Who's to Blame for Citation Biases? Tracing Citation Practices Back to Reviewer Preferences

Anonymous ACL submission

Abstract

Insularity, recency bias and over-representation of English have increasingly been recognized as critical challenges in AI research, potentially limiting the poly-vocal nature of human intelligence. These biases are often reproduced through citation practices. Far from being a matter of pure individual choice, such practices are shaped by the expectations and pressures of reviewers. Drawing on large-scale review data of major machine learning conferences from OpenReview, we quantify the influence of peer review on citation biases. Our analysis reveals that reviewers reinforce existing biases by suggesting authors to cite more newer, intra-field, and English-centric works in their publications. Our study thus contributes to a deeper understanding of the dynamics within peer review and citation practices, raising awareness for improving transparency and accountability in academic publishing. [1]

1 Introduction

Citation practices play a crucial role in shaping which topic gain visibility within academic communities. They also reflect and reinforce certain biases such as who gets cited and which topics are prioritized. Observed citation biases in Machine Learning (ML) and Natural Language Processing (NLP) include a tendency toward increased intra-field referencing (Wahle et al., 2023), a preference for more recent publications (Wahle et al., 2024), and a strong English-language bias (Hershcovich et al., 2022). These biases are often attributed solely to authors, yet not all citation decisions rest entirely in their hands. Significant influences can come from reviewers' preferences, a factor that is frequently overlooked in discussions about citation bias (Figure 1). Although several factors during the peer review process may

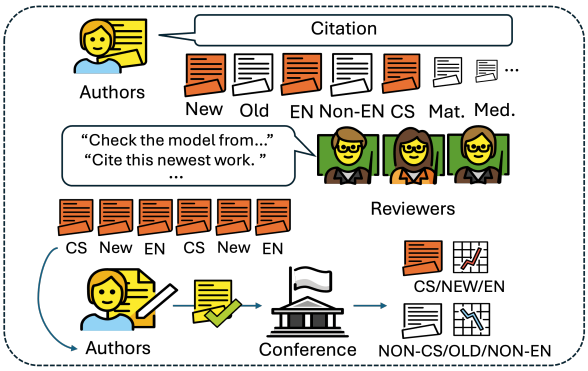

Figure 1: A simplified flow of the incorporation of English-centric, recency, and insularity biases into the research landscape via the peer review suggestions.

contribute to the observed recency, insularity, and English-centric biases in NLP and ML research, our study highlights that reviewer-recommended citations, as a form of "scientific currency" (Khurana, 2023) post the most direct and measurable effect. Previous works that look into this direction mainly focused on reviewer's self-citation practice, however, reviewer-recommended citations often go beyond self-promotion. For instance, reviewers might, based on their personal preferences or standards of paper quality, suggest papers from U.S. institutions, or popular topics such as LLMs.

In this work, we investigate citation recommendations emerged during peer reviews at major NLP and ML conferences. We examine how often reviewers suggest additional citations and the extent to which authors incorporate. We further analyze the characteristics of the recommended citations, including their institutional affiliations, publication years, and disciplinary focus. Finally, we explore whether the number of recommended citations is associated with the paper's final decision.

Our analyses show that citation recommendations form an essential component of reviewing practice. More than **55%** of all the submitted papers received at least one citation recommendation

[1] All the datasets and the code to reproduce the experiments are available on GitHub: *link anonymized for review*

and up to **79%** of the authors incorporated at least one of such citations. Our results also indicate that reviewers tend to reinforce skewed citation patterns, favoring papers from U.S. institutions and recent work (past 3 years). Despite ML's growing reach across disciplines, reviewers increasingly overlook related fields. Notably, rejected papers receive more citation suggestions, implying that reviewers use them to signal dominant paradigms. Although several authors and guidelines have urged more inclusive citation practices, breaking out of these entrenched patterns remains difficult when gatekeepers expect alignment with established citation norms.

2 Related Work

Biases in citation practices and in peer review have each been studied extensively. Until recently, however, it has been difficult to directly connect these two phenomena. The increasing transparency of peer review processes makes it possible to look into this relationship by examining additional citations emerged during the peer review process.

Citation Patterns. Researchers have studied citation biases from various dimensions such as authors' affiliation (Abdalla et al., 2023) and self-citation practices (Sala and Brooks, 2008). AI as a fast-growing field, has also attracted great attention in citation studies. Wahle et al. (2023, 2024) highlighted increasing recency and insularity bias in NLP citations, raising concerns about a lack of diversity. Gnewuch (2024) quantifies citational influence from industry on research trajectories. Recent research has also examined how LLMs heightened bias towards highly cited papers (Algaba et al., 2025b) and recent publications with shorter titles and fewer authors (Algaba et al., 2025a).

Peer Review. Peer review has also been studied extensively, including the geographic distribution of reviewers (Snell and Spencer, 2005; Lee et al., 2013), the influence of author reputation (Bornmann and Daniel, 2008; Okike et al., 2016), and the impact of publication venues (Squazzoni et al., 2020). With the recent rise of generative AI, Liang et al. (2024) suggest that up to 16.9% of peer reviews in ML conferences are written by LLMs. Ebadi et al. (2025) thus underscores the need for clear reviewer guidelines to address the ethical concerns and practical challenges in the era of LLMs.

Reviewer-Recommended Citations Levis and Leentjens (2020); Thombs et al. (2015) found that reviewers often request citations to their works. Similarly, Fong and Wilhite (2017) suggest citing a reviewer's own work proactively significantly increase review scores even after controlling for paper quality and reviewer expertise. Although these studies have explored the influence of peer review on citation practices, all of them focused on self-citation, leaving a gap in understanding the broader patterns of reviewer-suggested citations.

3 Methodology

Identifying citations recommended by reviewers requires processing largely unstructured text from reviewer comments, which posts challenges compared to detecting self-citations in a paper. Therefore, we tailored our methods to combine manual annotation, LLM-based classification, and regex-based approaches to accurately extract information.

3.1 Data Collection

We used the OpenReview API to obtain peer review data and paper submission information (OpenReview.net, 2024). The dataset consists of **~59,1k** reviews and **~14k** papers of major ML and NLP conferences ranging from 2013 to 2024, including: EMNLP, ICLR and NeurIPS. The choice was made due to the prominence of the conference and accessibility of structured peer review data. In particular, the main venues of our study includes EMNLP 2023, ICLR 2023, NeurIPS 2023 and 2024, with limited data from ICLR 2013, 2014, 2017 and 2019 (See Appendix §A.1 for detailed statistics).

3.2 Annotation and Data Analysis

Deciding whether a reviewer suggested additional literature is non-trivial, as recommendations can be implicit (e.g., "please check Joe Doe's newest paper"), making regex heuristics error-prone. We therefore used LLM-based classification, testing five prompts on `LLaMA 3.1 8B` and seven prompts on `LLaMA 3.1 70B` (prompts in Appendix § A.3). To select the optimal model-prompt combination for each venue, we manually annotated a balanced gold-standard sample of 100 reviews (50 positive, 50 negative) from EMNLP, ICLR, and NeurIPS, then evaluated each combination using Accuracy, F_1, and AUC. The best AUC ranged from 0.87 to 0.94. We used the best model-prompts to classify all reviews per venue, filtered those containing recommended citations, and manually cleaned a ran-

dom sample to obtain a citation list for later analysis. Sample size was determined with 95% confidence and 5% margin of error (Thompson, 2012). Using this clean list, we could examine whether authors incorporated reviewer-recommended citations in their final PDFs (§4.1).

To analyze reviewer-recommended citations, we used different methods per bias category. For institutional distribution, we obtained affiliation data of submitted papers via the OpenReview API, and for recommended citations, we collected the first author's institution from Google Scholar (§4.2). To assess recency bias, we used regex to detect typical citing patterns (e.g., (Author, 2022b)", arXiv:2202...", etc.), extracted publication years, and compared them to the average reference year in the submitted papers (§4.3). For insularity trends, recommended citations were queried via the Semantic Scholar (S2) API (Semantic Scholar, n.d.) to obtain primary fields of study from S2's classifier (≈86% accuracy) (Wahle et al., 2023) (§4.4). We also tested correlations between paper decisions and citation suggestions: one by acceptance and presence of citations recommendations, the other by number of reviewers suggesting citations.(§4.5).

4 Results

We first set the foundation by showing how often reviewer-recommended citations occur and how likely the authors incorporate such suggestions (§4.1). Building on this, we then follow up with the characteristics of such citations, including their affiliation (§4.2), recency (§4.3), topical distribution (§4.4), and their correlation with paper decision (§4.5).

4.1 Overview of Reviewer-Recommended Citation and Author Incorporation

The percentage of papers receiving at least one citation suggestion from their reviewers ranges from 58.59% to 79.32% (Table 1), suggesting that **for most the submitted papers, at least one of their reviewers suggest citing other paper(s)**. Another interesting tendency is that reviewers of ICLR have become more dedicated to recommending citations to the authors over the years. The result also presents the high proportion of authors who actually included at least such citations in the final submission. This implies **reviewer recommendations have a strong effect on the citation practices in published papers**.

Venue(Year)	Review Rec.	Paper Rec.	Include.
EMNLP'23	31.48%	67.48%	62.17%
ICLR'23	36.99%	79.32%	73.38%
NeurIPS'23	22.09%	64.27%	74.24%
NeurIPS'24	21.44%	58.59%	79.18%
Historical ICLR (avg. '13–'19): 24.6% / 63.7% / –			

Table 1: Citation recommendation from reviewers and incorporation by authors. *Review Rec.*: % of reviews suggesting extra citations; *Paper Rec.*: % of papers that received at least one suggested citation; *Include.*: % of authors that included at least one such citation. (§4.1)

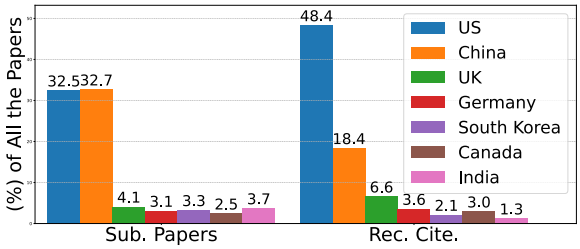

Figure 2: Country distribution of submitted papers vs. reviewer-recommended citations *Sub. Papers*: % of all the submitted papers; *Rec. Cite.*: % of all the reviewer-recommended citations. (§4.2)

4.2 Institutional Distribution

Across all the venues, we observe a consistent over-representation of U.S. among citations recommended by reviewers. While the share of submissions originating from the U.S. is 32.5% on average, its share among reviewer-recommended citations is substantially higher (48.4%) (Figure 2). Another interesting finding from the diachronic ICLR data is that, although other countries, particularly China, have experienced a substantial increase in paper submissions, the share of reviewer-recommended citations attributed to U.S. institutions has remained consistently high (Figure 3). One could argue that papers from the U.S. are better thus getting brought up more often, however, their acceptance rate is not higher compared to that of other countries (See Table **??** in Appendix §B.1).

4.3 Recency Bias Citations Recommendation

Figure 4 shows that across all the venues, **reviewer-recommended citations are much more recent than those already cited by the authors**. Specifically, both the author-cited and reviewer-recommended references in EMNLP papers skew newer compared to ICLR and NeurIPS. Besides, a direct comparison between NeurIPS 2023 and NeurIPS2024 shows an age decrease in both rec-

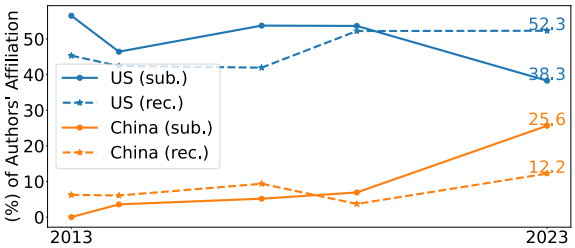

Figure 3: U.S. vs. China in submitted papers and reviewer-recommended citations over the years.(§4.2)

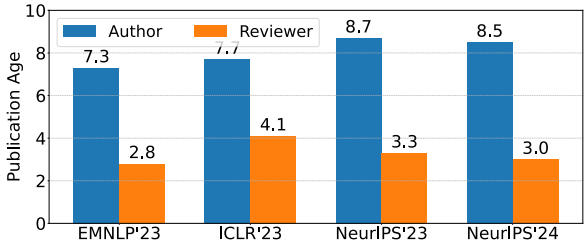

Figure 4: Average ages of authors' reference list vs. reviewer-recommended citations across venues. Higher values indicate older references. (§4.3)

ommended citations and cited references in submitted papers, meaning both authors and reviewers prefer more newer papers in NeurIPS 2024. To quantify the magnitude and significance of the age difference, we computed Cohen's d and Welch's t-test (Cohen, 1988; Welch, 1947) with results being *t-statistic = 10.87*, *p-value < 0.001* and *Cohen's d = 7.69*. **This confirms that reviewer suggestions are not only newer on average, but that the difference is statistically significant and substantially large**.

4.4 Insularity in Citations Recommendation

To answer this question, we use ICLR data as it offers multiple years of reviews. The results show that both reviewer-recommended and author-cited references largely favor Computer Science (CS) papers, which is not surprising given the domain of the conferences. Despite being the majority consistently, the proportion of CS papers in ICLR 2023 has increased ~10% in both author self-cited literature and reviewer-recommended citations compared to 10 years ago (Figure 5). **This result suggests a narrowing of interdisciplinary engagement from reviewer's side**, indicating that the observed insularity in NLP and ML (Wahle et al., 2023) could be amplified by reviewers overlooking neighboring domains which were more visible in earlier years of the venue.

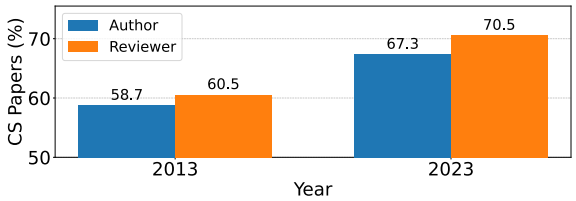

Figure 5: Growing insularity observed in both author self-cited literature and reviewer-recommended citations over the decade (ICLR 2013 - 2023). (§4.4)

4.5 Citation Recommendation and Rejection

Our analysis shows that **rejected papers tend to receive more citation recommendations compared to their accepted counterparts**. In the binary version of the test, the result of $\chi^2 = 5.97$ and $p = 0.0145$ indicates that rejected papers are more likely to get at least one reviewer-recommended citations. In version with finer-grained counts, the $\chi^2 = 35.66$ with $p < 0.0001$, again shows a significant relationship between rejection and the number of citation suggestions. (See Figure 10 and 11 in Appendix §B.3.) This result suggests that reviewers do not request additional citations at random, but rather employ them as a form of critique or a gatekeeping mechanism.

5 Conclusion

Our analysis shows that reviewer-recommended citations are common and influential in top-tier ML and NLP conferences. At ICLR 2023, 79% of authors added reviewer-suggested citations to their camera-ready papers. These recommendations favor U.S. institutions and more recent work, reinforcing institutional and recency biases. EMNLP shows the strongest recency effect, with an average citation age of 2.8 years, reflecting prior observations of "citation amnesia" in NLP (Wahle et al., 2024). From 2013–2023, ICLR citations show growing disciplinary insularity, increasingly citing within CS. Notably, rejected papers receive more citation suggestions, suggesting reviewers may use them as critique or gatekeeping tools.

Overall, our findings suggest that peer review actively shapes citation patterns and, by extension, the visibility and direction of research. Addressing these biases requires greater transparency in citation recommendations, reviewer awareness of their cumulative impact, and the development of review guidelines or tools that promote more balanced and inclusive citation practices.

6 Limitations

While this work offers new insights into reviewer-recommended citations and its influence on citation biases, a few limitations must be acknowledged.

One limitations is data source. Although OpenReview does provide a rich dataset for certain conferences, such as ICLR, it does not have access to many rejected papers in major venues or provide consistent data granularity across years. These problems limited the depth of analysis that could have been conducted for this study.

Additionally, citation recommendations are most of the time embedded in unstructured text fields, such as "questions to authors" or "weakness" section. Only EMNLP 2023 has a field dedicating to missing references. This resulted in relying on LLMs to extract information, which is inherently bias-prone.

Lastly, we included manual annotations for the data processing to ensure a more reliable results. However, it is by no means error- or bias-proof. Furthermore, the scale of manual annotation was necessarily limited due to time and resource constraints.

References

Salwa Abdalla, Moustafa Abdalla, Mohamed Saad, David Jones, Scott Podolsky, and Mohamed Abdalla. 2023. Ethnicity and gender trends of uk authors in the british medical journal and the lancet over the past two decades: a comprehensive longitudinal analysis. *EClinicalMedicine*, 64.

Andres Algaba, Vincent Holst, Floriano Tori, Melika Mobini, Brecht Verbeken, Sylvia Wenmackers, and Vincent Ginis. 2025a. How deep do large language models internalize scientific literature and citation practices? *Preprint*, arXiv:2504.02767.

Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. 2025b. Large language models reflect human citation patterns with a heightened citation bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6829–6864, Albuquerque, New Mexico. Association for Computational Linguistics.

Lutz Bornmann and Hans-Dieter Daniel. 2008. What do we know about the role of peer review in research funding and publication? *Journal of Documentation*.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

S. Ebadi, H. Nejadghanbar, A. R. Salman, and 1 others. 2025. Exploring the impact of generative ai on peer review: Insights from journal reviewers. *Journal of Academic Ethics*.

Eric A. Fong and Allen W. Wilhite. 2017. Authorship and citation manipulation in academic research. *PLOS ONE*, 12(12):1–34.

Max Martin Gnewuch. 2024. What impact does big tech funding have on ai research? a scholarly document analysis.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Mark P. Khurana. 2023. *Citations as Currency: How Does Scientific Capital Influence the Trajectory?*, page 11–15. Cambridge University Press.

Carole J Lee, Cassidy R Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17.

Bryan Levis and Albert FG Leentjens. 2020. Comparison of self-citation by peer reviewers in a medical journal. *BMC Medical Research Methodology*, 20(1):1–5.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel Mcfarland, and James Y. Zou. 2024. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 29575–29620. PMLR.

Kanu Okike, Kyle T Hug, Mininder S Kocher, and S Leopold. 2016. The effect of author prestige on peer review: An experimental study. *BMJ*, 355.

OpenReview.net. 2024. Openreview: Open peer review platform. `https://openreview.net`. Accessed: 2025-05-10.

Sergio Della Sala and Joanna Brooks. 2008. Multi-authors' self-citation: A further impact factor bias? *Cortex*, 44(9):1139–1145.

Semantic Scholar. n.d. Semantic scholar api. `https://api.semanticscholar.org/`. Accessed: 2025-05-14.

Linda Snell and J. Spencer. 2005. Reviewers' perceptions of the peer review process in medical education. *Medical Education*, 39(1):90–97.

Flaminio Squazzoni, Giangiacomo Bravo, Francisco Grimaldo, Daniel García-Costa, Mike Farjam, and Bahar Mehmani. 2020. Peer review and editorial decision-making in the time of covid-19. *Science and Public Policy*, 49(5):791–801.

Brett D. Thombs, Alexander W. Levis, Ilya Razykov, Achyuth Syamchandra, Albert F.G. Leentjens, James L. Levenson, and Mark A. Lumley. 2015. Potentially coercive self-citation by peer reviewers: A cross-sectional study. *Journal of Psychosomatic Research*, 78(1):1–6.

Steven K. Thompson. 2012. *Sampling*, 3rd edition. Wiley Series in Probability and Statistics. Wiley.

Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif Mohammad. 2023. We are who we cite: Bridges of influence between natural language processing and other academic fields. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12896–12913, Singapore. Association for Computational Linguistics.

Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif M. Mohammad. 2024. Citation amnesia: On the recency bias of nlp and other academic fields. *Preprint*, arXiv:2402.12046.

Bernard L. Welch. 1947. The generalization of "student's" problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35.

A Appendix: Data

A.1 Appendix: Data Statistics

Venue	Reviews	Subs.	Avail.	Accept	Reject
EMNLP'23	6,449	4,909	2,020	2,011	9
ICLR'23	14,351	4,874	3,796	1,574	2,222
NeurIPS'23	15,175	12,345	3,395	3,218	177
NeurIPS'24	16,650	15,671	4,238	4,036	202

Table 2: Summary statistics of reviews and papers for main studied venues. *Subs.* = total submissions; *Avail.* = submissions with accessible data.

A.2 Data Collection

For each submission, the metadata collated via the OpenReview API include standard bibliographic details such as paper ID, title, author list, keywords, abstract, TLDR, venue, and file links (PDF, supplementary materials, BibTeX). Each paper is associated with one or more reviews containing detailed textual and numerical feedback. Review-level metadata mainly includes review ID, review summary, strengths, weaknesses, questions, limitations, ethics flags, numerical ratings (e.g., reviewer confidence, contribution, presentation, soundness), review timestamps (creation, modification, decision dates), and various platform-level indicators such as review authorship, signature, visibility, and reply status. The naming of the metadata varies across the venues.

Venue	Reviews	Tokens	Available Papers
ICLR 2019	4,332	~2.08M	1,511
ICLR 2017	1,511	~0.52M	490
ICLR 2014	548	~0.16M	69
ICLR 2013	373	~0.11M	67

Table 3: Statistics for collected reviews and papers of historical ICLR data.

To reproduce the review dataset and submission dataset, one can follow the following steps:

1. Register at OpenReview

2. Locate the venue ID in the web link (usually after "group?id="), e.g., EMNLP/2023, and use it to access the conference via OpenReview API. OpenReview has different API versions, API v2.0 and v1.0. One should use the matching version for the targeted venue.

3. By fetching different notes, one can get peer reviews, decisions, rebuttals, and paper submissions, etc. All data are stored in JSON format by default. We converted the JSON to CSV to support easy merging and analysis.

4. In cases where multiple retrieval scripts introduce data overlap, records can be merged using unique paper or review identifiers.

Detailed instructions to use OpenReview API can be found in the official documentation.[2]

[2] OpenReview API documentation

A.3 Data Annotation and Processing

Prompt	
A	Does this peer review explicitly suggest the authors of the paper to cite any specific literature?
B	Does this peer review suggest the authors of the paper to refer to any other literature?
C	Does this peer review suggest the authors of the paper to refer to any other additional literature? Answer yes or no at the beginning.
D	Does this peer review suggest the authors of the paper to refer to specific literature that are not already discussed in the original paper?
E	Does this peer review suggest the authors of the paper to refer to specific literature that are not already discussed in the original paper? Note that sometimes the reviewers mention some literature in their reviews but those could be already included in the original paper.
F	You are reviewing a peer review for a research paper submitted to OpenReview. Your task is to determine whether the peer review suggests the authors should cite additional relevant papers that are not already included in the original manuscript. Please answer with "Yes" if the review suggests new citations, or "No" if it does not. Then, provide the following: List any suggested papers or references not cited in the manuscript. Briefly explain why those citations might be necessary, based on the reviewer's comments.
G	Does this peer review explicitly say the paper is missing relevant literature? Or should the paper compare to relevant baselines/benchmarks? Answer with yes or no at the very beginning, and give brief explanation.

Table 4: Prompts for detecting reviewer-recommended citations in peer reviews for (§3)

When evaluating prompt-model combinations, AUC captures how well the model distinguishes between reviews that recommend additional literature and those that do not. Among all prompt-model combinations, the one with the highest AUC (breaking ties by Accuracy for more true positives) was selected for each venue. For example, as shown in Figure 6 and Figure 7 on EMNLP 2023, `Llama 70B prompt F` achieved AUC = 0.94 and the highest Accuracy of 0.91, so we used that model-prompt combination to classify all reviews in EMNLP 2023.

B Additional Information for Results

B.1 Additional Information for §4.2

Country	Accept	Reject	Total	Acc. Rate (%)
South Korea	27	22	49	55.10
China	220	254	474	46.41
UK	36	46	82	43.90
Australia	13	17	30	43.33
US	283	372	655	43.21
India	22	29	51	43.14
Singapore	17	23	40	42.50
Canada	16	22	38	42.11
Germany	22	39	61	36.07
France	18	32	50	36.00

Table 5: Acceptance Rates for Top 10 Countries (ICLR 2023). §4.2

B.2 Additional Information for §4.3

To zoom in on how recency bias in citation practice can come from reviewer's suggestion, we created Figure 8 and Figure 9. Figure 8 reveals that in the past, reviewers in ICLR are less likely to recommend recent papers compared to nowadays. When comparing all the recent venues in Figure 9, they show similar patterns. Notably, the spikes around 2000 in ICLR 2023, NeurIPS 2023 and 2024 in Figure 9 result from a set of foundation works in ML field, therefore it is absent in the EMNLP 2023 graph.

B.3 Additional Information for §4.5

For §4.5, the Chi-square statistic is computed using the formula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

where O_{ij} denotes the observed number of papers with decision outcome i and citation suggestion category j (e.g., none, one, two, etc.), and E_{ij} represents the expected count for that cell under the null hypothesis of independence between citation suggestions and paper decisions. Here, r corresponds to the number of decision categories, and c to the number of citation suggestion categories defined in the contingency table. The results are visualized via Figure 10 and Figure 11.

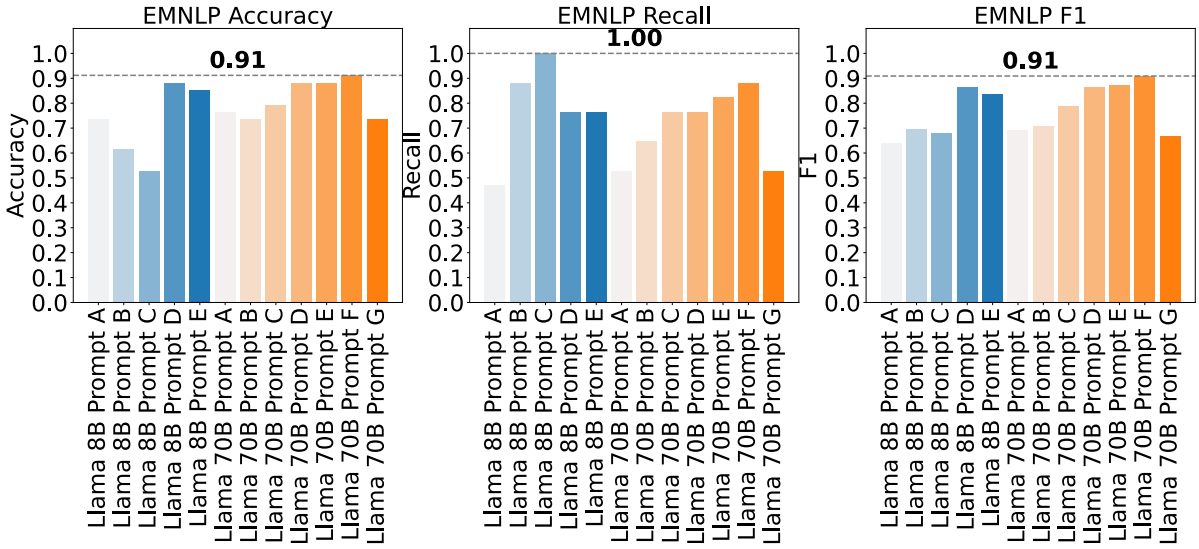

Figure 6: Performance of different models-prompts for EMNLP 2023 on different metrics (left to right: Accuracy, Recall, F1 score). The naming pattern of x-ticks is "Model X Prompt Y". **Accuracy** is particularly relevant, as the goal is to get as many correct predictions as possible.

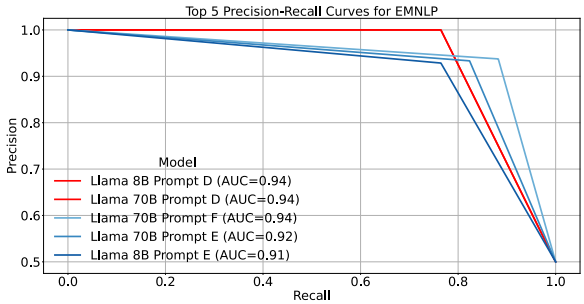

Figure 7: Top 5 Precision-Recall Curves for EMNLP 2023. The naming pattern of the legends is "Model X Prompt Y". The curves quantify the overall ability of the model-prompt combinations to discriminate between whether a review recommends extra citations. The larger the **AUC** the better. Best model-prompt combinations are in red.

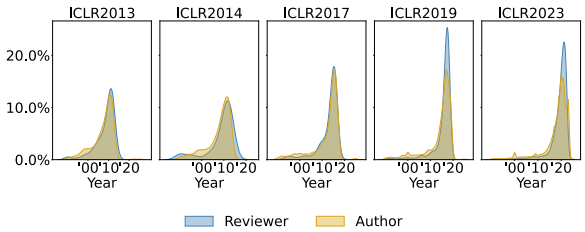

Figure 8: Year distribution of author's reference list and reviewer-recommended citations after 1990. The plot shows the increasing pattern of reviewer recommending more recent citations based on ICLR 2013 to 2023 (§4.3)

C Add-On Experiments

Motivation To look deeper into the dynamics between recency bias and peer review process in NLP and ML research, we conducted an add-on analysis to see whether reviewers reward papers that have a "newer" reference list. If peer review outcomes are biased toward recency, this could be an incentive for authors to cite more newer papers, potentially reinforcing short-term trends.

Methodology To assess the relationship between citation recency and review outcomes, we first quantified the "recency" of a paper's references. For each submission, we extracted the publication years of all cited papers from the reference list in the PDF, we then computed the average citation age. This was then compared against peer review outcomes, including whether the paper was accepted or rejected, and numerical reviewer scores when available (on different scales depending on categories). Note that only ICLR 2023 has relatively balanced accepted (1,574) and rejected (2,222) data available, which makes the overall results representable. However, other venues or subsets of venues were also examined to give a more comprehensive overview. The numbers of papers in each subsets is stated in Table 6. The reason why EMNLP 2023 only studied as a whole but NeurIPS 2023 & 2024 separately is because the former only has 9 rejected papers available, thus unnecessary to examine it separately; while both NeurIPS 2023 and 2024 have reasonable amount yet small proportion of rejected papers, it makes more sense to treat

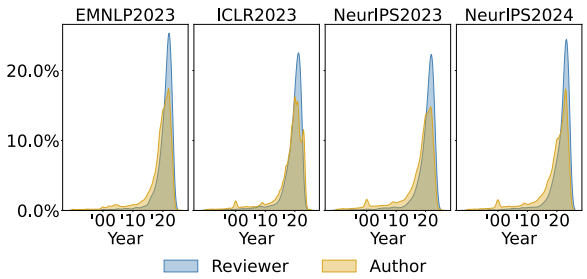

Figure 9: Year distribution of author's reference list and reviewer-recommended citations after 1990. The peaks indicate concentration of data points. The y-axis indicates the percentage of papers from particular years among all the cited/recommended citations of the studied venue. (§4.3)

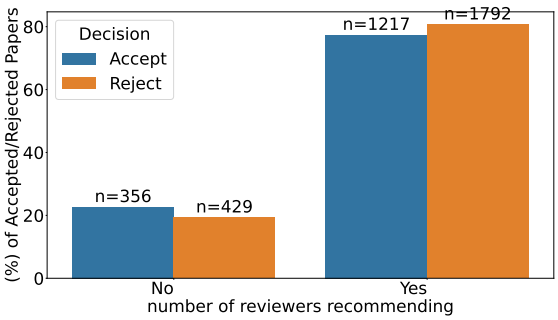

Figure 10: Correlation between paper decision and getting citation recommendation from reviewers. (§4.5

accepted and rejected data separately.

Venue (Subsets)	Number of Papers
ICLR 2023 (Total)	3,796
ICLR 2023 (Accepted)	1,574
ICLR 2023 (Rejected)	2,222
EMNLP 2023 (Total)	2,020
NeurIPS 2023 (Accepted)	3,218
NeurIPS 2023 (Rejected)	177
NeurIPS 2024 (Accepted)	4,036
NeurIPS 2024 (Rejected)	202

Table 6: Venue subsets studied for the add-on experiment, Appendix §C

Results Figure 12 show the average categorical and final review scores for ICLR 2023 papers, grouped by the average publication years of the reference list. To determine whether the distribution of review scores differs significantly among papers with different average citing years, we again conducted Chi-square tests. For average technical novelty, the Chi-square statistic was 51.28 with a p-value of 0.0221. This result indicates a statistically significant difference at the 5% level. Similarly, the average empirical novelty produced a

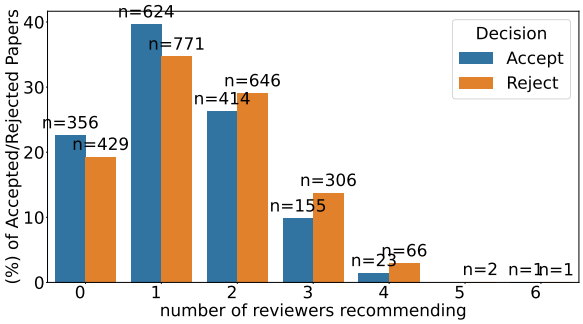

Figure 11: Correlation between paper decision and the number of citation recommendation (§4.5)

Chi-square statistic of 55.26 (p = 0.0089), which is significant at the 1% level. This implies even stronger evidence of non-random distribution in scores. For the final scores as shown in Figure 12, the trend is clearer. Namely,**when papers cite more recent literature, they are more likely to be assigned higher final scores by the reviewers in ICLR 2023**. The Chi-square test was again conducted, with Chi-square statistic being 142.221 (p = < 0.0001), demonstrating extreme significance at the 0.01% level.

However, this result does not automatically mean citing more recent papers leads to paper acceptance, as the ICLR 2023 (Total) includes both accepted and rejected papers. To make the analysis more rigorous, we also conducted the same analysis to both ICLR 2023 (Accepted) and ICLR 2023 (Rejected) separately. It reveals that only the scores of rejected papers correlate with their average citing years (See Table 7). In other words, **in ICLR 2023, if the quality of a paper is leaning towards inadequacy, reviewers tend to give it higher scores if it cites more recent papers**. On the other hand, a high-quality paper usually is not particularly awarded for citing more newer literature.

Interestingly, reviewers of NeurIPS 2023 and 2024 show different behavior compared to ICLR 2023. To take NeurIPS 2024 as an example (See Figure 13), the more "recent" a reference section is, the lower scores its paper received. If looking at individual categorical scores, papers that cited older papers received higher soundness scores. The significance of the above two observations are backed by Chi-square test as shown in Table 8. The trend for higher soundness score and final scores for papers that have higher citation ages is also significant in NeurIPS 2023 (Accepted). It is therefore safe to say, **reviewers often view papers that cited more older papers as more convincing and well-**

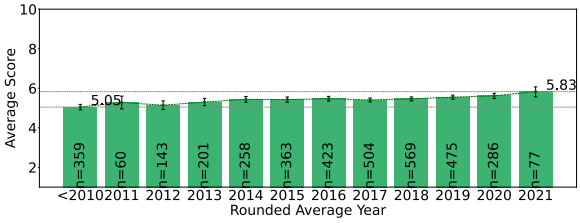

Figure 12: Total reviewer scores for different avg. citing years in ICLR 2023 (Total). Average Citing Year = Sum of years of all cited works in a paper / Number of citations in that paper. The result is rounded to the nearest integer. Consecutive years with few data points are aggregated (e.g., "<2010"). (Appendix §C)

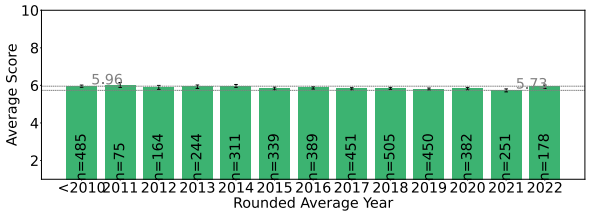

Figure 13: Total reviewer scores for different avg. citing years in NeurIPS 2024 (Total). Average Citing Year = Sum of years of all cited works in a paper / Number of citations in that paper. The result is rounded to the nearest integer. Consecutive years with few data points are aggregated (e.g., "<2010"). (Appendix §C)

established.

The different correlation between the citation age and review scores between ICLR and NeurIPS could therefore be due to conference-dependent reasons, or the lack of rejected papers of NeurIPS. One possible speculation is the grading system of ICLR 2023 focus more on novelty compared to that of NeurIPS 2023 & 2024 by only having novelty-oriented criteria in the meta-review. However, further analysis in the future is needed to confirm this assumption.

Venue (Subsets)	χ^2 (Tech)	p (Tech)	χ^2 (Emp)	p (Emp)	χ^2 (Final)	p (Final)
ICLR 2023 (Total)	37.530	**0.0207**	47.421	**0.0013**	142.221	**< 0.0001**
ICLR 2023 (Accepted)	30.991	0.0963	27.180	0.2045	58.209	0.3581
ICLR 2023 (Rejected)	32.728	0.0658	38.281	**0.0170**	106.422	**0.0012**

Tech = Technical novelty, Emp = Empirical novelty, Final = Final scores, χ^2 = Chi-square statistic, p = p-value.

Table 7: Chi-square test for different ICLR 2023 subsets. (Appendix §C)

Venue (Subsets)	χ^2 (Snd.)	p (Snd.)	χ^2 (Exc.)	p (Exc.)	χ^2 (Re.)	p (Re.)	χ^2 (Final)	p (Final)
EMNLP 2023 (Total)	27.747	0.8361	39.209	0.3280	36.615	0.4401	36.011	0.0548
Venue (Subsets)	**χ^2 (Con.)**	**p (Con.)**	**χ^2 (Pre.)**	**p (Pre.)**	**χ^2 (Snd.)**	**p (Snd.)**	**χ^2 (Final)**	**p (Final)**
NeurIPS 2023 (Accepted)	30.225	0.1774	22.105	0.5730	42.531	**< 0.0001**	79.967	**0.0026**
NeurIPS 2023 (Rejected)	39.212	**0.0133**	34.612	**0.0425**	18.550	0.6729	42.296	0.5449
NeurIPS 2024 (Accepted)	27.158	0.2971	14.425	0.9365	57.613	**0.0001**	97.364	**< 0.0001**
NeurIPS 2024 (Rejected)	26.262	0.3400	32.738	0.1097	19.968	0.6986	42.075	0.7132

Snd. = Soundness, Exc. = Excitement, Re. = Reproducibility, Con. = Contribution, Pre. = Presentation, Final = Final Score, χ^2 = Chi-square statistic, p = p-value.

Table 8: Chi-square test for different EMNLP and NeurIPS subsets (Appendix §C)