Master's Thesis

Quantifying Biases in Peer Review: Analyzing Reviewer Suggestions in Artificial Intelligence Publications

Master's Thesis for Joint Honours MA in Digital Humanities
with Applied Computer Science
Chair for Scientific Information Analytics
University of Göttingen

Submission date: June 30th, 2025

By: Zhuojing Huang
From: Göttingen
Matriculation number: 12305011

First examiner: Prof. Dr. Bela Gipp
Second examiner: Dr. Terry Lima Ruas
Supervisor: Jan Philip Wahle

University of Göttingen
Chair for Scientific Information Analytics

Contents

I hereby declare that I have written this thesis independently without any help from others and without the use of documents or aids other than those stated. I have mentioned all used sources and cited them correctly according to established academic citation rules.

Göttingen, June 30th, 2025

Quantifying Biases in Peer Review: Analyzing Reviewer Suggestions in Artificial Intelligence Publications

ZHUOJING HUANG, University of Göttingen, Germany

Abstract

The peer review process is central to academic publishing, yet its influence on the citation behaviour of authors remains underexplored. Drawing on large-scale review data and submitted papers of top-tier AI conferences from OpenReview, this study conducts a multifaceted analysis of reviewer-author interactions around citation practices. It investigates the role of citation recommendations made by reviewers, and analyses the characteristics of these suggested papers as well as how these impact the final bibliography of accepted and rejected papers. Using statistical methods, traditional NLP tools and LLMs, this research quantifies the frequency, recency and topical distribution of citation suggestions over the years, as well as the rate at which the authors incorporated such recommended works. It also examines whether citation recommendation behaviour correlates with paper decision and investigates the influence of citation recency on review scores. Key findings suggest that peer reviewers could be a possible source of recency biases in AI research field, as a majority of authors end up incorporating recommended papers from reviewers which are much newer compared to those they cited in their work. Another influence of such recommendations is the increasing insularity of AI-related papers due to an increasing proportion of papers from Computer Science being recommended. Additionally, the analysis shows statistically significant differences in citation recommendation patterns between accepted and rejected papers. This thesis contributes to a deeper understanding of the dynamics within peer review and citation practices. It gives hints for improving transparency and accountability in academic publishing and offers new directions for research into reviewer influence on published papers. [1]

1 Introduction

As one of the fastest-growing scientific fields, Artificial Intelligence (AI) is characterized by fast publication cycles and a strong emphasis on conference proceedings. Top-tier conferences not only push the frontiers of the field but also potentially shape the direction of future research. Scientific research is inherently interconnected, with no single domain existing in isolation. However, the modes of influence are numerous and complex, but one notable marker of scientific influence is citations [50]. Using a subfield of AI – Natural Language Processing (NLP) – as example, Wahle et al. [50] note that the field of NLP has a rise in intra-field references and a decrease in citation age. Despite the widespread impact of NLP technologies, especially Large Language Models (LLMs), there is insufficient engagement with literature outside computer science, particularly from fields like psychology, social science, and linguistics [50]. This lack of interdisciplinary integration is especially concerning given the risks posed to, e.g., marginalized communities, and it can also hinder the researchers from taking inspirations from other scientific domains.

The reasons behind the recency and insularity bias of NLP or broader AI publications are various. One such source could be peer reviews (See Figure 1 for an simplified example). In the context of scientific publication, peer review plays a critical role in deciding which papers can be accepted by the venues. It posts influence on the research fields not only by its "gate-keeping" nature, but also through suggesting authors to adopt certain methodologies or to refer to a particular theory. In other words, peer review can, for example, promote technical contribution over methodological ones, recent papers over older ones, specific applications over other, and trendy topics over foundational ones (e.g., LLMs). Among the various forms of feedback reviewers provide, one influential yet under-examined kind is the suggestion to cite additional literature. These citation recommendations have the potential to subtly influence the trajectory of future research, reinforce dominant sub-areas, or influence research visibility. Despite their significance, little systematic research has been done. To fill the gap, this thesis undertakes

[1] All the datasets and the code to reproduce the experiments are available on GitHub

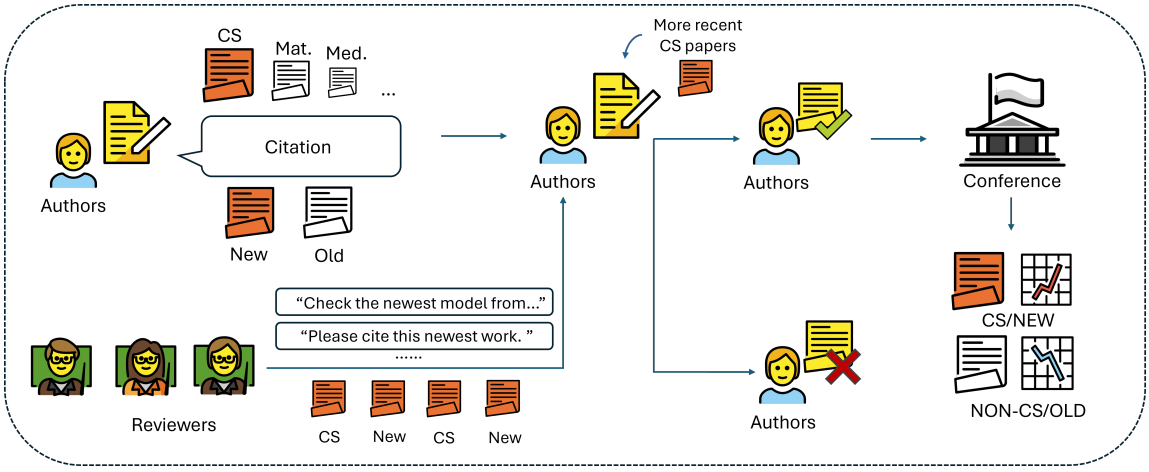


Fig. 1. A simplified flow illustrating the incorporation of insularity and recency biases into the AI reasearch landscape via the peer review suggestions.

a large-scale quantitative analysis of citation recommendations in peer review, with a particular focus on major AI conferences. This study aims to quantify how often such recommendations occur, the characteristics of the recommended papers, authors' reactions towards such recommendation, and finally, the reviewers' evaluation towards adopting newer literature.

To acquire the data for the thesis, I compiled new datasets with peer reviews and corresponding submitted papers with comprehensive metadata through OpenReview API. The dataset consists of ~59,1k reviews and ~14k papers of major AI venues ranging from 2013 to 2024. By analysing the datasets, my thesis follows the following three guidelines to explore how potential biases could be adopted through literature recommendations in the process of peer review:

1. Quantitative Analysis of Citation Recommendations. This component investigates the frequency, patterns, and distribution of citation suggestions made by reviewers. It examines how often reviewers recommend additional citations, the types of works being cited (e.g., recent papers, papers from specific fields), and whether there are observable trends over time. The goal is to identify whether papers with certain characteristics are disproportionately recommended, suggesting potential systemic bias or preferential citation behaviour.

2. Analysis of Reviewers' Suggestions and Their Influence on Research Focus. This analysis evaluates the thematic alignment between reviewers' citation suggestions and the original focus of the submitted manuscripts. It examines whether reviewers tend to influence authors toward specific research areas, methodologies, or frameworks. Additionally, this section explores whether the frequency or nature of citation suggestions correlates with paper outcomes. Specifically, whether rejected papers tend to receive more citation recommendations than accepted ones. This could suggest that reviewers give more citation suggestions to papers they think are lower quality or not a good fit for the conference.

3. Authors' Responses to Reviewers' Citation Suggestions. This section studies how authors respond to citation recommendations in the final, camera-ready versions of papers. It quantifies the extent to which suggested citations are adopted. Another point that this section investigates is, whether papers citing more recent works receive better peer review outcomes, i.e., higher review scores, compared to those that cite older, more fundamental papers.

The contribution of this thesis mainly consists of three aspects: (1) introduce a novel methodology combining heuristic rules and Large Language Models (LLMs) to automatically detect and quantify citation recommendations in peer reviews, (2) compile a new, large-scale dataset of ~59.1K reviews and ~14K AI conference papers ranging from 2013 to 2024 via the OpenReview API, and (3) provide a comprehensive analysis of the frequency, patterns, and impact of citation suggestions, including their potential role in reinforcing recency bias, field insularity, and shaping research directions in the field of AI.

The thesis reveals that citation recommendations are a common part of peer reviews, occurring in over 25% of reviews in most of the venues, with more than 50% of all the submitted papers receiving at least one paper recommendation. These suggestions disproportionately favour recent papers, typically from the past one to two years of the studied venue. Besides, the reviewers recommend more and more works from the field of Computer Science (CS) in the past decade, with the tendency to ignore under-represented yet essential fields as regard to AI development. There is also a notable correlation between rejection and the number of citation suggestions, indicating that reviewers may use these suggestions as a means of signaling deficiencies or redirecting a paper's focus. In terms of author response, more than 60% of the suggested citations are incorporated into the camera-ready versions of accepted papers.

These findings suggest that citation suggestions in peer reviews serve not only as technical recommendations but also as subtle ways of influence. The tendency to promote recent and intra-domain citations reinforces the insularity and recency biases already observed in the field, potentially narrowing the diversity of AI research. Ultimately, peer review appears to play a more active role in shaping scientific discourse than previously assumed.

2 Related Work

Analysis of Citation Patterns. The study of citation practices can date back to the mid-20th century [12]. Researchers have studied citation patterns from different angles, such as geographic location of authors [37], institutional affiliation [1], researcher reputation [10], and demographic factors [1, 5, 9, 26, 30]. Ismail et al. [20] through citation analysis and network visualization, present a domain-specific bibliometric analysis on identity studies, demonstrating how global trends and scholarly impacts can be mapped over a decade. Other dimensions of analysis include paper length [14], perceived quality [7], academic discipline [11], language of publication [25], publication venue [52], self-citation practices [38], instances of plagiarism [17, 49], and institutional diversity [1].

As a fast-growing field, AI has attracted great attention in citation studies with several open-access datasets on citation patterns [29, 31, 50, 51]. Wahle et al. [51] highlighted increasing recency and insularity bias in NLP citations, raising concerns about a lack of diversity. Gnewuch [18] quantifies citational influence from industry on research trajectories. Recent research has also examined how AI tools themselves, especially LLMs, influence citation practices. For example, Algaba et al. [4] found that LLMs not only replicate human citation patterns but also exhibit heightened bias towards highly cited papers, potentially reinforcing the Matthew effect. Building on this, Algaba et al. [3] showed that LLM-generated references tend to favour recent publications with shorter titles and fewer authors, indicating that such models may reshape citation dynamics across scientific domains.

While these studies provide deep insights into the factors influencing citation behaviour, they often focus on authors' choices post-publication. Thus, current research usually overlook how peer review may influence these decisions.

Peer Review in Scientific Research. Similar to citation, peer review has been a longstanding focus of academic research. Early studies in peer review process examines its reliability and objectivity. Mahoney [28] conducted an experimental study highlighting confirmatory bias in peer review, demonstrating that reviewers tend to favour manuscripts aligning with their own theoretical perspectives. Later research has explored various dimensions of peer review, including geographic distribution of reviewers [41, 47], institutional affiliation [23], and the influence of author reputation and seniority [6, 32]. Other studies have investigated aspects like the

length and tone of review reports [53], disciplinary norms [34], and language proficiency of authors [15]. The impact of publication venues on peer review practices has also been examined by Squazzoni et al. [42]. With the recent rise of generative AI, the interaction between peer review process and LLMs are also being explored. Jin et al. [21] introduces a simulation framework using LLM agents to model peer review dynamics, revealing significant variations in paper decisions due to reviewer biases. Ebadi et al. [13] underscores the need for clear guidelines and policies, as well as their proper dissemination among researchers, to address the ethical concerns and practical challenges raised by using LLMs in peer reviews.

Researchers emphasize the need for ongoing evaluation and improvement of the peer review system with the concerns about the reliability and bias in peer-reviews [19, 40], thus efforts to make it more fair and inclusive have also been made. Ross-Hellauer [36] provided an overview of open peer review, discussing its potential to bring transparency and accountability to the peer review process. Kern-Goldberger et al. [22] investigate how double-blind reviewing correlates with increases in women as first authors, suggesting policy shifts to promote diversity. In a recent study, Lu et al. [27] introduce a data-driven methodology for identifying and categorizing aspects within peer reviews. Their work shows the potential of aspect-based analysis in improving the consistency and quality of the peer review process, including detecting automated review generation. Recent efforts have begun to examine the cognitive shortcuts that may degrade the quality of peer reviews under time pressure. Purkayastha et al. [35] propose LAZYREVIEW dataset, which consists of peer-review sentences with vague praise, unsupported criticism, or irrelevant commentary. Their work highlights the challenges LLMs face in detecting these issues in a zero-shot setting, but also demonstrates that fine-tuning on LAZYREVIEW with instructions significantly improves performance, leading to more comprehensive and actionable reviews.

Despite considerable attention to characteristics, biases and transparency in peer review, little empirical work has examined how reviewers' suggestions, particularly citation recommendations, may shape the trajectory of accepted papers or subtly inject systemic biases in citation norms.

Citations Within Peer Review Contexts. The intersection of peer review and citation practices particularly addresses potential biases and ethical concerns. Fong and Wilhite [16] showed widespread coerced or superfluous citations, driven mostly by pressures for publication and funding, with variation across disciplines, ranks, and demographics. Levis and Leentjens [24] investigated self-citation practices among peer reviewers and found that reviewers often request citations to their own work, which is also coercive in some cases [45], raising questions about the objectivity and fairness of the review results. Stelmakh et al. [44] provided empirical evidence for citation bias in peer review by examining whether reviewers are more likely to recommend papers that cite their own work. A study of peer review in major computer science conferences also revealed evidence of citation bias, showing that citing a reviewer's work can significantly increase review scores even after controlling for paper quality and reviewer expertise [43]. These findings suggest a significant correlation, reinforcing concerns about self-interest in the review process. In a more constructive approach, Zong and Xie [54] evaluated whether open peer review improves citation outcomes for authors. Their results suggest that increased transparency may reduce certain forms of bias, although the effect was modest. Altogether, these studies show the complex dynamics between peer review, citation practices, and academic incentives.

While some studies have explored the influence of peer review given to citation practices, most of them have focused on instances where reviewers promote their own work, leaving a gap in understanding the broader patterns of reviewer-suggested citations.

Research Gap and Scope of This Thesis. Building upon the existing literature on citation practices, peer review processes, and their intersection, this thesis investigates potential biases emerged from reviewer-suggested citations within AI research. With self-citation being the most studied citation bias emerged from peer reviews, other citation bias formed through similar manner remains unexplored. The trend of increasing recency and insularity biases in AI citation patterns [51] could, however, be influenced by peer review citation suggestions. Focusing on peer review data from top-tier AI venues ranging from 2013 to 2024, the thesis aims to quantify

how often and under what conditions reviewers recommend citations to the submitted papers. It also looks into whether the pattern of recommended literature has changed over time. Through a series of structured research questions, the thesis offers an empirical analysis of reviewer citation suggestions and contributes to understanding how such suggestions may reinforce certain biases within the rapidly evolving field of AI.

3 Methodology

3.1 Data Source and Scope

To obtain peer review data and paper submission information from AI conferences, I used OpenReview.net. OpenReview is a platform designed to promote transparency and openness in scientific communication, particularly in the peer review process. It supports open access to papers and reviews, as well as ongoing discussion [33]. Crucially, OpenReview provides a REST API that facilitates structured access to submission records, review assignments, comments, and reviewer metadata. The API also comes with detailed documentation, which enables easy data collection at scale.

This study analyses peer reviews and citation patterns across multiple high-profile venues in AI research field, including:

- EMNLP – Conference on Empirical Methods in Natural Language Processing
- ICLR – International Conference on Learning Representations
- NeurIPS – Conference on Neural Information Processing Systems

In particular, the main venues of the thesis includes EMNLP 2023, ICLR and 2023, and NeurIPS 2023 and 2024 (see Table 1 for stats overview). The choice was made due to the prominence of the conference, volume of paper submissions, and accessibility of structured peer review and submission data. Note that there is a discrepancy between total submissions and available papers, as well as between the number of accepted papers and rejected papers in EMNLP 2023, NeurIPS 2023 and NeurIPS 2024. This is possibly due to authors' withdrawal of their rejected papers or not consenting to public access at OpenReview. In addition, thanks to the availability of older data, though relatively limited, ICLR also provides historical submission and review tracking, including ICLR 2013, 2014, 2017 and 2019 (See Table 2). These data are used in some of the later research questions in the thesis.

In total, the collected dataset comprises 59,389 reviews with approximately 28.53 million tokens. Across all the examined venues, there are 24,855 total submissions, with 18,655 papers available via OpenReview API. The dataset includes detailed metadata for each paper and review. This collection supports comprehensive and representative analyses of trends in reviewer suggested literature and changes in reviewer behaviour across time and venues.

Venue (Year)	Reviews	Token Counts	Total Submissions	Available Papers	Accepted	Rejected
EMNLP 2023	6,449	~2.59M	4,909	2,020	2,011	9
ICLR 2023	14,351	~7.06M	4,874	3,796	1,574	2,222
NeurIPS 2023	15,175	~7.83M	12,345	3,395	3,218	177
NeurIPS 2024	16,650	~8.18M	15,671	4,238	4,036	202

Table 1. Main venues in the thesis: statistics for collected reviews and papers across venues.

3.2 Data Collection and Processing

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).

Venue (Year)	Reviews	Token Counts	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 2. Historical data of ICLR: statistics for collected reviews and papers across venues.

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. The submission data and review data are separated by default. However, scripts to merge them are included in the GitHub together with other code to process and analyse the data. [2]

Table 3 is an example of EMNLP 2023 review data. Each review entry has not only free-text fields like the paper's main contributions, reasons to accept or reject, and questions for the authors, but also a variety of standardized rating fields. These include numerical scores for soundness, excitement, reproducibility, and reviewer confidence, typically on a 1–5 scale. Metadata of EMNLP 2023 for each review also includes ethical concerns, justification if any, missing references if applicable, writing and presentation comments, and timestamps for review creation, decision, and last modification. Each review is linked to specific paper submissions through unique IDs and includes OpenReview-specific fields such as reviewer anonymity status, reader visibility, signature, and the invitation used to post the review. Licensing information (e.g., CC BY 4.0) and tracking metadata (like forum/thread IDs and domain) are also included in the data acquired through OpenReview API.

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

(1) Register on OpenReview (required for API v2.0)
(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. I 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.[3]

4 Experiments

4.1 Research Questions

The objectives of this thesis are organized into three sequential parts, each corresponding to a different stage in the citation-feedback cycle. Part A establishes descriptive baselines for reviewer citation recommendations, Part B investigates how these recommendations influence research focus and review outcomes, and finally Part C evaluates authors' responses and the ultimate impact of integrating the potential bias into the paper.

[2] All code available on GitHub
[3] OpenReview API documentation

Field	Content
Paper Topic and Main Contributions	The authors propose an induction-augmented framework that utilizes inductive knowledge derived from LLMs and retrieved documents for better implicit reasoning. They enhance RAG with an inductor module, and propose IAG-GPT and IAG-Student models. Experiments show strong performance on CSQA2.0 and StrategyQA.
Reasons to Accept	Addresses a non-trivial problem with a novel approach. Experiments are extensive and results are reasonable.
Reasons to Reject	Concerns about generalization: inductive info added even when not needed; rigid prompt structure; large gap between IAG-GPT and IAG-Student performance. More experiments on larger models recommended.
Questions for the Authors	1. Would jointly fine-tuning the generator and inductor help? 2. How are μ and σ computed (line 228)?
Soundness	4: Strong
Excitement	4: Strong
Reproducibility	4: Could mostly reproduce
Ethical Concerns	No
Reviewer Confidence	4: Quite sure
Justification for Ethical Concerns	NULL
Missing References	NULL
Typos, Grammar, Style, and Presentation Improvements	NULL
Creation Date (cdate)	1691040000000
Decision Date (ddate)	NULL
Details	NULL
Domain	EMNLP/2023/Conference
Forum	zwqDROxClj
Invitations	[EMNLP/2023/Conference/Submission1595/-/Official_Review, EMNLP/2023/Conference/-/Edit]
License	CC BY 4.0
Modified Date (mdate)	1701460000000
Nonreaders	[]
Number	1
Official Date (odate)	NULL
Published Date (pdate)	NULL
Readers	[everyone, EMNLP/2023/Conference/Submission1595/Reviewer_HrRJ]
ReplyTo	zwqDROxClj
Signatures	[EMNLP/2023/Conference/Submission1595/Reviewer_HrRJ]
TCDate	1691040000000
TMDate	1701460000000
Writers	[EMNLP/2023/Conference, EMNLP/2023/Conference/Submission1595/Reviewer_HrRJ]

Table 3. An example of review entry from EMNLP 2023

Part A. Quantitative Analysis of Citation Recommendations. I measured how often reviewers suggest additional literature, characterizing those suggestions by both the age of references as well as their main disciplines, and comparing them to the paper's original citations. Specifically, I ask:

A1. How often do reviewers recommend additional literature as part of their review? (QA1)
A2. How does the age of the recommended references compare to the distribution of ages of references that the paper had cited? (QA2)

A3. **(a)** How often do the recommended references fall into the field of Computer Science (CS) as opposed to other fields?
(b) Is the distribution of fields of study for recommended references different from that of the original references in submissions? (QA3)

Part B. Analysis of Reviewers' Suggestions and Their Influence on Research Focus. Building on the quantitative statistics in Part A, I examined the topical patterns in reviewer recommendations and also test the correlations between citation recommendation and reviewers' final decisions. The questions are:

B1. In which topic areas are papers recommended by peer reviewers, and is there a bias toward specific topics? (QB1)
B2. Does paper decision correlate with citation recommendation frequency; that is, do rejected papers receive more citation suggestions than accepted papers? (QB2)

Part C. Authors' Responses to Reviewers' Citation Suggestions. Finally, I assessed how authors incorporate reviewer-suggested citations and whether citations in the camera-ready version reflect reviewer recommendations. Also, I evaluated if incorporating more recency into the citation results in higher review scores. The questions are:

C1. How often do authors incorporate the citing suggestions from the reviewers? (QC1)
C2. Do papers that heavily cite recent work receive better peer review outcomes (e.g., being accepted or having higher reviewer scores) than those with a more balanced or older reference profile? (QC2)

4.2 Main Experiments

4.2.1 *(QA1) How often do reviewers recommend additional literature as part of their review?*

Motivation. This research question gives a general overview of how often reviewers suggest citing additional literature, thereby setting a foundation for the later analyses. Confirming that citation recommendations are indeed a common phenomenon ensures that the findings of the thesis are representative.

Method. To determine whether a certain reviewer recommended the authors to refer to extra literature is not trivial. It is a tricky task because there could be various ways of suggesting a certain paper, dataset, benchmark, etc. For example, one could suggest extra literature by saying "please check Joe Doe's newest paper", or "the newest benchmark on this task is available". Therefore, a keyword- or regex-based heuristics is prone for errors. For higher accuracy and a better foundation for later experiments, I used an LLM-based classification. I tested both `LLaMA 3.1 8B` and `LLaMA 3.1 70B` [48] via LM Studio with different prompts on EMNLP 2023, ICLR 2023, NeurIPS 2023 and NeurIPS 2024. I tested five prompts on the `8B` model and seven prompts on the `70B` model, saving each model's outputs in separate CSV files. The prompts can be found in Appendix A.

To choose the optimal model-prompt combination for each venue, I manually annotated a balanced gold-standard sample of 100 reviews (50 positive, 50 negative) evenly drawn from EMNLP, ICLR, and NeurIPS. After removing manually annotated tags, I ran each candidate model-prompt combination on this sample and computed standard binary classification metrics, including **Accuracy**, **Recall**, **F_1 score**, and Area Under the ROC Curve (**AUC**).

The **Precision-Recall (PR) curve** plots Precision against Recall at various classification thresholds.

The **PR AUC** quantifies the overall trade-off between precision and recall, and is defined as:

$$\text{PR AUC} = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}) \tag{1}$$

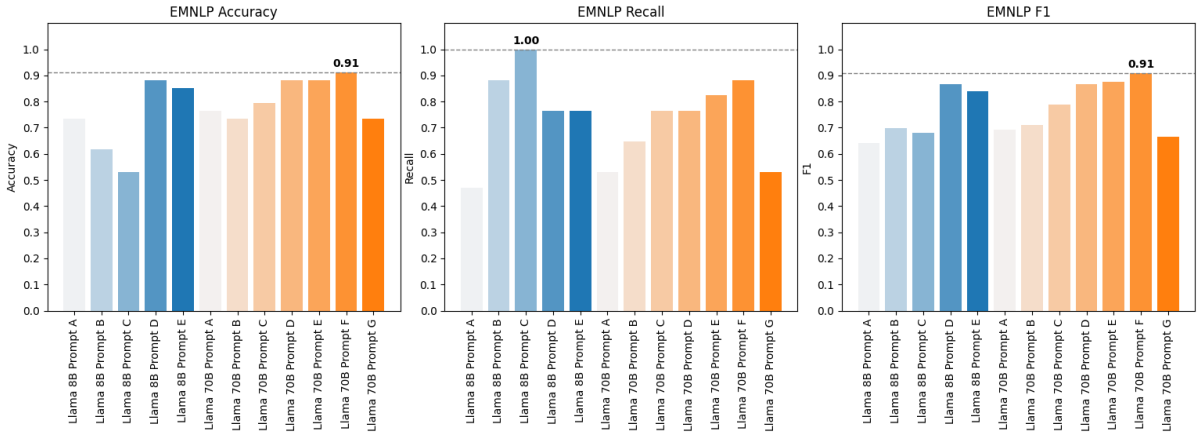


Fig. 2. Performance of different models-prompts for EMNLP 2023 on different metrics (left to right: Accuracy, Recall, F1 score). The naming pattern of x-ticks is "Model X Prompt Y". **Accuracy** is particularly relevant, as the goal is to get as many correct predictions as possible. (QA1, 4.2.1)

In this context, 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 2 and Figure 3 on EMNLP 2023, `Llama 70B prompt F` achieved AUC = 0.94 and the highest Accuracy of 0.91, so I used that model-prompt combination to classify all reviews in EMNLP 2023.

Results & Discussion. The proportions of reviews and papers containing at least one citation recommendation are presented in the following table:

For most of the venues, between 21% to 37% of the peer reviews contain citation suggestions, which confirms that reviewers suggesting extra literature is both frequent and universal in AI conferences. While the percentage of individual reviews containing citation recommendations varies from 21.44% (NeurIPS 2024) to 36.99% (ICLR 2023), the percentage of papers receiving at least one such recommendation is notably higher, ranging from 58.59% to 79.32%. This discrepancy suggests that even if individual reviewers do not frequently recommend citations, most papers still receive at least one such suggestion from among their set of reviewers.

Although older data from ICLR like ICLR 2013 and ICLR 2014 have only less than 1,000 reviews in total, thus too little to be representative for this research question, they show similar pattern as their more recent counterparts. Another interesting tendency is that along the years, reviewers of ICLR have become more dedicated in recommending extra literature to the authors. Although instead of ICLR 2013 and 2014, ICLR 2017 has the lowest reviews with recommendations and papers with recommendations percentage among all the available ICLR data, the growing trend is clear (See Figure 4). Besides, as stated earlier, ICLR 2013 and ICLR 2014 have too little data available, which could introduce a bigger error range. That is to say, citation recommendation is indeed a universal and frequent phenomenon in AI research field, which could also have a growing trend based on the analysis on historical data of ICLR.

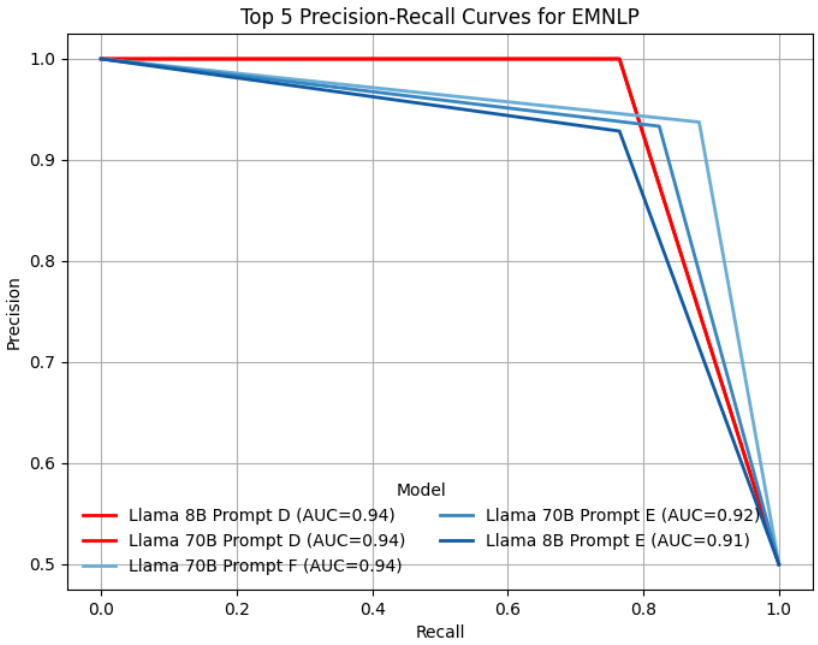


Fig. 3. Top 5 Precision-Recall Curves for EMNLP 2023. The naming pattern of the legends is "Model X Prompt Y". The curves quantify the overall ability of the model-prompt combinations to discriminate between whether a review recommends extra citations. The larger the **AUC** the better. Best model-prompt combinations are in red. (QA1, 4.2.1)

Venue (Year)	Reviews	Review with Rec.	Papers	Paper with Rec.
EMNLP 2023	6,449	31.48%	2,020	67.48%
ICLR 2023	14,351	36.99%	3,796	79.32%
NeurIPS 2023	15,175	22.09%	3,395	64.27%
NeurIPS 2024	16,650	21.44%	4,238	58.59%
Historical ICLR Data				
ICLR 2013	373	24.93%	67	67.16%
ICLR 2014	548	16.97%	69	63.77%
ICLR 2017	1,511	23.56%	490	54.49%
ICLR 2019	4,332	33.86%	1,511	69.34%

Table 4. Overview of extra literature recommendation across venues. Review with recommendation accounts for the percentage of peer reviews that suggest additional papers. Paper with recommendation calculate the percentage of submitted papers that got recommended additional papers by at least one reviewer. (QA1, 4.2.1)

4.2.2 *(QA2) How does the age of the recommended references compare to the distribution of ages of references that the paper had cited?*

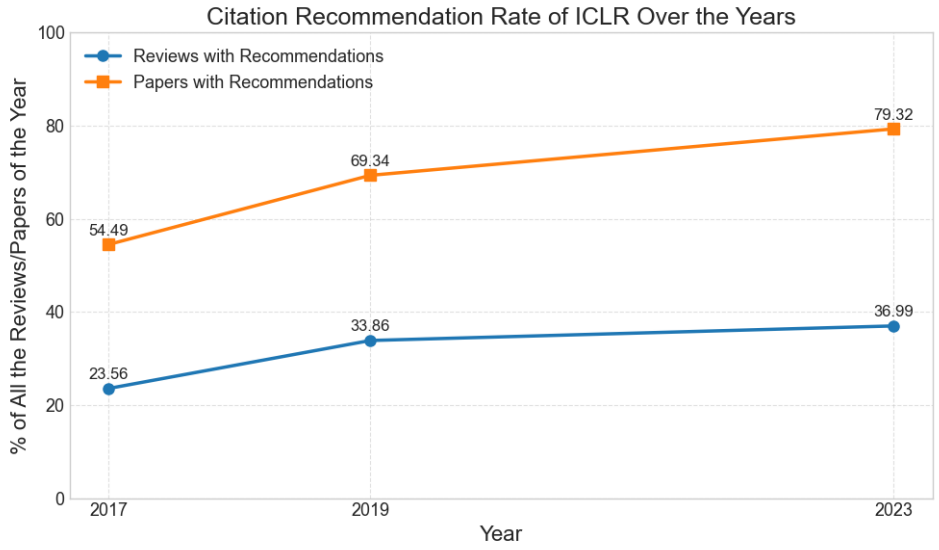


Fig. 4. Citation recommendation rate (%) of ICLR from 2017 to 2023. Although ICLR 2013 and 2014 are also available, the limited data might introduce bigger error range. (QA1, 4.2.1)

Motivation. Understanding the age distribution of reviewer-recommended references relative to those already cited by authors reveals whether peer review encourages citation of recent research, foundational literature, or a balanced mix. This is crucial for assessing whether the recency bias in AI-related publications may originate, in part, from peer review.

Method. For each citation suggestion identified in *QA1* (Section 4.2.1), I extracted the publication years and compared them to the average publication year of the references cited by the paper. Since for this particular research question, only the years are relevant, I ignored the other information at this stage and only looked for years with typical citing patterns, such as "(Author, 2022b)", "arXiv:2202...", ". 2022" or ". (2022)". Apart from comparing the difference between average citing age, I also calculated the medians and visualised the data distribution of the citing years to rule out the possibility of extremely unbalanced citing years.

Venue (Year)	Avg. Cited Age	Avg. Rec. Age	Med. Cited Age	Med. Rec. Age
EMNLP 2023	7.3	2.8	6.4	2
ICLR 2023	7.7	4.1	6.9	3
NeurIPS 2023	8.7	3.3	7.9	2
NeurIPS 2024	8.5	3.0	7.8	1

Table 5. Average cited vs. recommended ages across venues. To rule out the possibility of extremely imbalanced year distribution, medians are also calculated. The higher the age, the older the papers. (*Cited Age = Publication Year - Year of the Cited Paper*; *Recommended Age = Publication Year - Year of the Recommended Paper.*) (QA2, 4.2.2)

Results & Discussion. Across all venues, reviewer-recommended citations are, on average, much more recent than those already cited by the authors (See Table 5). This trend is consistent across EMNLP, ICLR, and NeurIPS, though its magnitude varies. Specifically, both the cited and recommended references in EMNLP papers skew newer compared to ICLR and NeurIPS. Besides, a direct comparison betwenn NeurIPS 2023 and NeurIPS2024 shows a age decrease in both cited and recommended papers, meaning both authors and reviewers prefer more newer papers in NeurIPS 2024-Detailed visualizations of these distributions are provided in Figure 5 and Figure 6. To quantify the magnitude and significance of the age difference, I computed both Cohen's d and Welch's t-test (relevant equations can be found in Equation 2 to 6):

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \tag{2}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{3}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{5}$$

$$p = P(T_{df} > |t|) \tag{6}$$

$\bar{X}_1, \bar{X}_2$ are the sample means of the two groups, i.e., cited age and recommended age

s_1, s_2 are the standard deviations,

n_1, n_2 are the sample sizes for each group.

The results indicate a highly significant difference between the average age of cited references (mean = 8.05 years) and reviewer-recommended references (mean = 3.3 years), with a **t-statistic = 10.87** and a **p-value < 0.001**. Moreover, the effect size is extremely large, with **Cohen's d = 7.69**. This confirms that reviewer suggestions are not only newer on average, but that the difference is both statistically and substantively large.

It is also noteworthy that in all venues analyzed, both the references cited by authors and those suggested by reviewers have a strong concentration within the three years leading up to the paper's submission, as shown in Figure 6. In addition, for ICLR and NeurIPS papers, there is a small citing peak around the year of 2000 as also shown in Figure 6, which is due to a set of foundation papers that are more prominent to the field of general machine learning, thus the peak does not exist in ENMLP papers for it mainly targets the field of NLP.

The findings suggest that peer reviewers contribute to the community's focus on recent work, reinforcing recency bias in citation behaviour. While emphasizing new research ensures awareness of the latest advancements, it may also lead to under-citation of older, yet still influential and foundational work. The particularly strong recency trend in reviewer suggested papers observed in EMNLP – indicated by low citation age of 2.8 years – implies that the NLP research community may be especially susceptible to these dynamics, which aligns with the findings by Wahle et al. [50].

4.2.3 *(QA3) a) How often do the recommended references fall into the field of CS as opposed to other fields? b) Is the distribution of fields of study different from that of papers cited in the submissions?*

Motivation. **(a)** Other than recency bias, prior work [50] also suggested that fields like NLP have grown increasingly insular in the past decades, often citing within the domain of CS. This research question therefore investigates whether peer reviewers contribute to that insularity by recommending more and more citations from within the CS communities over the years.

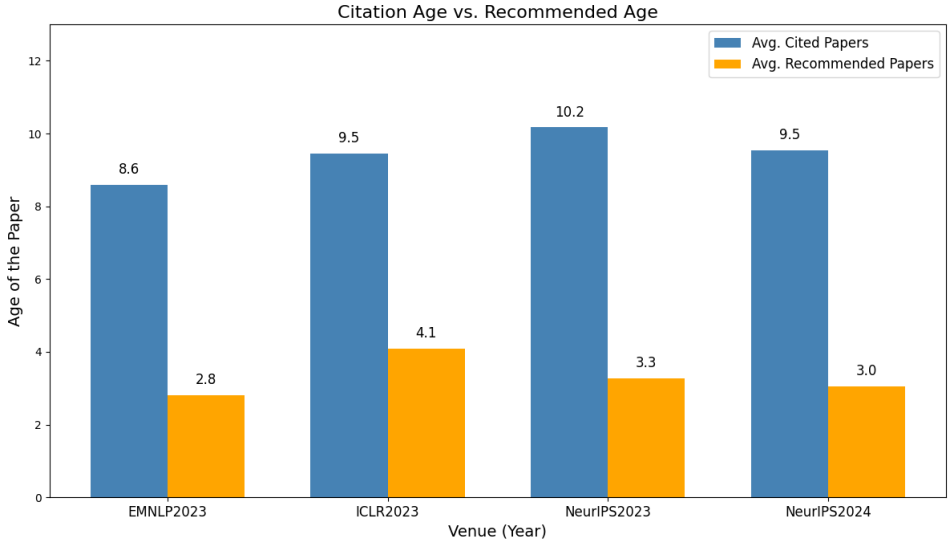


Fig. 5. Cited and recommended ages across venues. The higher, the age the older the papers. (*Cited Age = Publication Year - Year of the Cited Paper*; *Recommended Age = Publication Year - Year of the Recommended Paper.*) (QA2, 4.2.2)

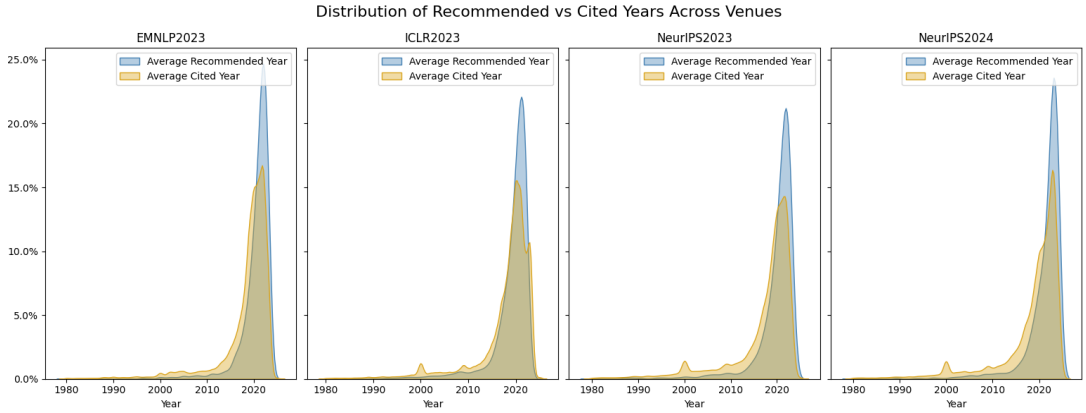


Fig. 6. Year distribution of cited and recommended papers after 1980. The peaks indicate concentration of data points. The y-axis indicates the percentage of papers from particular years among all the cited/recommended papers of the studied venue. (QA2, 4.2.2)

(b) It is equally important to determine whether the distribution of fields among reviewer-recommended literature differs from the distribution found in the original citations of the submitted papers. Such a comparison could show whether reviewer influence pushes papers toward narrower or simply different domains or subdomains.

Method. Looking at multiple years of data provides a meaningful understanding, as increasing proportion of cited or recommended CS papers emerge over time. The point of this research question is to evaluate if certain fields of study have witnessed change in proportion as regard to being recommended to authors or being cited by

authors. Hence, a conference that has reasonable amount of old data is required, which leaves ICLR the only choice in this study, as it also has data from the years of 2013, 2014, 2017 and 2019.

(a) To identify the field of recommended references, I first decided which reviews have recommended papers based on the results of *QA1* (4.2.1)). Next, I extracted only the paper information from the whole review text. A rule-based regex approach was initially attempted to extract recommended paper titles, but turned out to be insufficient due to the high inconsistency of citation formats in reviews. I then adopted an LLM-based information extraction approach using `llama-70B`, which is proven to be more effective than its `8B` counterpart in *QA1*, and venue-specific prompts to identify recommended paper titles, authors, and years. However, due to the diverse forms of recommending extra literature, ambiguous or unclear cases like “latest version of John Doe’s work”, or "Jane Doe’s paper on the same topic", are frequently encountered, making it extremely difficult or nearly impossible to use solely LLM to extract reliable information.

In order to make the analysis more reliable and not to introduce more potential biases from LLM, I decided to only sample parts of the dataset and manually annotate recommended literature. The sample size is calculated by Calculator.net [8] with the confidence level being 95% and margin of error 5%. Essentially, it uses the formula:

$$n_0 = \frac{Z^2 \cdot p \cdot (1-p)}{E^2} \tag{7}$$

where $Z = 1.96$ (corresponding to 95% confidence), $p = 0.5$ (assuming maximum variability), and $E = 0.05$ (representing a 5% margin of error) [46]. Based on the total number of recommended papers, I sampled corresponding amount of papers for each venue. In order to have a buffer zone for the next step, I annotated more than the amount according to Calculator.net [8]. The annotated dataset is also available at the GitHub repository of this thesis [4].

Manually annotated recommended papers, typically including the paper title, name of the benchmarks, or the models, were submitted to the Semantic Scholar (S2) API [39] to retrieve the corresponding S2 Paper ID for each reference. Once the unique identifiers were obtained, additional API calls were used to extract structured metadata for each paper, including its full title, abstract, primary and secondary fields of study as defined by Semantic Scholar’s classifier. Its field-of-study classifier is based on abstracts and titles. This process enabled a consistent labelling of fields of study (e.g., “Computer Science,” “Mathematics,” “Cognitive Science,” etc.) for subsequent quantitative analysis with an accuracy about 86% [50].

(b) To determine the field distribution of the references originally cited in the submissions, each paper was first processed using a combination of the Python `PyPDF2` library and regular expression-based extraction. This approach targeted the "References" section of each PDF to isolate citation entries. Given the formatting inconsistencies and occasional encoding issues common in PDF documents, the extracted reference strings were not always cleanly separated or standardized. To address these issues, `llama-70B` was used again to assist in cleaning and parsing the reference data. Specifically, the LLM was prompted to extract and format key citation metadata, particularly paper titles, author names, and publication years, from the raw strings. Manual checking after LLM processing was done to prevent data being altered.

After preprocessing, the cleaned citation entries were matched to records in the S2 database via their public API. This matching process is identical to what was done to retrieve paper information for recommended papers. Namely, it involved querying each citation by title (and, when needed, by additional metadata like authors and year) to retrieve the corresponding S2 Paper ID. Once the correct record was identified, use the API again to retrieve the associated fields of study for each cited work.

Step **(a)** and **(b)** together allows direct comparison between the field distribution of citations originally included by the authors and those recommended by peer reviewers.

[4] Annotated datasets available on GitHub

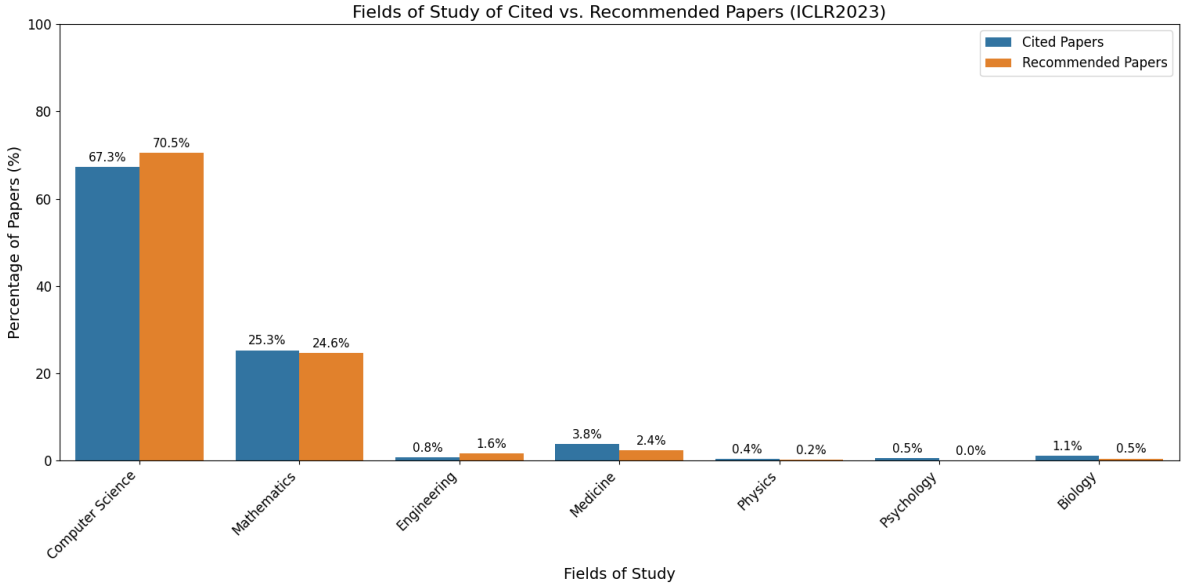


Fig. 7. Fields of study of cited and recommended papers in ICLR 2023. The graph shows the Top 7 fields of study. (QA3, 4.2.3)

Results & Discussion. The results for both **(a)** and **(b)** show that the majority of both recommended literature and cited references fall within the field of CS, which is not surprising given the domain of the conference. However, this also reinforces a degree of insularity, as such patterns can narrow our scholarly horizons and limit cross-disciplinary exploration. To use ICLR 2023 as an example, CS as the most recommended and cited field composes around 70% of all the fields, followed by Mathematics and Medicine as the second and third largest fields in both cases, with around 20% for the former and less than 5% for the latter (Figure 7).

Even though CS has always been the absolute majority of cited and recommended papers in the ICLR context, the CS dominance of ICLR 2023 has increased ~10% compared to that of ICLR 2013 (See Figure 8). This trend is although not consistent over the decade, with ICLR 2019 being an exception, ICLR 2023 ended up having the most CS-related papers in recommended papers. This result could suggest a narrowing of interdisciplinary engagement from reviewer suggestions, potentially at the cost of diverse perspectives from neighboring domains like neuroscience, cognitive science, or the social sciences, which were more visible in earlier years of the conference.

4.2.4 *(QB1) In which topic areas are the papers recommended by peer review? Is there a bias toward recommending specific topics?*

- a) In which topic is the submitted paper
- b) In which topic is the suggested citation by a reviewer
- c) How do peer reviewers' focus on current trends in AI lead to citing recent work?

Motivation. As shown in previous *QA3* (4.2.3), when papers in all the studied venues cited or got recommended extra literature, over 60% or even 70% of then falls into CS-related field. However, CS itself is also a very broad subject with tens of subfields. AI as one of them also includes various research areas. This research question therefore zooms in to investigate whether peer reviewers tend to recommend citations within specific topic areas, such as currently trending topics in AI such as Reinforcement Learning or generative AI. If reviewers

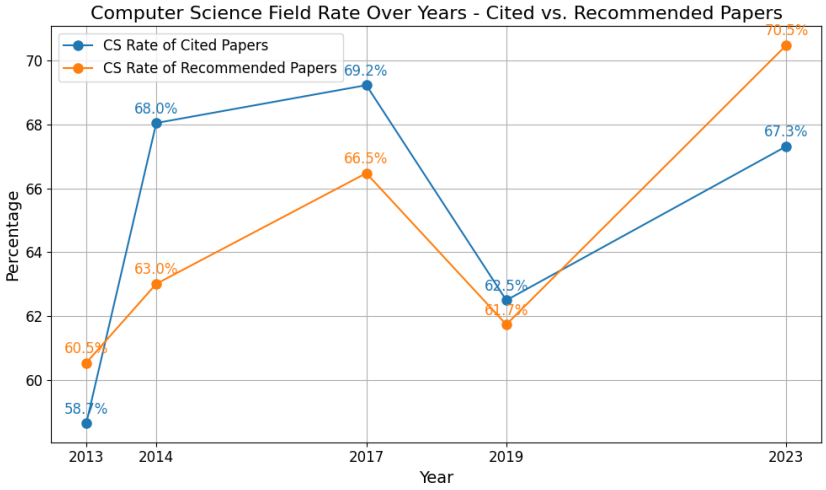


Fig. 8. Percentage of both CS-related cited and recommended papers across ICLR 2013 - 2023. (QA3, 4.2.3)

predominantly suggest citations from certain areas, this behaviour may amplify insularity biases by further concentrating attention on a narrow range of topics. It is also important to assess whether reviewers are recommending citations from the same topical area as the submission, or guiding papers toward new directions.

Method. To answer this question, I compared the topical distributions of submitted papers and the citations recommended by reviewers. For each of the venue in this research question–EMNLP 2023, ICLR 2023, and NeurIPS 2024–I began with clean lists of references from the submitted papers and reviewer-suggested papers compiled from earlier stages of the thesis. For EMNLP 2023, I extracted the most common bigrams from titles of both cited and recommended papers, then manually assigned topics based on those phrases. To decide the topics, I used the ARR taxonomy [2] as a reference. In case of ambiguous bigrams, I allowed multiple topic labels. I annotated the top 200 instances of recommended citations and cited references. For ICLR 2023 and NeurIPS 2024, the topic label of each submission was taken directly from the "primary topic" field of the OpenReview metadata. For recommended citations in these two venues, I manually reviewed the full paper titles, rather than keyword patterns as for EMNLP 2023, to assign topics, aligning with the topic schema used in submission data. I annotated 400 citation titles per venue.

For each venue, I evaluated whether the recommended citation has the same topics distribution as the submission overall. On top of this analysis, I also looked closer into whether paper with a certain topic usually got recommended with literature from the same topic. For this point, only the top 3 topics of the submitted papers are examined.

Results & Discussion. Based on the topic comparison as shown in Figure 9, 10 and 11, several clear trends emerge regarding the kinds of topics peer reviewers tend to recommend papers from, compared to the topics of submitted papers.

At ICLR 2023, there is a noticeable shift in focus from the topics of the submitted papers to those of the recommended ones. Most obviously, "Deep Learning and Representational Learning," while dominating the submitted papers, appears far less frequently among recommended citations. On the other hand, topics such as "Applications," "Unsupervised and Self-Supervised Learning," and "Probabilistic Methods" are favoured in the recommendations compared to the submissions. This suggests that reviewers may encourage authors toward

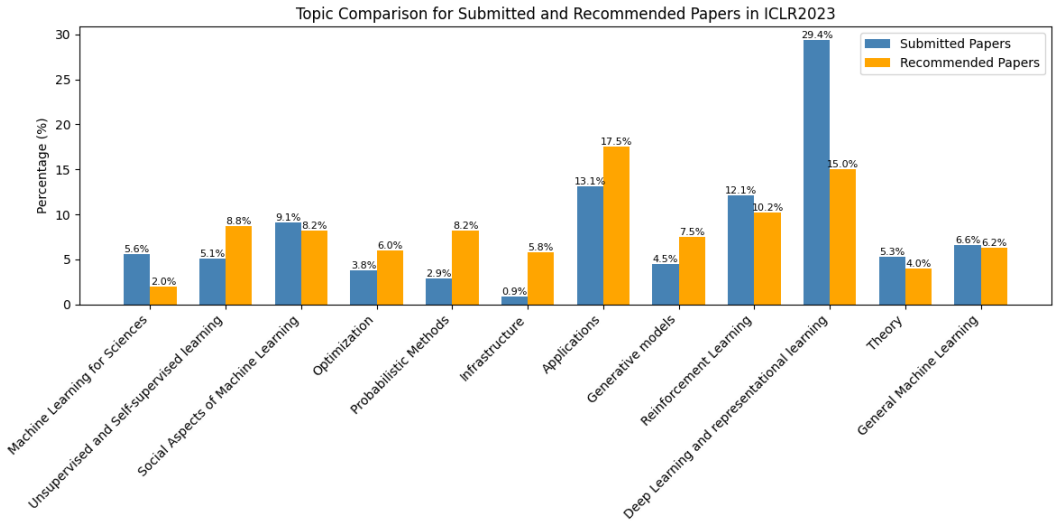


Fig. 9. Distribution of Cited and Recommended Topics in ICLR 2023 (QB1, 4.2.4)

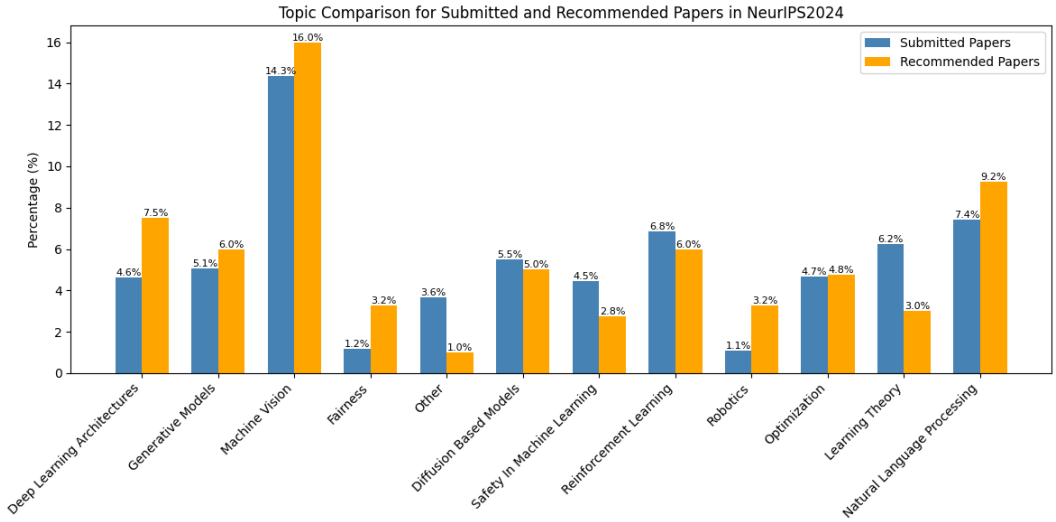


Fig. 10. Distribution of Cited and Recommended Topics in NeurIPS 2024 (QB1, 4.2.4)

broader or more diverse areas. The possible motivation could be encouraging grounding deep learning work with more applied or methodological contributions.

In the case of NeurIPS 2024, a few topics show clear reviewer preferences. For example, “Graph Neural Networks” (GNN) appear frequently in both submitted and recommended papers, but their proportion is slightly higher among recommendations, which indicates a consistent and possibly increasing interest from reviewers

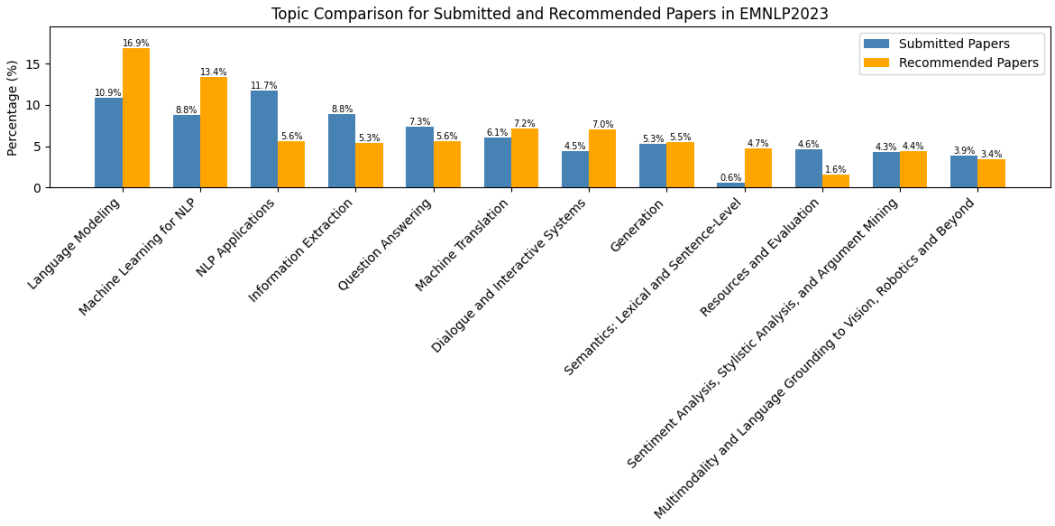


Fig. 11. Distribution of Cited and Recommended Topics in EMNLP 2023 (QB1, 4.2.4)

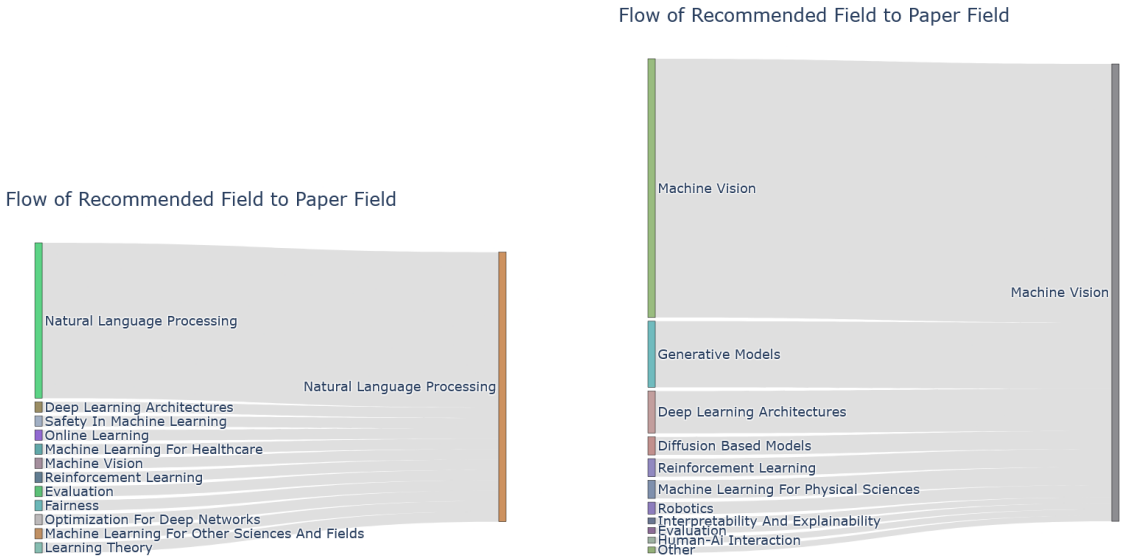
in that area. Topics like “Natural Language Processing” and “Deep Learning architectures” also have higher frequencies in recommended citations than in submissions, while topics such as “Machine Learning for Other Science and Fields” and “Learning Theory” are rather underrepresented in recommendations. The overall trend in this venue suggests that while reviewers may be reflecting current trends like GNN, applications in NLP (possibly LLM related), and large architectures, they also appear to amplify certain niche, potentially pushing submissions toward these topics. However, lack of interest in “Machine Learning for Other Science and Fields” among reviewers is a sign of decreasing interaction with fields outside of Computer Science.

For EMNLP 2023, the influence of trends in recommendation practices is even more apparent. “Language Modeling” is significantly overrepresented among recommended papers relative to submissions, despite already being a major theme in submitted work. “Machine Learning for NLP” and “Machine Translation” are also slightly more emphasized in reviewer suggestions. On the other hand, application-orieted topics like "Question Answering", “Information Extraction” and “NLP Applications” appear more in submitted papers than in recommended ones, suggesting that reviewers might be de-emphasizing more traditional or practical areas in favor of newer, more model-focused paradigms.

Across all the venues, a common pattern exits: peer reviewers frequently suggest papers from different subsets of the field. This can reinforce topical shift, increasing the field’s insularity by guiding citation practices toward another set of areas. It also indicates that reviewers might contribute to the increasing emphasis on recency and technical novelty–possibly encouraging authors to frame their work in ways that align with the current hype or perceived cutting-edge directions of the field.

Further results on the dynamics between incoming fields and the main current topics of NeurIPS 2024 are visualised via the example of "Natural Language Processing" and "Machine Vision" in Figure 12 [5]. In both examined topics, there is a notable trend of relatively strong within-topic suggestion, meaning that around half of the NLP papers tend to be suggested to other NLP papers, and the same pattern holds for other top fields. However, some cross-topic recommendation do appear. It is noteworthy that generative model is a relatively big incoming field

[5]See Appendix B.3 for Sankey Diagrams of other topics and venues



(a) NLP (b) Machine Vision

Fig. 12. Incoming fields to **NLP** and **Machine Vision** papers in NeurIPS 2024. The left column of each Sankey diagram is the fields of study across all the papers that got recommended to the papers identified as **NLP** or **Machine Vision** (right column) by the authors themselves.

towards machine vision in NeurIPS 2024, which could indicate a trend of mono- or multimodal generative vision models (e.g., vision-language models). The diverse influence on NLP papers in NeurIPS 2024 also suggests that NLP is attracting great attention from different perspective, such as fairness, safety, LLM as agent (which often being classified as reinforcement learning papers), etc.

4.2.5 *(QB2) Does the paper decision correlate with citation recommendation, i.e., does rejected papers get more citation recommendations compared to accepted papers?*

Motivation. This question aims to uncover whether peer reviewers use citation suggestions as a corrective mechanism, namely, giving more recommendations to weaker papers, or whether citation suggestions are evenly distributed regardless of paper quality. Understanding this relationship could help clarify whether citation suggestions primarily aim to improve underperforming work, or highlighting work in the same topical area. It also provides insight into whether citation recommendation behaviour subtly reflects reviewers' perceptions of paper quality.

Method. To test whether there is a correlation between paper decisions and the number of citation suggestions made by reviewers, I used a Chi-square test of independence on contingency tables. The first test categorizes papers by acceptance status (accepted or rejected) and by whether they received at least one citation suggestion. The second test provides a more granular breakdown by categorizing papers according to their exact number of reviewers who made citation recommendations. This allows me to assess whether rejected papers are more likely to receive higher numbers of citation suggestions.

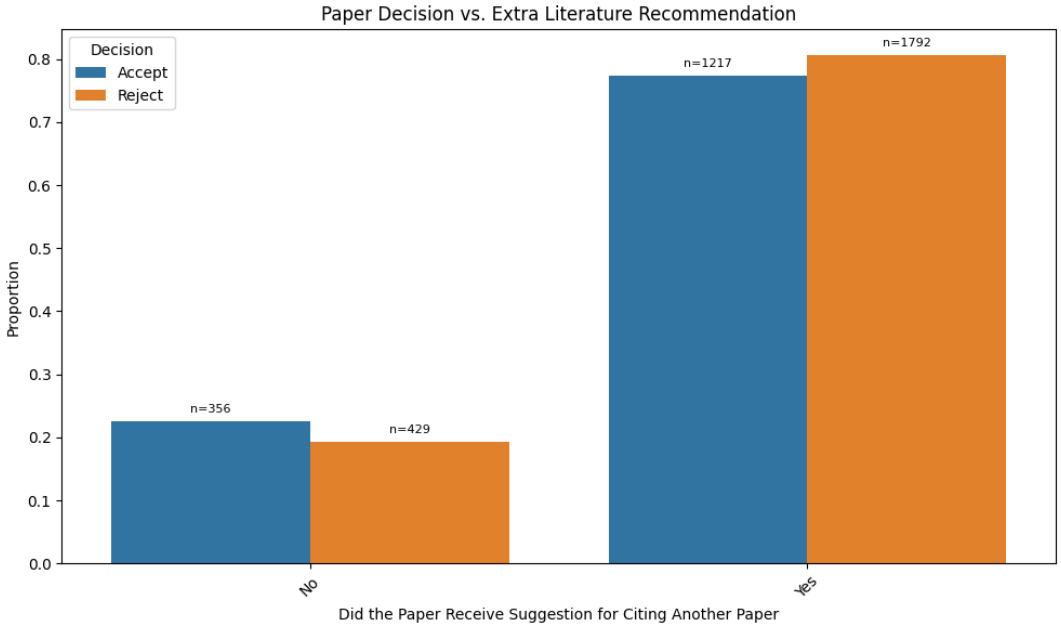


Fig. 13. Correlation between paper decision and getting citation recommendation from reviewers. (QB2, 4.2.5)

The Chi-square statistic is computed using the formula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{8}$$

where O_{ij} denotes the observed number of papers with decision outcome i (e.g., accepted or rejected) 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.

Results & Discussion. In the binary version of the test (0 vs. 1 or more reviewers recommending citations), the chi-square statistic was 5.97 with a p-value of 0.0145, indicating a statistically significant difference, meaning rejected papers were more likely to receive at least one citation recommendation than accepted ones. In the extended version with finer-grained counts (0 through n reviewers making recommendations), the chi-square statistic was even stronger at 35.66 with a p-value < 0.0001, again showing a significant relationship between rejection and the number of citation suggestions. Both versions are visualised via Figure 13 and 14.

These results suggest that reviewers are more likely to recommend citations for papers that are ultimately rejected. This supports the idea that citation suggestions may serve a corrective role. In other words, reviewers propose additional literature when they perceive the submission to be a weaker candidate for the conference. It also indicates that reviewers may use citation recommendations to encourage deeper engagement with related work when the submission appears underdeveloped to them. Ultimately, this finding highlights that citation

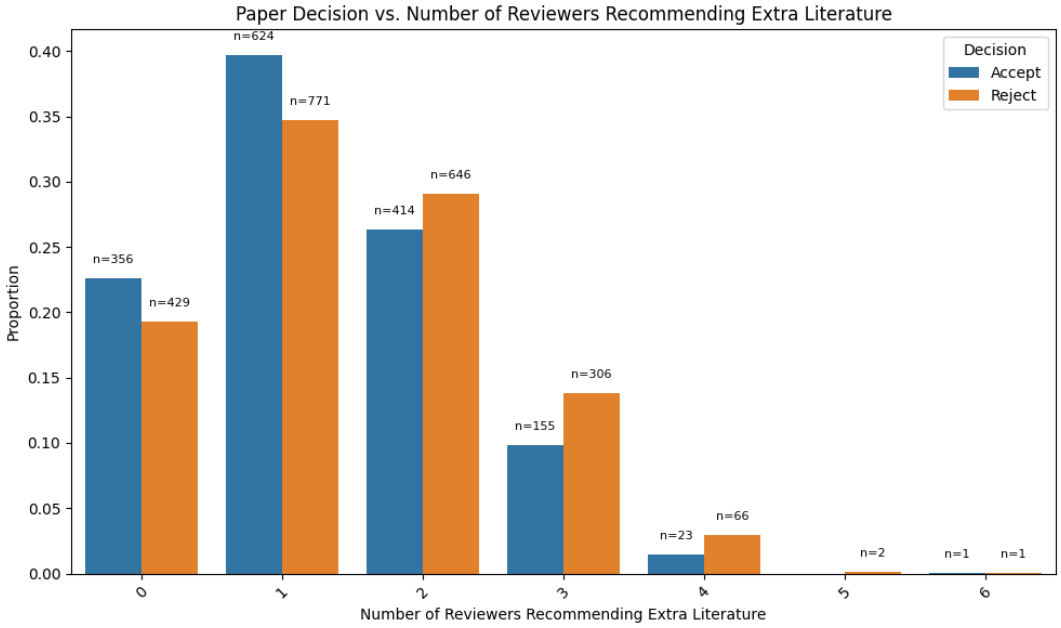


Fig. 14. Correlation between paper decision and the number of citation recommendation (QB2, 4.2.5)

suggestions are not purely neutral references to related work but may function as subtle signals of perceived deficiencies in a paper.

4.2.6 *(QC1) How often do authors incorporate the citing suggestions from the reviewers?*

Motivation. The previous experiments has shown that the reviewers often suggest extra literature to the authors, and sometimes the recommended papers are even from different topic areas comparing to the submitted paper itself. However, it remains unclear whether these suggestions influence the final paper. As a last step, it is important to examine whether or not the authors actually integrated those recommended papers.

Method. For determining whether authors actually incorporated the suggested citations, I cross-referenced submission IDs from the rebuttal data with both the list of recommended papers and the reference sections of the final camera-ready PDFs. By checking whether the recommended paper appeared in the final bibliography, I could confirm actual inclusion.

Results & Discussion. Table 6 presents the high proportion of authors who actually included the recommended references in the final submission, with incorporation rates ranging from 62.17% to 79.18%. Although NeurIPS 2023 and 2024 together are still not adequate for a comprehensive chronological study, NeurIPS 2024 does show an increase in the inclusion rate of the recommended papers. This implies most of the authors choose to incorporate the suggested citations from the reviewers and this might also be a growing trend. This finding suggests that the recency and insularity biases raised by recommendation of extra literature are very likely to be adopted by the authors.

Venue (Year)	Inclusion Rate
EMNLP 2023	62.17%
ICLR 2023	73.38%
NeurIPS 2023	74.24%
NeurIPS 2024	79.18%

Table 6. Percentage of authors who actually cited the recommended papers in their final submission. (QC1, 4.2.6)

4.2.7 *(QC2) Do papers that heavily cite recent work receive better peer review outcomes/ higher review scores (i.e., being accepted or having higher reviewer scores) than those with a more balanced or diverse citation age?*

Motivation. At this stage, the thesis has discovered that reviewers indeed recommend more recent literature and authors do tend to incorporate them into the camera-ready version of their papers, but whether reviewers reward papers that align closely with current trends remains unknown. 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. This research question helps reveal whether author's citation behaviour is being shaped by considerations to appeal to reviewer expectations.

Method. To assess the relationship between citation recency and review outcomes, I first quantified the "recency" of a paper's references. For each submission, I extracted the publication years of all cited papers from the reference list in the PDF, I 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 and rejected 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 7. 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 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 7. Main Venues Subsets Studied for QC2, 4.2.7

Results & Discussion. Figure 15 and Figure 16 show the average categorical and final review scores for ICLR 2023 papers, grouped by the citing year of referenced works. Figure 15 illustrates that papers citing older

literature tend to receive the lowest average scores in both technical and empirical novelty, whereas papers referencing more recent work show the highest scores. However, the figure does not reveal a clear or consistent trend, and the overall differences in scores are relatively small. To determine whether the distribution of review scores differs significantly among papers with different average citing years, I again conducted Chi-square tests as shown in Equation 8. Note that the score bins of final scores differ from that of individual scores: for technical and empirical novelty scores, the bins are from 1 to 4, while the final scores are from 1 to 8. Although the highest possible rating of final score is 10, only very few papers received this score from some but not all of the reviewers, making the average score usually below 8. Therefore, to avoid introducing redundant categories, the score bins of final rating is set to 1 to 8. 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 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 16, the trend is clearer. Namely, when papers cite more recent literature, they are more likely to be assigned higher final scores by the reviewers. 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. But this initial 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, I also conducted the same analysis to both ICLR 2023 (Accepted) and ICLR 2023 (Rejected) separately. It turns out that only the scores of rejected papers correlate with their average citing years (See Table 8). In other words, in ICLR, 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 awarded for citing more newer literature.

Interestingly, reviewers of NeurIPS 2023 and 2024 show different behaviour comparing to ICLR 2023. To take NeurIPS 2024 as an example (See Figure 17), the more "recent" a reference section is, the lower scores its paper received. If looking at individual categorical scores in Figure 18, 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 9. 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-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 8. Chi-square test for different ICLR 2023 subsets. (QC2, 4.2.7)

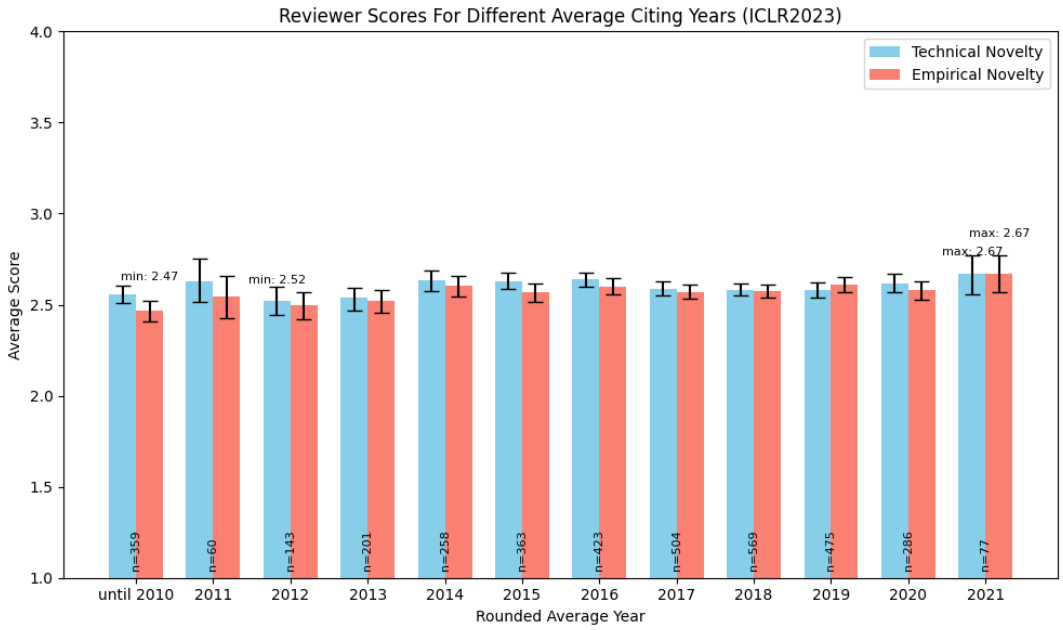


Fig. 15. Categorical reviewer scores for different average citing years in ICLR 2023 (Total). Average Citing Year = Sum of years of all cited works in a paper / Number of citations in that paper. The result is rounded to the nearest integer. Consecutive years with few data points are aggregated (e.g., "until 2010"). (QC2, 4.2.7)

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 9. Chi-square test for different EMNLP and NeurIPS subsets (QC2, 4.2.7)

5 Conclusion

This thesis explored the relationship between peer review and citation practices in the context of top-tier AI conferences. Through a quantitative analysis based on data from OpenReview, the analysis particularly focused on how reviewers recommend citations, how authors respond to such recommendations, and what implications these interactions have on AI research and paper outcomes.

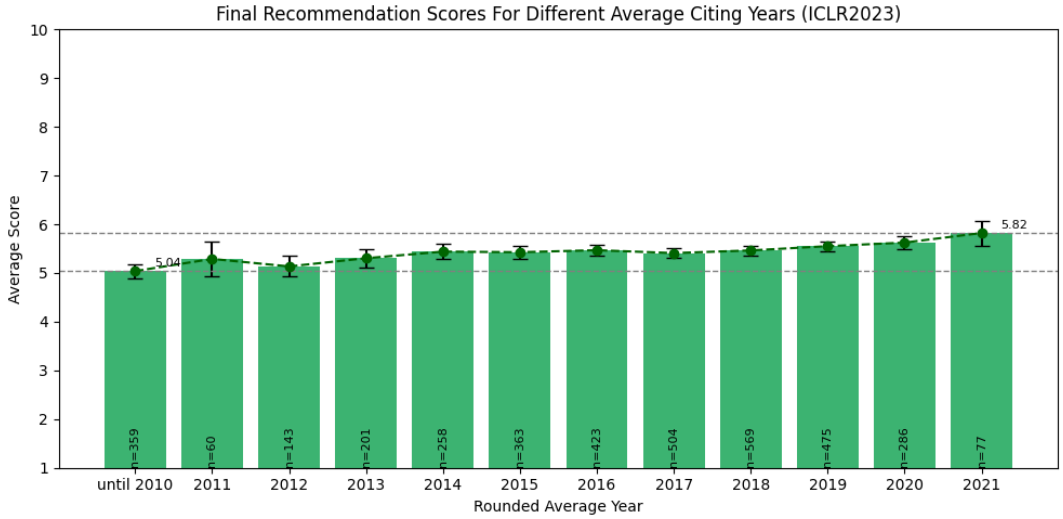


Fig. 16. Total reviewer scores for different avg. citing years in ICLR 2023 (Total). Average Citing Year = Sum of years of all cited works in a paper / Number of citations in that paper. The result is rounded to the nearest integer. Consecutive years with few data points are aggregated (e.g., "until 2010"). (QC2, 4.2.7)

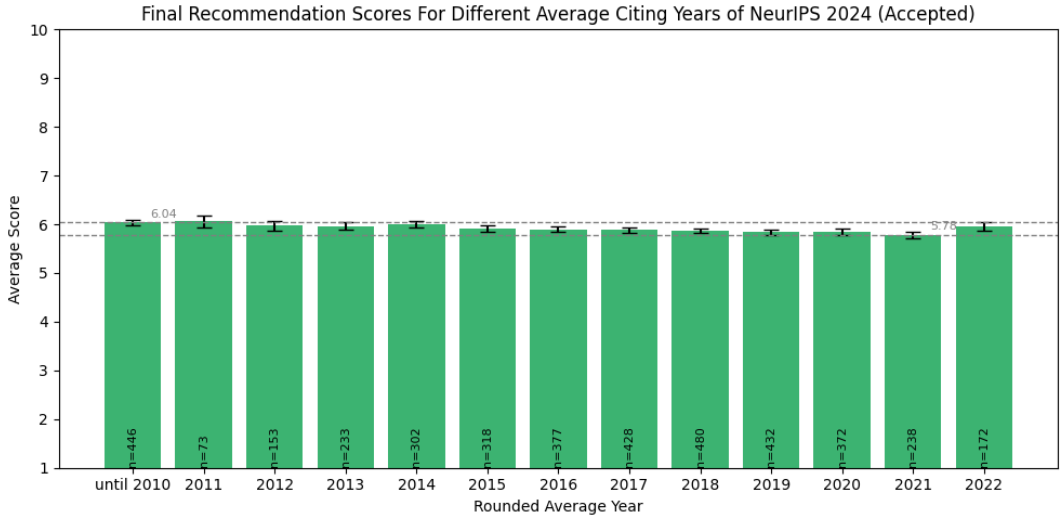


Fig. 17. Total reviewer scores for different avg. citing years in NeurIPS 2024 (Accepted). Average Citing Year = Sum of years of all cited works in a paper / Number of citations in that paper. The result is rounded to the nearest integer. Consecutive years with few data points are aggregated (e.g., "until 2010"). (QC2, 4.2.7)

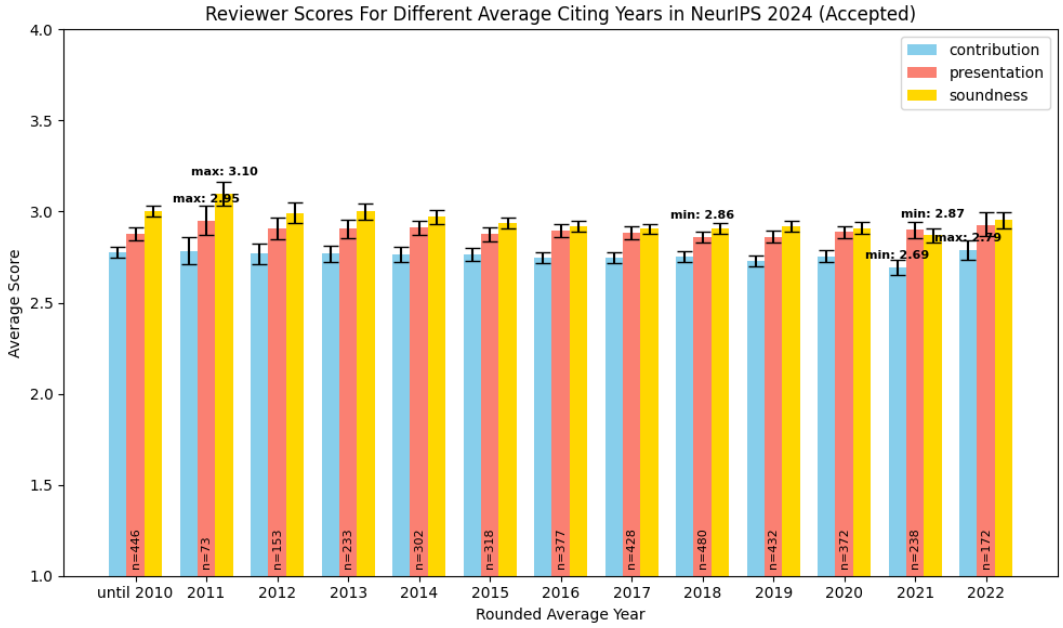


Fig. 18. Categorical reviewer scores for different avg. citing years in NeurIPS 2024 (Accepted). Average Citing Year = Sum of years of all cited works in a paper / Number of citations in that paper. The result is rounded to the nearest integer. Consecutive years with few data points are aggregated (e.g., "until 2010"). (QC2, 4.2.7)

5.1 Key Findings

By going back to the three categories set at the beginning of Section 4 (the *Experiments* section), this section aims to draw some short conclusions on the results of the main experiments.

Part A. Quantitative Analysis of Citation Recommendations. First, reviewer citation recommendations are widespread and frequent across top-tier AI conferences like ICLR, NeurIPS, and EMNLP. In all the venues examined in this thesis, more than half of the papers got at least one citation recommendation from the reviewers. In ICLR 2023, this number researches 79.32%. Reviewers also tend to recommend more recent papers compared to the papers cited by the authors themselves, with the biggest citation age gap being 5.5 years – found in NeurIPS 2024. This partially contributes to the observed recency bias in the AI research field. Besides, by tracking historical data of ICLR, the thesis also revealed that the insularity of recommended literature has increased over the years, namely, more and more recommended papers come from the field of Computer Science. It is also noteworthy that recommended literature usually have higher proportion of CS papers compared to that of cited papers in the manuscript.

Part B. Analysis of Reviewers' Suggestions and Their Influence on Research Focus. Within the AI field itself, these recommendations sometimes come from topics that are not fully aligned with the core subject of the submitted manuscript in many cases, which suggests that citation suggestions may sometimes reflect a reviewer's own expertise or biases rather than the paper's central contribution. As a result, this shift of topic could lead to the slight change of trajectory of the future research. The thesis also revealed interesting correlations between paper acceptance and citation recommendation behaviour. The analyses indicate that rejected papers tend to receive slightly more citation recommendations, and in finer-grained tests, this relationship is also statistically significant.

This finding suggests that reviewers may be more inclined to recommend additional citations, perhaps as a form of critique or gate-keeping, when they are leaning toward rejection, potentially using citation recommendations as a means to assert influence or signal perceived shortcomings in the submission.

Part C. Authors' Responses to Reviewers' Citation Suggestions. By comparing reviewers' recommendation and authors' final submissions, the thesis checked if authors actually cited the recommended work. It turns out that the actual inclusion of the suggested papers in final versions are high, exceeding 60% in all cases and peaking at nearly 80% in NeurIPS 2024. This finding points out that authors do integrate more recent papers or papers within the same field in their final version, resulting an actual influence of the scholarly publication. Lastly, the analysis on the correlation between citation age of papers and their final decision revealed noteworthy insights into how the temporal distribution of citations may affect their reception. Although the evidence does not point to a strong causal relationship between citation recency and peer review scores or acceptance outcomes, the findings suggest that reviewer preferences might vary depending on the perceived balance between foundational works and cutting-edge literature, especially in the cases of inadequate papers.

5.2 Implications and Future Research

Overall, these findings deepen our understanding of peer review as an interactive and multifaceted process that shapes not only what is published but also what is cited and eventually what is under the spotlight overtime. By uncovering how citation recommendations are given and adopted during peer review, this thesis positions the reviewer as an active participant in shaping the trajectory of AI research. Practically, the insights from this study also have implications for improving the peer review process. Greater transparency around citation suggestions, e.g., reasoning behind them, could help reduce potential biases.

This thesis also opens several directions for future research. One would be to study citation suggestion behaviour over time with more historical data, to trace how they evolve in response to shifts in conference regulations or broader academic norms. Another would be expanding the scope to other fields outside of AI. This could offer a clearer and bigger picture of whether AI is the only field that has witnessed growing recency and insularity biases stemming from peer reviews.

Ultimately, by examining the often-overlooked citation-related recommendations in peer reviews, it sheds light on the subtle but consequential ways in which peer review might affect the trajectory of future research.

6 Limitations

While this thesis offers new insights into reviewer citation suggestions and its influence on recency and insularity biases, several limitations must be acknowledged. These limitations include data availability, methodological constraints, and interpretative ambiguities, which should be considered when evaluating the results and drawing broader conclusions.

6.1 Limitations of Data Sources

One of the primary limitations in this thesis is the reliance on publicly available data from the OpenReview platform. Although OpenReview does provide a rich dataset for certain conferences, such as ICLR, which has both old data dating back to 2013 as well as abundant rejected papers with reviews and submitted PDFs, it does not have access to many rejected papers in major venues or provide consistent data granularity across years. For example, EMNLP and NeurIPS do not have enough data for rejected papers as ICLR does, nor do they have available data earlier than 2019, which limited the depth of analysis that could be conducted for these venues. The lack of data on rejected papers could be, however, due to authors' own deletion. For the limited amount of historical data, it is because some venues only started using OpenReview in the past few years. Nevertheless, these problems limited the depth of analysis that could have been conducted for this thesis.

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. Although some citation suggestions are still given in other fields. This resulted in relying on regular expressions or LLMs to extract information, which is inherently bias-prone and may lead to both false positives and false negatives.

6.2 Methodological Assumptions and Biases

To make the analysis tractable, some methodological simplifications were necessary, but these come with trade-offs. For instance, the analysis assumes that if a recommended paper appears in the reference section of the camera-ready version, it was directly influenced by the reviewer's suggestion. This assumption does not account for independent author decisions. In cases where authors would have cited a work regardless of reviewer recommendation, the effect of the reviewer's input may be exaggerated.

In addition, for certain research questions, I used manual annotations to ensure a more reliable results compared to using automated methods. However, it is by no means bias-proof. Especially when there is only one annotator, the method could introduce potential annotator subjectivity. Furthermore, the scale of manual annotation was necessarily limited due to time and resource constraints, potentially affecting the generalization of the findings derived from these subsets.

6.3 Temporal and Behavioural Changes

Another limitation is the static nature of the thesis with respect to time. Reviewer behaviour, citation norms, and author response strategies evolve over time. For instance, increasing awareness of reviewer bias, new guidelines from conference organizers, or the new AI writing assistance may all affect how citations are suggested and incorporated. Despite these influences, the thesis does not explore how these variables interact longitudinally or how changes in peer review formats, e.g., double-blind versus open review, might impact citation suggestion behaviours.

Acknowledgment

I sincerely thank Dr. Terry Lima Ruas and Jan Philip Wahle for their insightful discussions, valuable ideas, and continuous support during the process of this thesis. Besides, I am also grateful to them for providing group API keys to LM Studio hosted on GippLab server and Semantic Scholar, which were essential for conducting parts of this research. I would also like to express my gratitude to GWDG for granting API with access to some LLMs, which also played an important role in supporting this work.

References

[1] 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 (2023).

[2] ACL Rolling Review. 2025. Area Keywords at ARR. https://aclrollingreview.org/areas

[3] Andres Algaba, Vincent Holst, Floriano Tori, Melika Mobini, Brecht Verbeken, Sylvia Wenmackers, and Vincent Ginis. 2025. How Deep Do Large Language Models Internalize Scientific Literature and Citation Practices? arXiv:2504.02767 [cs.DL] https://arxiv.org/abs/2504.02767

[4] Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. 2025. Large Language Models Reflect Human Citation Patterns with a Heightened Citation Bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 6829–6864. https://aclanthology.org/2025.findings-naacl.381/

[5] Ian Ayres and Fredrick E Vars. 2000. Determinants of citations to articles in elite law reviews. *The Journal of Legal Studies* 29, S1 (2000), 427–450.

[6] 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* (2008).

[7] Gualberto Buela-Casal and Izabela Zych. 2010. Analysis of the relationship between the number of citations and the quality evaluated by experts in psychology journals. *Psicothema* (2010), 270–276.

[8] Calculator.net. 2025. Sample Size Calculator. https://www.calculator.net/sample-size-calculator.html.

[9] Paula Chatterjee and Rachel M Werner. 2021. Gender disparity in citations in high-impact journal articles. *JAMA Network Open* 4, 7 (2021), e2114509–e2114509.

[10] François Collet, Duncan A Robertson, and Daniela Lup. 2014. When does brokerage matter? Citation impact of research teams in an emerging academic field. *Strategic Organization* 12, 3 (2014), 157–179.

[11] Rodrigo Costas, Maria Bordons, Thed N Van Leeuwen, and Anthony FJ Van Raan. 2009. Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of individual researchers. *Journal of the American Society for Information Science and Technology* 60, 4 (2009), 740–753.

[12] Derek John de Solla Price. 1962. *Science Since Babylon.* Yale University Press, New Haven, CT.

[13] S. Ebadi, H. Nejadghanbar, A. R. Salman, and others. 2025. Exploring the Impact of Generative AI on Peer Review: Insights from Journal Reviewers. *Journal of Academic Ethics* (2025). https://doi.org/10.1007/s10805-025-09604-4

[14] Matthew E Falagas, Angeliki Zarkali, Drosos E Karageorgopoulos, Vangelis Bardakas, and Michael N Mavros. 2013. The impact of article length on the number of future citations: a bibliometric analysis of general medicine journals. *PloS One* 8, 2 (2013), e49476.

[15] John Flowerdew. 2001. Attitudes of journal editors to nonnative speaker contributions. *TESOL Quarterly* 35, 1 (2001), 121–150.

[16] Eric A. Fong and Allen W. Wilhite. 2017. Authorship and citation manipulation in academic research. *PLOS ONE* 12, 12 (12 2017), 1–34. https://doi.org/10.1371/journal.pone.0187394

[17] Bela Gipp and Norman Meuschke. 2011. Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. In *Proceedings of the 11th ACM Symposium on Document Engineering* (Mountain View, California, USA) *(DocEng '11)*. Association for Computing Machinery, New York, NY, USA, 249–258. https://doi.org/10.1145/2034691.2034741

[18] Max Martin Gnewuch. 2024. What Impact Does Big Tech Funding Have on AI Research? A Scholarly Document Analysis. (2024).

[19] Serge P.J.M. Horbach and Willem Halffman. 2019. The hidden role of editors in scientific misconduct. *Accountability in Research* 26, 3 (2019), 154–184.

[20] Ahmad Ismail, Hardiyanti Munsi, Andi Muhammad Yusuf, and Pawennari Hijjang. 2025. Mapping One Decade of Identity Studies: A Comprehensive Bibliometric Analysis of Global Trends and Scholarly Impact. *Social Sciences* 14, 2 (2025). https://doi.org/10.3390/socsci14020092

[21] Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring Peer Review Dynamics with LLM Agents. 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, 1208–1226. https://doi.org/10.18653/v1/2024.emnlp-main.70

[22] Adina R. Kern-Goldberger, Richard James, Vincenzo Berghella, and Emily S. Miller. 2022. The impact of double-blind peer review on gender bias in scientific publishing: a systematic review. *American Journal of Obstetrics and Gynecology* 227, 1 (2022), 43–50.e4. https://doi.org/10.1016/j.ajog.2022.01.030

[23] 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 (2013), 2–17.

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

[25] Rodrigo Pessoa Cavalcanti Lira, Rafael Marsicano Cezar Vieira, Fauze Abdulmassih Gonçalves, Maria Carolina Alves Ferreira, Diana Maziero, Thais Helena Moreira Passos, and Carlos Eduardo Leite Arieta. 2013. Influence of English language in the number of citations of articles published in Brazilian journals of Ophthalmology. *Arquivos Brasileiros de Oftalmologia* 76 (2013), 26–28.

[26] Anaïs Llorens, Athina Tzovara, Ludovic Bellier, Ilina Bhaya-Grossman, Aurélie Bidet-Caulet, William K Chang, Zachariah R Cross, Rosa Dominguez-Faus, Adeen Flinker, Yvonne Fonken, and others. 2021. Gender bias in academia: A lifetime problem that needs solutions. *Neuron* 109, 13 (2021), 2047–2074.

[27] Sheng Lu, Ilia Kuznetsov, and Iryna Gurevych. 2025. Identifying Aspects in Peer Reviews. arXiv:2504.06910 [cs.CL] https://arxiv.org/abs/2504.06910

[28] Michael J Mahoney. 1977. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research* 1, 2 (1977), 161–175.

[29] Joseph Mariani, Gil Francopoulo, and Patrick Paroubek. 2019. The NLP4NLP Corpus (I): 50 Years of Publication, Collaboration and Citation in Speech and Language Processing. *Frontiers in Research Metrics and Analytics* 3 (2019). https://doi.org/10.3389/frma.2018.00036

[30] Saif M Mohammad. 2020. Gender gap in natural language processing research: Disparities in authorship and citations. *arXiv preprint arXiv:2005.00962* (2020).

[31] Saif M. Mohammad. 2020. NLP Scholar: A Dataset for Examining the State of NLP Research. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry

Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 868–877. https://aclanthology.org/2020.lrec-1.109/

[32] 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 (2016).

[33] OpenReview.net. 2024. OpenReview: Open Peer Review Platform. https://openreview.net.

[34] David Pontille and Didier Torny. 2015. Font of knowledge: Peer review and the epistemic authority of reviewers. *Science, Technology, Human Values* 40, 3 (2015), 327–350.

[35] Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. 2025. LazyReview A Dataset for Uncovering Lazy Thinking in NLP Peer Reviews. arXiv:2504.11042 [cs.CL] https://arxiv.org/abs/2504.11042

[36] Tony Ross-Hellauer. 2017. What is open peer review? A systematic review. *F1000Research* 6 (2017), 588.

[37] Mukund Rungta, Janvijay Singh, Saif M Mohammad, and Diyi Yang. 2022. Geographic citation gaps in NLP research. *arXiv preprint arXiv:2210.14424* (2022).

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

[39] Semantic Scholar. n.d.. Semantic Scholar API. https://api.semanticscholar.org/.

[40] Richard Smith. 2006. Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine* 99, 4 (2006), 178–182.

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

[42] 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 (2020), 791–801.

[43] Ivan Stelmakh, Charvi Rastogi, Ryan Liu, Shuchi Chawla, Federico Echenique, and Nihar B. Shah. 2023. Cite-seeing and reviewing: A study on citation bias in peer review. *PLOS ONE* 18, 7 (07 2023), 1–16. https://doi.org/10.1371/journal.pone.0283980

[44] Ivan Stelmakh, Vibhor Rastogi, and Nihar B Shah. 2019. Cite-seeing and reviewing: A study on citation bias in peer review. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2751–2759.

[45] 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 (2015), 1–6. https://doi.org/10.1016/j.jpsychores.2014.09.015

[46] Steven K. Thompson. 2012. *Sampling* (3rd ed.). Wiley.

[47] Andrew Tomkins, Min Zhang, and William D Heavlin. 2017. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences* 114, 48 (2017), 12708–12713.

[48] Hugo Touvron, Shruti Bhosale, Mikel Artetxe, Louis Martin, Kevin Lu, Eric Hambro, Faisal Azhar, Serkan Cabi, Tian Li, Zhanghao Wu, et al. 2024. LLaMA 3: Open Foundation and Instruction Models. arXiv:2404.14219 [cs.CL] https://arxiv.org/abs/2404.14219

[49] Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2023. Paraphrase Types for Generation and Detection. 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, 12148–12164. https://doi.org/10.18653/v1/2023.emnlp-main.746

[50] 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12896–12913. https://doi.org/10.18653/v1/2023.emnlp-main.797

[51] Jan Philip Wahle, Terry Ruas, Saif Mohammad, and Bela Gipp. 2022. D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2642–2651. https://aclanthology.org/2022.lrec-1.283/

[52] Jan Philip Wahle, Terry Ruas, Saif M Mohammad, and Bela Gipp. 2022. D3: A massive dataset of scholarly metadata for analyzing the state of computer science research. *arXiv preprint arXiv:2204.13384* (2022).

[53] Ankit Yadav and others. 2021. Peer review reports: A textual analysis of tone and length. *International Journal of Global Research and Review* 7, 2 (2021).

[54] Yu Zong and Xin Xie. 2020. Does open peer review improve citation count from a reviewer's perspective? *Journal of Informetrics* 14, 3 (2020), 101044.

A Prompts for *QA1*

Table 10 shows all the prompts used in 4.2.1.

Prompt	Content
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 10. Prompts for Detecting Literature Citation Suggestions in Peer Reviews (*QA1* 4.2.1)

B Additional Results and Graphs

B.1 Additional Graphs for *QA1*

This section shows the results of different model-prompt combinations for ICLR and NeurIPS, as a supplement for 4.2.1.

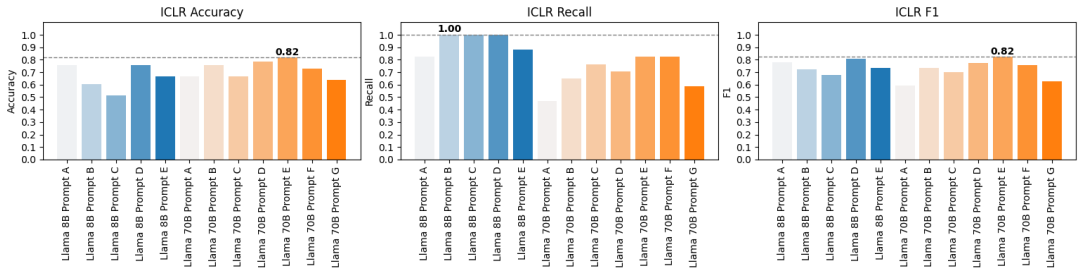


Fig. 19. Performance of different models-prompts for ICLR 2023 on different metrics (left to right: Accuracy, Recall, F1 score). The naming pattern of x-ticks is "Model X Prompt Y". **Accuracy** is particularly relevant, as the goal is to get more correct predictions. (QA1, 4.2.1)

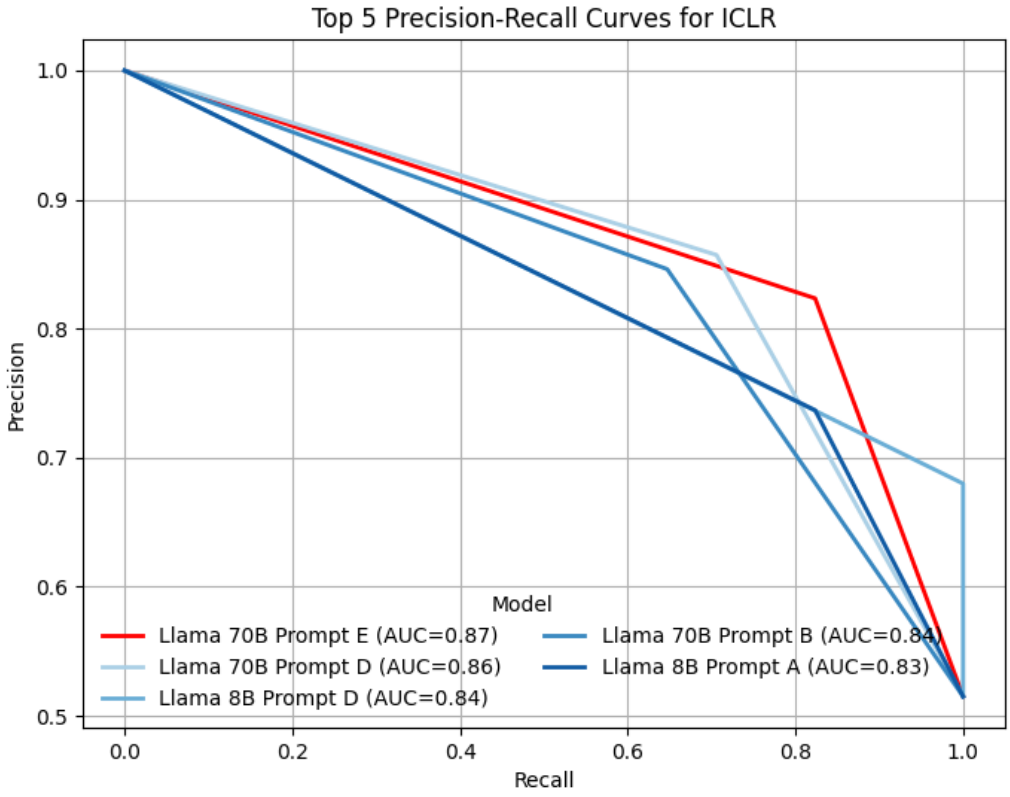


Fig. 20. Top 5 Precision-Recall Curves for ICLR 2023. (QA1, 4.2.1)

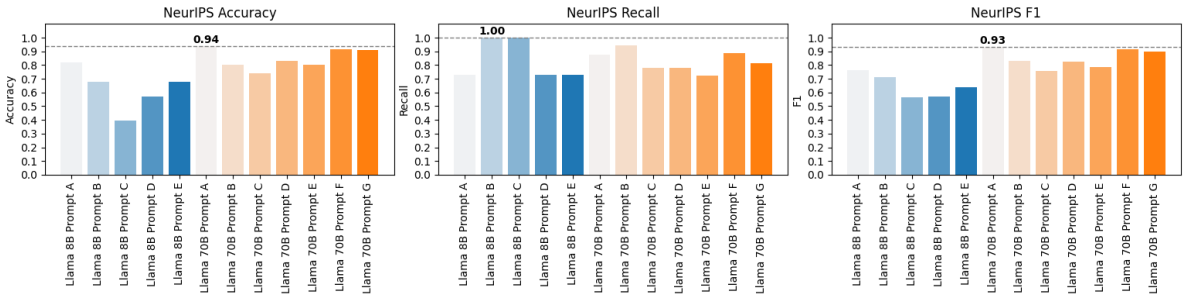


Fig. 21. Performance of different models-prompts for NeurIPS 2023 & 2024 on different metrics (left to right: Accuracy, Recall, F1 score). The naming pattern of x-ticks is "Model X Prompt Y". **Accuracy** is particularly relevant, as the goal is to get more correct predictions. (QA1, 4.2.1)

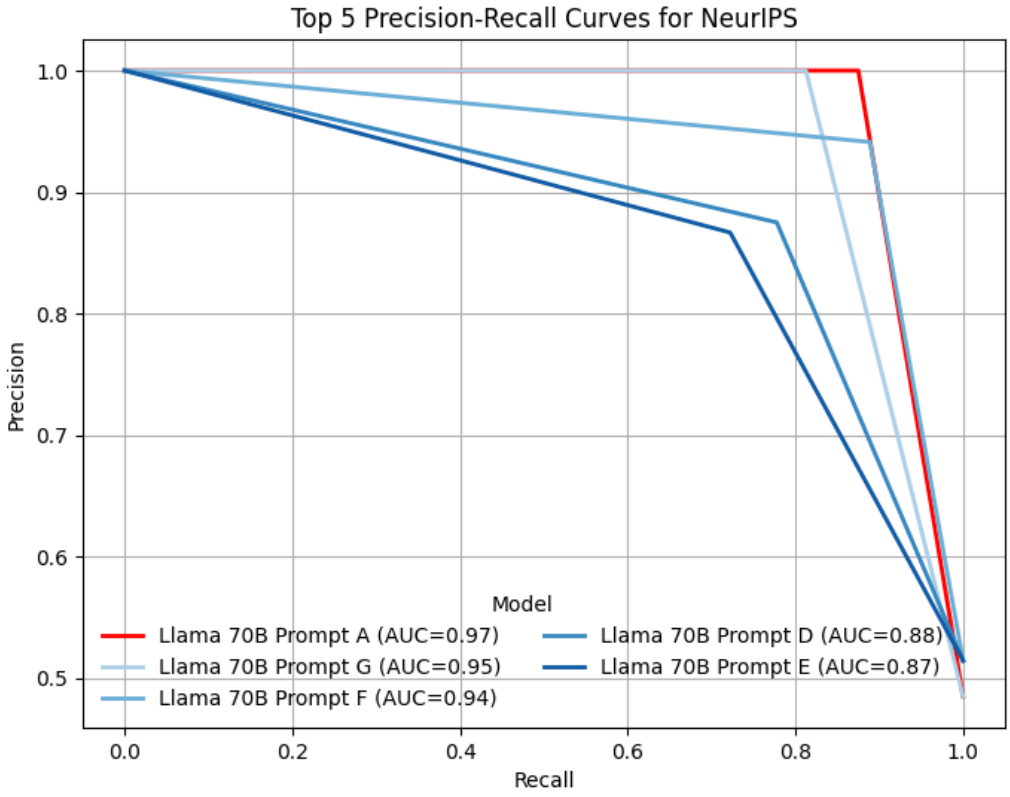


Fig. 22. Top 5 Precision-Recall Curves for NeurIPS 2023 & 2024 (QA1, 4.2.1)

According to Figure 19 and 20, the best model-prompt combination for ICLR 2023 is `Llama 70B prompt E` for having the highest AUC of 0.87. For NeurIPS 2023 and 2024 shown in Figure 21 and 22, `Llama 70B prompt A` is the best candidate, with an AUC of 0.97.

B.2 Additional Experiment for *Part B*

I conducted a supplementary experiment for Part B: What are the most common keywords and phrases used in a reject case as opposed to an accept case?

Motivation. This question investigates whether one common reason for rejection is the lack of engagement with relevant or recent literature. If reviewers often highlight insufficient citation coverage or missing comparisons with related work as part of the rejection reasons, that would offer evidence that citation practices can influence review outcomes.

Method. To explore this, I focused only on the ICLR 2023, as it provides the most complete data for both accepted and rejected submissions through OpenReview. I retrieved decision-level metadata in JSON format and identified the "decision" and "comment" (meta-review) entries for each paper. These entries were then merged

with the raw peer review data compiled during the analysis for *QA1*, and converted into a structured CSV format for downstream processing. Each row in the dataset included the final decision (accept or reject), and text fields for weaknesses and strengths.

The analysis proceeded in two main ways. First, I applied TF-IDF vectorization over the meta-review texts, separated by decision category (accept vs. reject), to identify distinguishing terms that tend to appear more frequently and distinctively in each group. Second, I conducted a frequency analysis on n-grams (uni-, bi-, and tri-grams) across decision rationales—again separated by decision class—to compare the most common phrases appearing in feedback for accepted versus rejected papers. In addition to general frequency analysis, I also performed category-specific filtering: for example, extracting the most common phrases in the "weaknesses" field of rejected papers, and comparing them to those in the "reasons to accept" fields of accepted papers. While I also attempted dimensionality reduction and clustering (using PCA and UMAP) on sentence embeddings of review texts to visualize distinctions in reviewer language between accepted and rejected papers, this approach was ultimately unproductive. The semantic overlap in review language, regardless of decision, was too high to yield meaningful separable clusters in low-dimensional space.

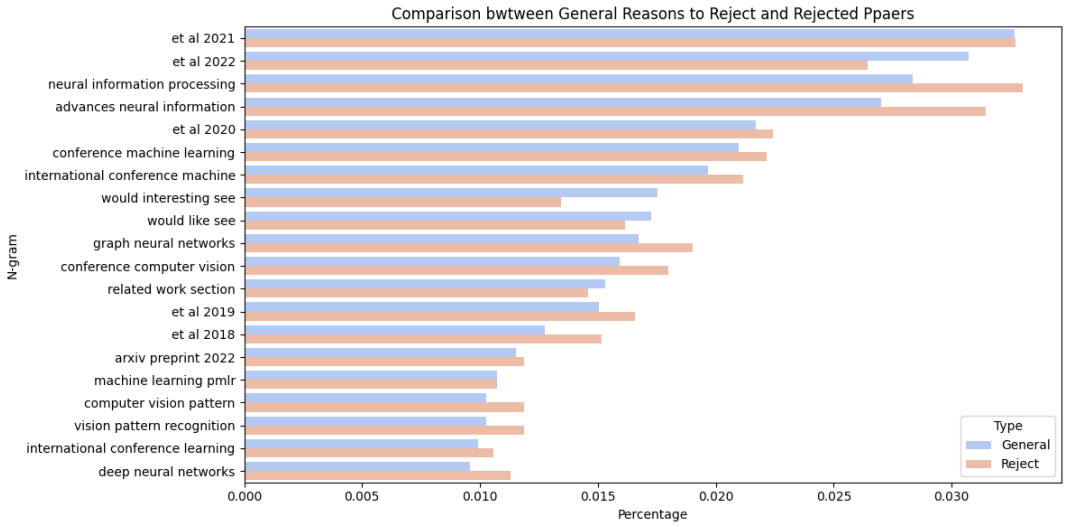


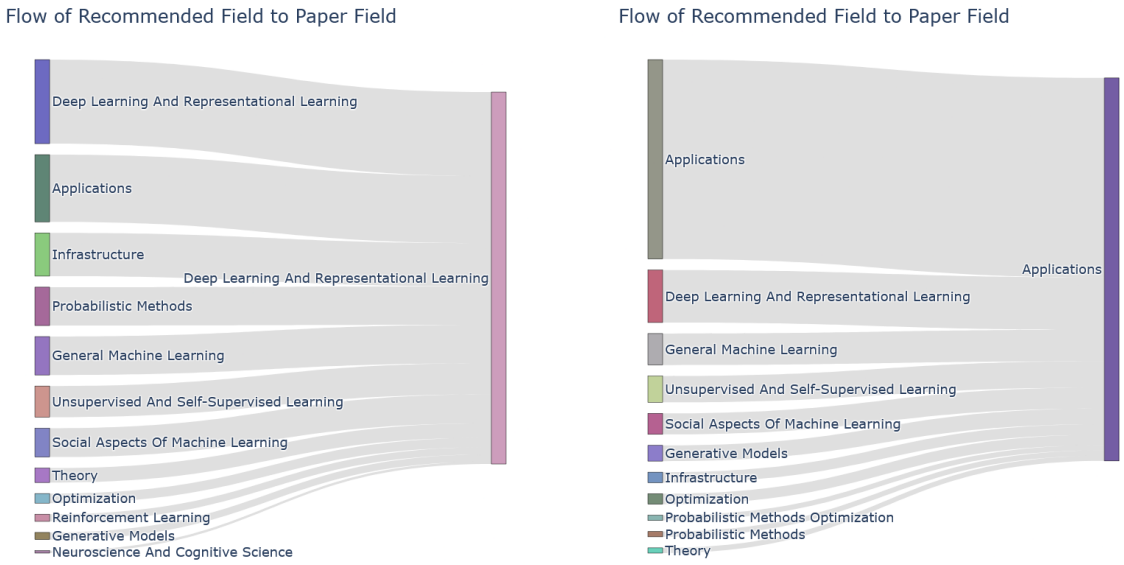
Fig. 23. Most Frequent Trigrams in Reviews of Rejected Papers vs. All the Papers in ICLR 2023.

Results & Discussion. The TF-IDF analysis revealed that reviews for rejected papers are more likely to include references to prior literature, often in the context of criticizing insufficient comparison or lack of citation to recent or foundational work. Phrases such as "et al.", or specific conference names (e.g., "NeurIPS", "ICLR") appear with notably higher weight in rejected reviews. In contrast, accepted papers more frequently include general phrases indicating strengths, such as "well-written", "novel approach", and "strong empirical results". The frequency analysis of n-grams showed similar trends (See Figure 23).

The findings suggest that one of the recurring reasons for rejection is the perception that a submission has not adequately acknowledged or compared itself with related work. Such citation-related criticisms appear to play an important role in peer review decisions. These results support earlier findings in this thesis regarding reviewer frequently suggest extra literature during the review process (*QA1*, 4.2.1) and it is often related to rejection (*QB2*, 4.2.5).

B.3 Additional Graphs for *QB1* (4.2.4)

Figure 24 shows two examples of citation flow in ICLR 2023. For papers fall into the category of "Deep Learning And Representational Learning", the incoming recommendations tend to be diverse and with relatively even distribution, suggesting that the field is attracting attention from various sub-field. For the field of "Application", authors are most likely to be recommended papers from the same category, possibly even similar types of application. However, literature with more theoretical background, especially those with the focus on learning algorithms, also compose as an important source for appliation papers in ICLR 2023.



(a) Deep Learning and Representational Learning

(b) Applications

Fig. 24. Incoming fields to **Deep Learning and Representational Learning** and **Applications** papers in ICLR 2023. The left column of each Sankey diagram shows the fields of study across all the papers that were recommended to the papers identified (right column) by the authors themselves. (QB1, 4.2.4)

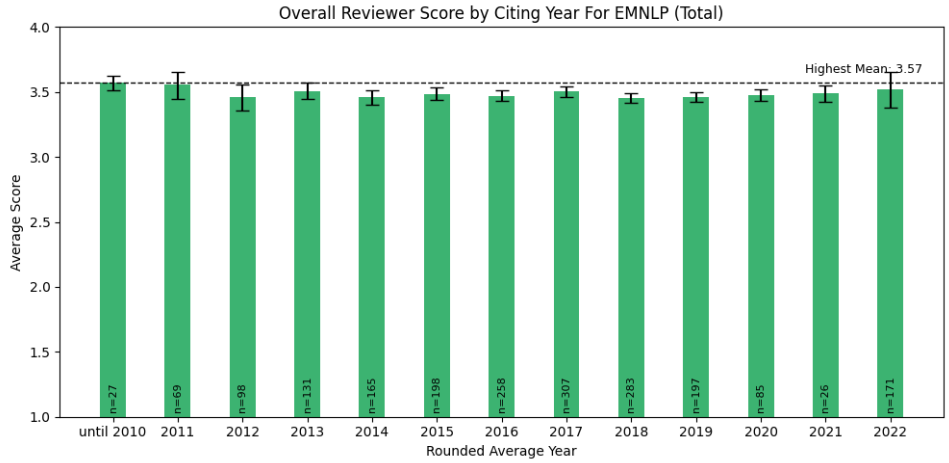


Fig. 25. Total reviewer scores for different avg. citing years in EMNLP 2023 (Total). Average Citing Year = Sum of years of all cited works in a paper / Number of citations in that paper. The result is rounded to the nearest integer. Consecutive years with few data points are aggregated (e.g., "until 2010"). (QC2, 4.2.7).

B.4 Additional Graphs for *QC2* (4.2.7)

Figure 25 and 26 show the average review scores of papers with different average citing years in EMNLP 2023 (Total). It is noteworthy that the highest average scores fall into papers that cite more older literature, and that papers that cite very recent literature also have one of the top average total scores. This finding aligns with the results of NeurIPS 2024 (Accepted) in Section 4.2.7). Since EMNLP 2023 (Total) only has 9 rejected data point, it can also be considered as an estimate of EMNLP 2023 (Accepted).

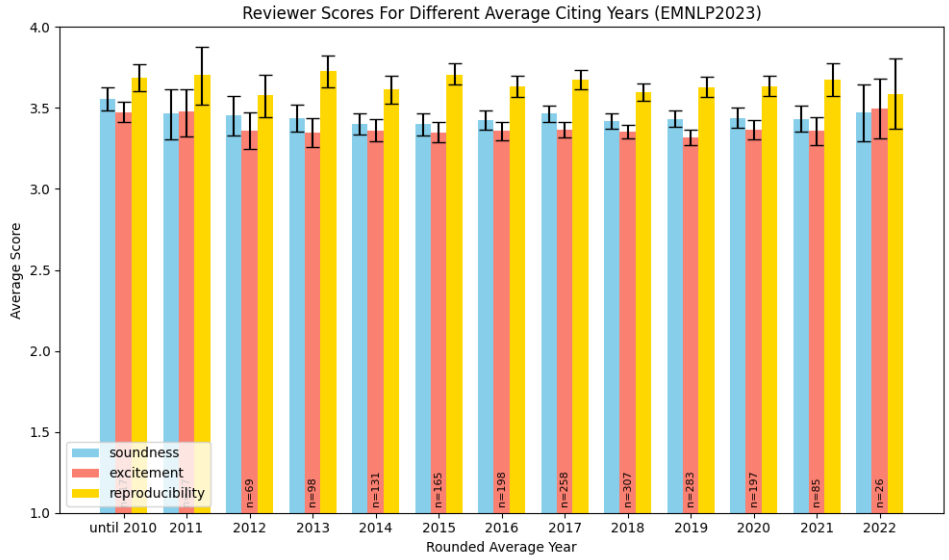


Fig. 26. Categorical reviewer scores for different avg. citing years in EMNLP 2023 (Total). Average Citing Year = Sum of years of all cited works in a paper / Number of citations in that paper. The result is rounded to the nearest integer. Consecutive years with few data points are aggregated (e.g., "until 2010"). (QC2, 4.2.7).

C AI Usage Card

AI Usage Card for Quantifying Biases in Peer Review: Analyzing Reviewer Suggestions in Artificial Intelligence Publications

PROJECT DETAILS	PROJECT NAME Quantifying Biases in Peer Review: Analyzing Reviewer Suggestions in Artificial Intelligence Publications	DOMAIN Master's Thesis	KEY APPLICATION Natural Language Processing
CONTACT(S)	NAME(S) Zhuojing Huang	EMAIL(S) zhuojing.huang@stud.uni-goettingen.de	AFFILIATION(S) Georg-August-Universität Göttingen
MODEL(S)	MODEL NAME(S) ChatGPT Llama	VERSION(S) 4o, 4.5 3.1	
IDEATION	GENERATING IDEAS, OUTLINES, AND WORKFLOWS	IMPROVING EXISTING IDEAS	FINDING GAPS OR COMPARE ASPECTS OF IDEAS
LITERATURE REVIEW	FINDING LITERATURE	FINDING EXAMPLES FROM KNOWN LITERATURE OR ADDING LITERATURE FOR EXISTING STATEMENTS	COMPARING LITERATURE
METHODOLOGY	PROPOSING NEW SOLUTIONS TO PROBLEMS	FINDING ITERATIVE OPTIMIZATIONS	COMPARING RELATED SOLUTIONS
EXPERIMENTS	DESIGNING NEW EXPERIMENTS	EDITING EXISTING EXPERIMENTS	FINDING, COMPARING, AND AGGREGATING RESULTS
WRITING	GENERATING NEW TEXT BASED ON INSTRUCTIONS Llama	ASSISTING IN IMPROVING OWN CONTENT OR PARAPHRASING RELATED WORK ChatGPT	PUTTING OTHER WORKS IN PERSPECTIVE
PRESENTATION	GENERATING NEW ARTIFACTS	IMPROVING THE AESTHETICS OF ARTIFACTS	FINDING RELATIONS BETWEEN OWN OR RELATED ARTIFACTS
CODING	GENERATING NEW CODE BASED ON DESCRIPTIONS OR EXISTING CODE ChatGPT	REFACTORING AND OPTIMIZING EXISTING CODE ChatGPT	COMPARING ASPECTS OF EXISTING CODE
DATA	SUGGESTING NEW SOURCES FOR DATA COLLECTION	CLEANING, NORMALIZING, OR STANDARDIZING DATA Llama	FINDING RELATIONS BETWEEN DATA AND COLLECTION METHODS
ETHICS	WHY DID WE USE AI FOR THIS PROJECT? Efficiency / Speed Scalability	WHAT STEPS ARE WE TAKING TO MITIGATE ERRORS OF AI? Manual validation; Prompts comparison and evaluation	WHAT STEPS ARE WE TAKING TO MINIMIZE THE CHANCE OF HARM OR INAPPROPRIATE USE OF AI? Manual validation

THE CORRESPONDING AUTHORS VERIFY AND AGREE WITH THE MODIFICATIONS OR GENERATIONS OF THEIR USED AI-GENERATED CONTENT