Novelty Assessment Report

Paper: Preference Leakage: A Contamination Problem in LLM-as-a-judge

PDF URL: https://openreview.net/pdf?id=grIvSXVJ65

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-04

Abstract

Large Language Models (LLMs) as judges and LLM-based data synthesis have emerged as two fundamental LLM-driven data annotation methods in model development. While their combination significantly enhances the efficiency of model training and evaluation, little attention has been given to the potential contamination brought by this new model development paradigm. In this work, we expose preference leakage, a contamination problem in LLM-as-a-judge caused by the relatedness between the synthetic data generators and LLM-based evaluators. To study this issue, we first define three common relatednesses between the data generator LLM and the judge LLM: being the same model, having an inheritance relationship, and belonging to the same model family. Through extensive experiments, we empirically confirm the bias of judges towards their related student models caused by preference leakage across multiple LLM baselines and benchmarks. Further analysis suggests that preference leakage is a pervasive and real-world problem that is harder to detect compared to previously identified biases in LLM-as-a-judge scenarios. All of these findings imply that preference leakage is a widespread and challenging problem in the area of LLM-as-a-judge.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Bias Detection in LLM-Based Evaluation Systems**

A total of **50 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Bias in LLM Content Generation and Outputs**
- **Bias in LLM-as-Evaluator Systems**
- **Evaluation Methodology and Benchmark Quality**
- **Comprehensive LLM Evaluation Studies**
- **Tangentially Related Topics**

Complete Taxonomy Tree

- Bias Detection in LLM-Based Evaluation Systems Survey Taxonomy
- Bias in LLM Content Generation and Outputs
  - General Social Bias Characterization in LLMs (4 papers)
  - [1] Bias and fairness in large language models: A survey (Isabel O. Gallegos, 2024) View paper
  - [2] A survey on fairness in large language models (Li Yingji, 2023) View paper
  - [5] The life cycle of large language models in education: A framework for understanding sources of bias (Jinsook Lee, 2024) View paper
  - [18] Should chatgpt be biased? challenges and risks of bias in large language models (Emilio Ferrara, 2023) View paper
  - Domain-Specific Bias Manifestations (6 papers)
  - [4] Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation (Jizhi Zhang, 2023) View paper
  - [6] Bias testing and mitigation in llm-based code generation (Dong Huang, 2024) View paper
  - [13] Bias Assessment and Mitigation in LLM-based Code Generation (Huang, 2023) View paper
  - [17] Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era (Sunhao Dai, 2024) View paper
  - [26] Large language model as attributed training data generator: A tale of diversity and bias (Yu Yue, 2023) View paper
  - [42] Fairt2i: Mitigating social bias in text-to-image generation via large language model-assisted detection and attribute rebalancing (Sato, 2025) View paper
  - Implicit and Latent Bias Detection Methods (3 papers)
  - [10] Fine-Grained Bias Detection in LLM: Enhancing detection mechanisms for nuanced biases (Mohanty, 2025) View paper
  - [15] Semantic and Structural Analysis of Implicit Biases in Large Language Models: An Interpretable Approach (Renhan Zhang, 2025) View paper
  - [19] Investigating Implicit Bias in Large Language Models: A Large-Scale Study of Over 50 LLMs (Kumar Divyanshu, 2024) View paper
  - Bias Measurement Frameworks and Benchmarks (6 papers)
  - [8] The large language model (LLM) bias evaluation (age bias) (Y Duan, 2024) View paper
  - [9] LIBRA: Measuring Bias of Large Language Model from a Local Context (Bo Pang, 2025) View paper
  - [12] Ranking of large language model (llm) regional bias (Y Duan, 2023) View paper
  - [44] Beats: Bias evaluation and assessment test suite for large language models (Alok Abhishek, 2025) View paper
  - [47] Evaluate Bias without Manual Test Sets: A Concept Representation Perspective for LLMs (Gao Lang, 2025) View paper
  - [49] Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model (Wang Sibo, 2024) View paper

  - Bias Detection and Mitigation Tools (5 papers)
  - [7] Bias Detection and Robustness Testing in Large Language Models: An Experimental Framework' (Talavera, 2025) View paper
  - [24] Biasguard: A reasoning-enhanced bias detection tool for large language models (Zhiting Fan, 2025) View paper
  - [31] BIASINSPECTOR: Detecting Bias in Structured Data through LLM Agents (Li Haoxuan, 2025) View paper
  - [34] Leveraging Large Language Models for Bias Detection and Mitigation in Data Analytics Models (Manideep Marripudugala, 2025) View paper
  - [43] Biasalert: A plug-and-play tool for social bias detection in llms (Chen, 2024) View paper
- Bias in LLM-as-Evaluator Systems
  - Positional and Ordering Biases (3 papers)
  - [3] Large language models are not fair evaluators (Cai, 2024) View paper
  - [22] Diagnosing bias and instability in llm evaluation: A scalable pairwise meta-evaluator (Catalin Anghel, 2025) View paper
  - [28] Judging the judges: A systematic study of position bias in llm-as-a-judge (Shi Lin, 2024) View paper
  - Self-Preference and Model Relatedness Biases ★ (2 papers)
  - [0] Preference Leakage: A Contamination Problem in LLM-as-a-judge (Anon et al., 2026) View paper
  - [37] Self-preference bias in llm-as-a-judge (Wataoka Koki, 2024) View paper
  - Fairness and Demographic Bias in Evaluation (3 papers)
  - [20] FairEval: Evaluating Fairness in LLM-Based Recommendations with Personality Awareness (Lian XiaoLi, 2025) View paper
  - [23] CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System (Yashar Deldjoo, 2024) View paper
  - [48] Challenging Fairness: A Comprehensive Exploration of Bias in LLM-Based Recommendations (Shahnewaz Karim Sakib, 2024) View paper
  - Context and Surface-Level Sensitivity Biases (3 papers)
  - [30] Don't Judge Code by Its Cover: Exploring Biases in LLM Judges for Code Evaluation (Moon Ji-Won, 2025) View paper
  - [33] Curse of knowledge: When complex evaluation context benefits yet biases llm judges (LI Weiyuan, 2025) View paper
  - [39] I Think, Therefore I Am Under-Qualified? A Benchmark for Evaluating Linguistic Shibboleth Detection in LLM Hiring Evaluations (Roosta, 2025) View paper
  - Evaluation Reliability and Consistency Issues (3 papers)
  - [16] Evaluating and Mitigating LLM-as-a-judge Bias in Communication Systems (Gao, 2025) View paper
  - [32] Justice or prejudice? quantifying biases in llm-as-a-judge (Ye Jiayi, 2024) View paper
  - [36] Large language models are inconsistent and biased evaluators (Stureborg, 2024) View paper
  - Cognitive and Decision Biases in LLM Agents (2 papers)
  - [27] Towards Cognitive Synergy in LLM-Based Multi-Agent Systems: Integrating Theory of Mind and Critical Evaluation (Adam Kostka, 2025) View paper
  - [38] LLM Agents Can Be Choice-Supportive Biased Evaluators: An Empirical Study (Zhuang Nan, 2025) View paper
- Evaluation Methodology and Benchmark Quality
  - Benchmark Limitations and Dataset Quality (3 papers)
  - [25] Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence (Timothy R McIntosh, 2024) View paper
  - [41] XFacta: Contemporary, Real-World Dataset and Evaluation for Multimodal Misinformation Detection with Multimodal LLMs (Han Zeyu, 2025) View paper
  - [45] Revisiting Multi-Modal LLM Evaluation (Jian Lu, 2024) View paper
  - Evaluation Validity and Reliability Concerns (2 papers)
  - [40] Audit-style framework for evaluating bias in large language models (Peter Baldwin, 2025) View paper
  - [50] Llm-evaluation tropes: Perspectives on the validity of llm-evaluations (Dietz, 2025) View paper
  - Alternative Evaluation Approaches (2 papers)
  - [14] PRE: A Peer Review Based Large Language Model Evaluator (Chu, 2024) View paper
  - [29] Building Trust in Mental Health Chatbots: Safety Metrics and LLM-Based Evaluation Tools (Park Jung In, 2024) View paper
- Comprehensive LLM Evaluation Studies (3 papers)
  - [21] A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets (Bari, 2023) View paper
  - [35] Does Reasoning Introduce Bias? A Study of Social Bias Evaluation and Mitigation in LLM Reasoning (Wu Xuyang, 2025) View paper
  - [46] Biased by Design? Evaluating Bias and Behavioral Diversity in LLM Annotation of Real-World and Synthetic Hotel Reviews (Maria C. Voutsa, 2025) View paper
- Tangentially Related Topics (1 papers)
  - [11] Creativity in LLM-based Multi-Agent Systems: A Survey (Lin Yi-cheng, 2025) View paper

Narrative

Core task: Bias detection in LLM-based evaluation systems. The field has organized itself around several major branches that reflect different facets of bias. One branch examines bias in LLM content generation and outputs, focusing on how models produce biased text across domains such as code generation, recommendations, and general language tasks. A second branch targets bias in LLM-as-evaluator systems, where models serve as judges or scorers of other outputs, introducing concerns about self-preference, positional effects, and inconsistent scoring. A third branch addresses evaluation methodology and benchmark quality, questioning whether existing test suites adequately capture bias phenomena or inadvertently perpetuate flawed assumptions. Comprehensive LLM evaluation studies provide broad empirical assessments that cut across multiple bias types, while tangentially related topics touch on fairness in adjacent areas like information retrieval or mental health applications. Representative works such as Not Fair Evaluators[3] and Self-Preference Bias[37] illustrate how evaluator-specific biases can distort automated assessment, whereas surveys like Bias Fairness Survey[1] and Fairness Survey[2] map the broader landscape of bias concerns in LLM deployment.

Particularly active lines of work explore the tension between using LLMs as convenient evaluators and the risk that these models favor their own outputs or exhibit systematic preferences tied to model architecture or training lineage. Studies on self-preference and model relatedness biases reveal that evaluators may assign higher scores to responses generated by themselves or by closely related models, undermining the objectivity of automated judgments. Preference Leakage[0] sits squarely within this cluster, investigating how subtle cues in generated text can inadvertently signal a model's identity to an evaluator, thereby triggering preferential scoring. This work complements Self-Preference Bias[37], which documents the phenomenon more broadly, and contrasts with efforts like FairEval[20] that propose mitigation strategies to reduce evaluator partiality. Together, these studies highlight an open question: whether LLM-based evaluation can be made sufficiently robust to replace human judgment, or whether inherent biases will always require careful calibration and transparency measures.

Related Works in Same Category

No comparison data available.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Preference leakage problem definition

Description: The authors formally define preference leakage as a new contamination issue that arises when the LLM used for synthetic data generation and the LLM used as an evaluator are related, causing systematic bias in evaluation scores. They identify three types of relatedness: being the same model, having an inheritance relationship, and belonging to the same model family.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 2: Empirical validation of preference leakage bias

Description: The authors perform comprehensive experiments using multiple LLM baselines and benchmarks (Arena-Hard and AlpacaEval 2.0) to empirically confirm that judge LLMs exhibit systematic bias toward their related student models. They introduce the preference leakage score metric to quantify this bias across different scenarios.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: Analysis of preference leakage mechanisms and characteristics

Description: The authors investigate the underlying mechanisms of preference leakage through recognition experiments and category analyses. They demonstrate that preference leakage is particularly hard to detect, especially affecting subjective questions and judgment dimensions, and that judge LLMs cannot reliably recognize their related student models' generations.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Preference Leakage: A Contamination Problem in LLM-as-a-judge View paper
- [1] Bias and fairness in large language models: A survey View paper
- [2] A survey on fairness in large language models View paper
- [3] Large language models are not fair evaluators View paper
- [4] Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation View paper
- [5] The life cycle of large language models in education: A framework for understanding sources of bias View paper
- [6] Bias testing and mitigation in llm-based code generation View paper
- [7] Bias Detection and Robustness Testing in Large Language Models: An Experimental Framework' View paper
- [8] The large language model (LLM) bias evaluation (age bias) View paper
- [9] LIBRA: Measuring Bias of Large Language Model from a Local Context View paper
- [10] Fine-Grained Bias Detection in LLM: Enhancing detection mechanisms for nuanced biases View paper
- [11] Creativity in LLM-based Multi-Agent Systems: A Survey View paper
- [12] Ranking of large language model (llm) regional bias View paper
- [13] Bias Assessment and Mitigation in LLM-based Code Generation View paper
- [14] PRE: A Peer Review Based Large Language Model Evaluator View paper
- [15] Semantic and Structural Analysis of Implicit Biases in Large Language Models: An Interpretable Approach View paper
- [16] Evaluating and Mitigating LLM-as-a-judge Bias in Communication Systems View paper
- [17] Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era View paper
- [18] Should chatgpt be biased? challenges and risks of bias in large language models View paper
- [19] Investigating Implicit Bias in Large Language Models: A Large-Scale Study of Over 50 LLMs View paper
- [20] FairEval: Evaluating Fairness in LLM-Based Recommendations with Personality Awareness View paper
- [21] A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets View paper
- [22] Diagnosing bias and instability in llm evaluation: A scalable pairwise meta-evaluator View paper
- [23] CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System View paper
- [24] Biasguard: A reasoning-enhanced bias detection tool for large language models View paper
- [25] Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence View paper
- [26] Large language model as attributed training data generator: A tale of diversity and bias View paper
- [27] Towards Cognitive Synergy in LLM-Based Multi-Agent Systems: Integrating Theory of Mind and Critical Evaluation View paper
- [28] Judging the judges: A systematic study of position bias in llm-as-a-judge View paper
- [29] Building Trust in Mental Health Chatbots: Safety Metrics and LLM-Based Evaluation Tools View paper
- [30] Don't Judge Code by Its Cover: Exploring Biases in LLM Judges for Code Evaluation View paper
- [31] BIASINSPECTOR: Detecting Bias in Structured Data through LLM Agents View paper
- [32] Justice or prejudice? quantifying biases in llm-as-a-judge View paper
- [33] Curse of knowledge: When complex evaluation context benefits yet biases llm judges View paper
- [34] Leveraging Large Language Models for Bias Detection and Mitigation in Data Analytics Models View paper
- [35] Does Reasoning Introduce Bias? A Study of Social Bias Evaluation and Mitigation in LLM Reasoning View paper
- [36] Large language models are inconsistent and biased evaluators View paper
- [37] Self-preference bias in llm-as-a-judge View paper
- [38] LLM Agents Can Be Choice-Supportive Biased Evaluators: An Empirical Study View paper
- [39] I Think, Therefore I Am Under-Qualified? A Benchmark for Evaluating Linguistic Shibboleth Detection in LLM Hiring Evaluations View paper
- [40] Audit-style framework for evaluating bias in large language models View paper

- [41] XFacta: Contemporary, Real-World Dataset and Evaluation for Multimodal Misinformation Detection with Multimodal LLMs View paper
- [42] Fairt2i: Mitigating social bias in text-to-image generation via large language model-assisted detection and attribute rebalancing View paper
- [43] Biasalert: A plug-and-play tool for social bias detection in llms View paper
- [44] Beats: Bias evaluation and assessment test suite for large language models View paper
- [45] Revisiting Multi-Modal LLM Evaluation View paper
- [46] Biased by Design? Evaluating Bias and Behavioral Diversity in LLM Annotation of Real-World and Synthetic Hotel Reviews View paper
- [47] Evaluate Bias without Manual Test Sets: A Concept Representation Perspective for LLMs View paper
- [48] Challenging Fairness: A Comprehensive Exploration of Bias in LLM-Based Recommendations View paper
- [49] Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model View paper
- [50] Llm-evaluation tropes: Perspectives on the validity of llm-evaluations View paper
- [51] Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges View paper
- [52] Holistic evaluation of language models View paper
- [53] LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models View paper
- [54] Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms View paper
- [55] Do LLM Evaluators Prefer Themselves for a Reason? View paper
- [56] Improving drug-drug interaction prediction via in-context learning and judging with large language models View paper
- [57] SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal Behaviors View paper
- [58] Time To Impeach LLM-as-a-Judge: Programs are the Future of Evaluation View paper
- [59] Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination View paper
- [60] One token to fool llm-as-a-judge View paper
- [61] Rankers, judges, and assistants: Towards understanding the interplay of llms in information retrieval evaluation View paper
- [62] Replacing judges with juries: Evaluating llm generations with a panel of diverse models View paper
- [63] Play favorites: A statistical method to measure self-bias in llm-as-a-judge View paper
- [64] Toward Trustworthy Difficulty Assessments: Large Language Models as Judges in Programming and Synthetic Tasks View paper
- [65] Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge View paper
- [66] A survey on data contamination for large language models View paper
- [67] Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs View paper
- [68] Evading data contamination detection for language models is (too) easy View paper
- [69] Benchmark data contamination of large language models: A survey View paper
- [70] Livebench: A challenging, contamination-free llm benchmark View paper
- [71] Does data contamination detection work (well) for llms? a survey and evaluation on detection assumptions View paper
- [72] Towards Reliable Benchmarking: A Contamination Free, Controllable Evaluation Framework for Multi-step LLM Function Calling View paper
- [73] Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models View paper
- [74] Time travel in llms: Tracing data contamination in large language models View paper