate every available IRR metric, such as Fleiss's Kappa and the Brennan-Prediger coefficient, our aim was to emphasize the importance of utilizing a diverse set of metrics rather than conducting an exhaustive analysis of all available IRR metrics.

3.2 Statistical Testing

In the context of evaluating LLM outputs for legal RAG systems, we carefully selected five key attributes to compare various nonparametric tests. Relative power, impact of sample size, number of repeated measurements, robustness and practical implications (see Table 2). **Relative Power** was chosen for its critical role in detecting subtle differences between LLM judges or RAG systems, which is essential given the nuanced nature of legal language and the potentially small but significant variations in system performance. **Sample Size Impact** was included due to practical constraints often faced in the legal evaluation of AI, where large datasets may not always be available or feasible to process. **Number of Repeated Measurements** attribute is crucial for aligning the statistical test with our experimental design, which typically involves comparing two systems or versions on the same set of legal queries. **Robustness** was selected to ensure the reliability of our results under various data conditions, acknowledging the diverse and sometimes unpredictable nature of legal text data. Finally, we included **Practical Implications** to provide context on each test's applicability, helping to bridge the gap between statistical theory and the practical challenges of evaluating legal AI systems. Based on these attributes, we determined that the Wilcoxon Signed Rank test [33] is the most appropriate nonparametric test for our specific context of LLM judge evaluation in legal RAG systems (see Table 2).

This test compares two related samples to assess the differences in population mean rank. Calculate the differences between the paired observations, rank them, and derive a test statistic by summing the positive and negative ranks separately. Its importance lies in its ability to analyze nonnormally distributed data, making it a robust alternative to the paired t-test. Additionally, we considered three multiple testing correction methods: Bonferroni [2], Benjamini-Hochberg (B-H) [1], and Holm-Bonferroni [13]. Each balances strictness of correction, error control, and statistical power differently. Bonferroni, the most conservative, offers strong family-wise error rate (FWER) control but lower power, suitable for critical decisions with few comparisons. B-H controls the false discovery rate (FDR) with higher power, which is ideal for exploratory analyses with numerous comparisons. Holm-Bonferroni provides a middle ground. Given

Attribute	Wilcoxon	Sign Test	Mann–Whitney	Friedman
Statistical Power	High	Moderate	High	Moderate
Sample Size Needed	Small	Small	Medium	Medium
Study Design	Paired	Paired	Independent	Repeated
Robustness	High	Very High	High	High
When to Use:				
Best For	A vs B comparison *(magnitude matters)*	A vs B comparison *(simple win/loss)*	Two groups *(independent data)*	Multiple systems *(3+ comparisons)*

Table 2: Statistical test comparison for LLM evaluation. Our choice: Wilcoxon for paired A/B testing with magnitude sensitivity.

our study's exploratory nature and multiple performance aspects, we selected the B-H method to balance error control and detection of significant differences.

Metric	Skewness
Relevance	−0.4895
Completeness	1.0569
Extrinsic Hallucinations	3.7119
Readability	2.1468
Correctness	0.0898
Inaccurate Hallucinations	5.1269

Table 3: Distribution skewness of evaluation metrics used in RQ2. The strong right skew in hallucination-related metrics motivates non-parametric tests.

We used the Wilcoxon signed-rank test with B-H correction to compare two systems (A and B) using various metrics on 117 queries of varying complexity. Using GPT-4o as a judge, we combined direct and pairwise assessments, performing 10 runs per query and taking the majority vote. This approach ensures rigor in identifying significant performance differences while mitigating false positives from multiple comparisons.

4 Results

Our analysis focused on two key aspects: evaluating different LLM judges for their reliability in assessing legal RAG systems (RQ1) and comparing the performance of two RAG systems using statistical methods (RQ2). For RQ1, we examined various inter-rater reliability metrics to determine their effectiveness in selecting appropriate LLM judges. For RQ2, we investigated the application of statistical methods, particularly the WSRT with B-H corrections, to compare system performance across multiple metrics.

4.1 RQ1 Finding: Evaluating LLM Judges

Our analysis of IRR metrics reveals specific recommendations for different evaluation scenarios in the legal domain. For general agreement assessment, K-Alpha proves effective with balanced rating distributions, while Gwet's AC2 is preferred for skewed distributions (Table 4). The correlation metrics in our study specifically measure different aspects of ranking consistency. Spearman's rank correlation evaluates how well LLM judges preserve the relative ordering of document relevance compared to human expert rankings (with GPT4o showing the highest correlation at 0.73), while Kendall's Tau measures pairwise ranking consistency, particularly important for maintaining proper precedential value ordering between documents. For specific tasks, Gwet's AC2 with quadratic weighting demonstrated superior performance (0.78 for GPT4o) in assessing relevance, while Cohen's Kappa remains adequate for binary decisions despite its limitations with skewed data.

When applying these metrics to evaluate different LLM judges, we observed varying strengths and weaknesses that highlight the importance of a multimetric approach. For example, Prometheus2 8x7B exhibits a higher K-Alpha (0.43) than the Llama model (0.32), suggesting a better overall agreement. However, for ordinal data where rank preservation is crucial, Llama's higher Spearman and Kendall Tau values indicate superior performance in maintaining relative ordering. Similarly, Mistral stands out with its higher agreement in K-Alpha and Gwet's linear coefficient, compared to models like Claude Opus3, which shows moderate performance across most metrics. These contrasting results demonstrate the potential pitfalls of relying on a single metric and reinforce our recommendation for using multiple metrics in combination, with particular emphasis on Gwet's AC2 for skewed distributions, rank correlations for ordering preservation, and task-specific metric selection based on the nature of the legal evaluation task.

> **Takeaway.** When selecting LLM judges for legal RAG evaluation, avoid relying on single metrics like Krippendorff's alpha in skewed distributions. Instead, use Gwet's AC2 for agreement assessment and rank correlation coefficients for ordering consistency—this multimetric approach reveals nuanced judge capabilities that single metrics miss.

4.2 RQ2 Finding: Comparing RAG Systems

Based on our comprehensive analysis, we recommend: (1) using non-parametric tests like the WSRT for comparing legal RAG systems due to the typically skewed nature of evaluation metrics, (2) applying B-H corrections when conducting multiple comparisons to control false discovery rates while maintaining statistical power, and (3) evaluating systems across multiple quality dimensions to capture the nuanced requirements of legal applications. These recommendations emerge from our systematic comparison of two RAG

LLM Judge	Percent Agr.	Cohen κ	Krippendorff α	Gwet's AC2 (lin)	Gwet's AC2 (quad)	Spearman	Kendall τ
GPT4o	**0.56**	**0.35**	**0.70**	**0.63**	**0.78**	**0.73**	**0.66**
LLaMA2-70B	0.26	0.07	0.32	0.26	0.47	0.68	0.61
Claude Opus3	0.40	0.21	0.43	0.35	0.48	0.64	0.56
Mistral	0.42	0.21	0.52	0.50	0.70	0.64	0.57
Prometheus2-7B	0.30	0.13	0.06	0.17	0.27	0.42	0.37
Prometheus2-8x7B	0.40	0.17	0.43	0.36	0.46	0.53	0.46

Table 4: LLM judge agreement and ranking consistency on Search Relevancy. Best per column in bold. κ: Cohen's kappa; α: Krippendorff's alpha. This supports the RQ1 findings discussed in the text.

systems, where we focused on both distribution characteristics and hypothesis testing with appropriate corrections.

Our statistical analysis examined the distribution of different evaluation metrics to inform our statistical approach. As shown in Table 3, metrics exhibited varying degrees of skewness, from slight left skewness in 'Relevance' (-0.49) to strong right skew in 'Inaccurate Hallucinations' (5.13), confirming the appropriateness of our choice of nonparametric tests.

Using the Wilcoxon signed rank test with B-H corrections for multiple comparisons, we conducted a comprehensive comparison of Systems A and B across all metrics. The results revealed distinct patterns of superiority between the systems. System B demonstrated significant advantages in relevance (adjusted p-value = 0.0358), completeness (adjusted p-value = 1.215e-18), and Correctness (adjusted p-value < 0.05). In contrast, System A showed superior performance in Extrinsic Hallucinations (adjusted p-value = 0.0204) and readability (adjusted p-value = 0.01997). Neither system showed a significant advantage in Inaccurate Hallucinations (adjusted p-values > 0.05 for both hypotheses).

These findings highlight the importance of considering multiple quality dimensions when evaluating legal RAG systems. While System B excelled in crucial accuracy-related metrics for legal reliability, System A demonstrated strengths in presentation and hallucination prevention. Balanced performance across different metrics reflects the inherent trade-offs in optimizing RAG systems for legal applications, where both factual accuracy and readability are essential for professional use.

Takeaway. Legal RAG systems exhibit inherent trade-offs between precision and presentation quality. The Wilcoxon Signed-Rank Test with Benjamini-Hochberg corrections provides the statistical rigor needed to detect these nuanced differences across multiple quality dimensions simultaneously.

5 Conclusion

In conclusion, this study demonstrates the viability and challenges of using LLMs as evaluative judges for domain-specific (e.g., legal) RAG systems. Our research makes three key contributions: First, we establish that a multimetric approach to evaluating LLM judges is essential, with different metrics capturing distinct aspects of reliability. Gwet's AC2 proved particularly effective for skewed distributions common in legal evaluations, while rank correlation metrics better captured ordering relationships crucial for legal precedent. Second, we demonstrate that robust statistical analysis, particularly the Wilcoxon Signed-Rank Test with Benjamini-Hochberg corrections, is crucial for meaningful system comparisons. This approach effectively balances statistical power with false discovery control in multiple comparison scenarios. Third, our findings highlight the importance of comprehensive evaluation frameworks that consider both quantitative metrics and domain-specific requirements. While LLM judges can significantly reduce evaluation time, their effective deployment requires careful consideration of reliability metrics, statistical methods, and domain-specific constraints.

References

[1] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[2] Carlo Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze* 8 (1936), 3–62.

[3] Robert L Brennan and Dale J Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement* 41, 3 (1981), 687–699.

[4] Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With Little Power Comes Great Responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 9263–9274. https://doi.org/10.18653/v1/2020.emnlp-main.745

[5] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shan Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. *ArXiv* abs/2308.07201 (2023). https://api.semanticscholar.org/CorpusID:260887105

[6] Shi-Yi Chen, Zhe Feng, and Xiaolian Yi. 2017. A general introduction to adjustment for multiple comparisons. *Journal of Thoracic Disease* 9 (06 2017), 1725–1729. https://doi.org/10.21037/jtd.2017.05.34

[7] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[8] Ehsan Doostmohammadi, Oskar Holmström, and Marco Kuhlmann. 2024. How Reliable Are Automatic Evaluation Methods for Instruction-Tuned LLMs? arXiv:2402.10770 [cs.CL] https://arxiv.org/abs/2402.10770

[9] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[10] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL] https://arxiv.org/abs/2312.10997

[11] Kilem L. Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *The British journal of mathematical and statistical psychology* 61 Pt 1 (2008), 29–48. https://api.semanticscholar.org/CorpusID:13915043

[12] Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *Brit. J. Math. Statist. Psych.* 61, 1 (2008), 29–48.

[13] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[14] M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1/2 (1938), 81–93. http://www.jstor.org/stable/2332226

[15] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1-2 (1938), 81–93.

[16] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. 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, 4334–4353. https://aclanthology.org/2024.emnlp-main.248

[17] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking Cognitive Biases in Large Language Models as Evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 517–545. https://doi.org/10.18653/v1/2024.findings-acl.29

[18] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

[19] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*.

[20] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[21] 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

[22] Evan Miller. 2024. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. arXiv:2411.00640 [stat.AP] https://arxiv.org/abs/2411.00640

[23] 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* (Barcelona, Spain) *(KDD '24)*. Association for Computing Machinery, New York, NY, USA, 2307–2317. https://doi.org/10.1145/3637528.3671883

[24] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. *ArXiv* abs/2404.13076 (2024). https://api.semanticscholar.org/CorpusID:269293311

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[26] Bobby Puglia. 2024. Setting Our Own Standards: Guidelines for AI-Powered Products. LinkedIn. https://www.linkedin.com/pulse/setting-our-own-standards-guidelines-ai-powered-products-bobby-puglia-thfme Accessed: [Insert access date here].

[27] Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation Metrics in the Era of GPT-4: Reliably Evaluating Large Language Models on Sequence to Sequence Tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8776–8788. https://doi.org/10.18653/v1/2023.emnlp-main.543

[28] Charles Spearman. 1961. The proof and measurement of association between two things. (1961).

[29] C Spearman. 2010. The proof and measurement of association between two things. *International Journal of Epidemiology* 39, 5 (10 2010), 1137–1150. https://doi.org/10.1093/ije/dyq191 arXiv:https://academic.oup.com/ije/article-pdf/39/5/1137/18481215/dyq191.pdf

[30] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. *ArXiv* abs/2404.18796 (2024). https://api.semanticscholar.org/CorpusID:269449458

[31] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 9440–9450. https://doi.org/10.18653/v1/2024.acl-long.511

[32] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. http://www.jstor.org/stable/3001968

[33] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin* 1, 6 (1945), 80–83.

[34] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL] https://arxiv.org/abs/1904.09675

[35] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2020, 29 pages.

[36] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631* (2023).

Appendix: Why Gwet's AC2 is More Reliable Under Skewed Labels

Inter-rater reliability coefficients are generally defined in terms of the *observed agreement* A_o and the *expected agreement* A_e that would be obtained "by chance." The generic form is:

$$\text{Coefficient} \;=\; \frac{A_o - A_e}{1 - A_e}.$$

Definitions.

- A_o: the empirical proportion of times that raters agree (possibly weighted for ordinal data).
- A_e: the chance agreement, estimated from the overall distribution of labels.
- $D_o = 1 - A_o$: the observed disagreement.
- $D_e = 1 - A_e$: the expected disagreement.

Krippendorff's α is commonly expressed in terms of disagreement:

$$\alpha \;=\; 1 - \frac{D_o}{D_e}.$$

Problem under skew. Suppose the ratings are highly imbalanced (e.g., 90% of responses fall into a single Likert category). Then the marginal probability distribution is dominated by that one category, which makes $A_e \to 1$. Consequently, $1 - A_e \to 0$ (equivalently, $D_e \to 0$). This causes the denominator in both κ and α to shrink toward zero, depressing the coefficient even when raters actually agree most of the time. This is the well-known *prevalence paradox*.

How AC2 differs. Gwet's AC2 avoids this instability by modeling chance disagreement directly. Instead of relying on squared marginals, it normalizes by the *available chance disagreement*, which is proportional to

$$1 - \sum_k p_k^2,$$

where p_k is the overall proportion of ratings in category k. This term measures how much variability is present in the label distribution. It only vanishes when all ratings fall into a single category, and it decreases at the same rate as the actual difficulty of the task.

Conclusion. Because AC2's denominator reflects the true amount of potential chance disagreement, it remains stable under skewed distributions. In contrast, κ and α normalize by a vanishing term when one category dominates, which can yield deceptively low reliability scores despite high observed agreement.