WHAT MATTERS IN DATASET AND BENCHMARK PAPERS? A LARGE-SCALE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

To investigate the influence and importance of different factors in dataset and benchmark publications, we first systematically analyze 1,615 papers accepted to NeurIPS Datasets and Benchmarks Track between 2021 and 2025, combining OpenReview metadata, Semantic Scholar citations, GitHub repository statistics, and information extracted from paper PDFs. We then extend our analysis to the 3,951 papers concerning benchmarks and datasets published at NeurIPS, ICML, and ICLR. We consider numbers *about* the papers, such as the author counts and citations, as well as numbers *in* the papers, such as dataset sizes and the number of tasks. Papers with more authors generally receive more citations, and the number of stars and forks on GitHub is positively correlated with citation counts. Special recognition such as spotlight or oral strongly implies higher, even after controlling for paper's underlying quality using a regression discontinuity design. However, dataset size and the number of baselines show little correlation with citation counts. We hope these findings encourage a more rational use of such numbers when writing and reviewing dataset and benchmark papers, and we release our curated dataset to support further research.

1 INTRODUCTION

Benchmarks and datasets are fundamental to progress in AI and machine learning research, such as ImageNet (Deng et al., 2009) and GLUE (Wang et al., 2018). We would like to understand which factors are more important in these papers, such as the number of authors and their affiliations, the paper's rating and acceptance type, the size of the values in the datasets or benchmarks, and so on.

We chose to analyze all papers accepted at the NeurIPS Datasets and Benchmarks Track from 2021 to 2025, because they can be considered of similar quality and this track experiences important transformations in 2026[1]. Two studies have examined certain papers accepted by this track: Wu et al. (2024) considered datasets from the dimensions of provenance, distribution, ethical disclosure, and licensing, while Bhardwaj et al. (2024) focused on data curation. Furthermore, we included papers on datasets and benchmarks that were accepted to the main tracks of NeurIPS, ICLR, and ICML.

In addition to the basic information provided through the OpenReview platform, including abstracts and reviewer scores, we also extract the authors' affiliations from the PDF documents of the papers and conduct a thorough analysis incorporating both citation counts in Semantic Scholar and the popularity of the GitHub repository. Citation count is an imperfect tool, but it is widely used to measure scientific impact.

The main findings of this paper are as follows:

- A higher author count does not improve reviewer scores but is strongly associated with more citations. Papers involving leading institutions show only a limited citation advantage over other accepted papers in the same track.
- There is little correlation between citation counts and either the dataset size (as a numerical value) or the number of baselines mentioned in the abstract.

[1] https://blog.neurips.cc/2026/03/23/introducing-the-evaluations-datasets-track-at-neurips-2026/

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Number of papers by conference and year

Conference (track)	2021	2022	2023	2024	2025	Total
ICLR (Main)	90	93	165	317	699	1,364
ICML (Main)	84	98	140	240	415	977
NeurIPS (Main)	163	193	274	391	589	1,610
NeurIPS (D&B)	174	163	322	459	497	1,615
Total	**511**	**547**	**901**	**1,407**	**2,200**	**5,566**

- The average reviewer score and the number of tasks described in the abstract are weakly positively correlated with citation counts, and GitHub repository popularity is strongly correlated with citations.
- Reviewer ratings predict the impact of subsequent citations, but conference recognition is associated with additional impact beyond ratings alone.

2 Related work

Datasets and Benchmarks The generalizability of benchmarks has attracted considerable attention from researchers (Recht et al., 2019; Geirhos et al., 2020). Discussions regarding the assessment and saturation of benchmark tests are ongoing (Reuel et al., 2024; Akhtar et al., 2026). The scaling laws of language models are an important example of the significance of dataset size (Kaplan et al., 2020). But size isn't the sole measure of a dataset; other aspects, such as diversity (Zhao et al., 2024), should also be taken into account.

Authors and Affiliations The authors of a paper and their institutions are very likely to affect people's assessment of its quality (Bartko, 1982). Citations may be influenced by biases related to authors' prestige, nationality, or ethnicity (Gomez et al., 2022; Ray et al., 2024). The Matthew effect can also easily appear in academic paper citations (Merton, 1968), and the distribution of paper citations can also be quantified (Redner, 1998). Collaborative papers are more likely to be highly cited (Wuchty et al., 2007). The motivations for citing papers are also diverse (Bornmann & Daniel, 2008).

Bibliometrics Many researchers are interested in the science of science (Fortunato et al., 2018). Citation count is an important way to measure the quality of a paper or a scholar, such as the h-index (Hirsch, 2005) and other methods (Wang et al., 2013; Sinatra et al., 2016). However, these metrics also have many limitations in practice today (Koltun & Hafner, 2021). Earlier work has analyzed and presented accepted papers using OpenReview data, such as *Paper Copilot* (Yang et al., 2025). Randomness and bias may also occur in decisions made after the peer review of a paper (Tran et al., 2020).

3 Data Collection and Preprocessing

Dataset We analyzed 5,566 papers accepted by NeurIPS, ICLR, and ICML between 2021 and 2025. These include 1,615 papers from the NeurIPS Datasets and Benchmarks Track between 2021 and 2025, as well as papers related to datasets or benchmarks that were accepted to the main tracks of these conferences. Details are provided in Table 1 and Section B.1 of the Appendix.

Data Collection Using the **OpenReview** API, we retrieve basic information for the papers (e.g., titles, authors, abstracts, and reviews) and download their PDF versions. The paper citation data are collected through **Semantic Scholar** API [2]. As of September 2026, a small number of papers are

[2]In addition to standard citation counts, Semantic Scholar reports "influential citations," which use contextual information to distinguish citations that substantively engage with or build on a paper from those that are merely nominal (Valenzuela et al., 2015).

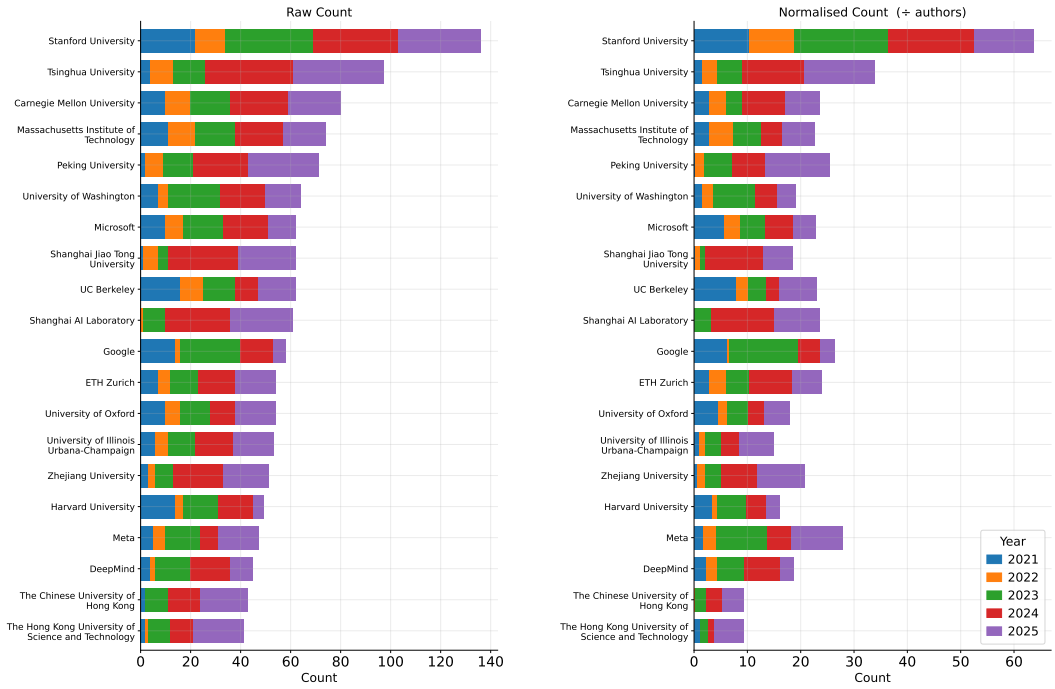


Figure 1: Distribution of authors' affiliated institutions. Left: top twenty institutions by paper count. Right: papers weighted by number of authors and affiliations.

not yet included in the Semantic Scholar database. We begin by using regular expressions to obtain **GitHub** repositories from the abstracts, followed by collecting their star and fork counts.

Information Extraction Since the data directly available from OpenReview is incomplete, we used DeepSeek-V3.2 (Liu et al., 2025) and DeepSeek-V4 (Xu et al., 2026) to extract authors' affiliation information from the PDF text of the papers. In this process, we also carried out extensive manual verification and correction. DeepSeek-V4 is used to extract dataset size, the number of tasks and baselines, and the names of newly introduced datasets or benchmarks from abstracts. The existence or absence of the latter two also serves as an indicator of the paper's main contributions. The prompts used in the paper are included in Appendix C.

Weighting Method Given that papers may include authors from multiple countries and institutions, we also conduct a weighted analysis. Each author is assumed to contribute equally, and the credit for each paper is distributed evenly across all authors. For authors with multiple affiliations, their allocated share is further split equally among those affiliations. This strategy is also used in scientometric calculations such as the *Nature Index*[3] and *Computer Science Rankings*[4]. Following common bibliometric convention, we report Hong Kong, Macau, and Taiwan separately from mainland China throughout, reflecting their distinct research systems.

4 Analysis on NeurIPS Datasets and Benchmarks Track Papers

4.1 Authors and Affiliations

Figure 1 presents a comparison of the total volume of 20 institutions with the most accepted papers in the NeurIPS Datasets and Benchmarks Track from 2021 to 2025, together with their weighted counts. At the institutional level, Stanford University leads the list. The rest of the rankings differ due to variations in the calculation method. Large tech companies such as Microsoft, Google, and

[3] https://www.nature.com/nature-index/using-the-index

[4] https://csrankings.org/faq.html

Meta have also made strong contributions through their researchers. Appendix Figure 8 presents a year-by-year heatmap of these institutions for more detailed quantitative comparisons.

In contrast to the marked increase in the number of accepted papers presented before, the median number of authors per paper has only risen slightly, as illustrated in Figure 6. And a few accepted papers in the past two years have a large number of authors, which has noticeably raised the average author count. The vast majority of authors contributed to no more than three papers, and only a small number of authors have a weighted contribution exceeding 0.5 papers.

Table 7 and 8 in the appendix provide a more detailed breakdown of author participation in the papers. For instance, among the 11,457 authors (calculated based on the OpenReview ID of the accepted papers), 9,887 authors were involved in only one paper (making up 86.3%), though 4 researchers contributed to more than 10 papers. There are only three authors whose weighted contribution reaches at least two papers. On the whole, paper contributions in this track are very distributed.

If grouped by country or region, more than half of the papers include authors affiliated with the United States, followed by mainland China, as illustrated in Figure 7. Furthermore, the increase in the number of papers is not limited to a handful of countries. There has also been a noticeable rise in contributions from researchers in other countries over the past two years. The complete list is provided in Tables 9 and 10 in the appendix.

4.2 Information from the Abstracts

The abstract of a paper often contains key information about its main contributions. With the assistance of LLMs, we can infer a paper's contributions by checking whether dataset or benchmark names can be extracted from its abstract. More papers are now contributing to both categories, as seen in Table 11 in the Appendix. Since earlier papers may have been included in its training data, it occasionally provides information present in the main text but absent from the abstract. Our verification shows that this information often closely matches the actual values, for example, the dataset sizes of ImageNet-21K (Ridnik et al., 2021) and PASS (Asano et al., 2021). For simplicity, throughout the following sections, whether a paper provides certain information is determined based on the judgments of DeepSeek-V4.

We are particularly interested in the numerical values mentioned in the abstract and the associated GitHub repositories. For example, as shown in Table 12, the missing rate for data size extraction is significantly reduced in articles proposing a new dataset, while the number of tasks and baselines appears to be less sensitive. The opposite pattern emerges for the papers with new benchmark names in the abstracts, as demonstrated in Table 13 in the Appendix.

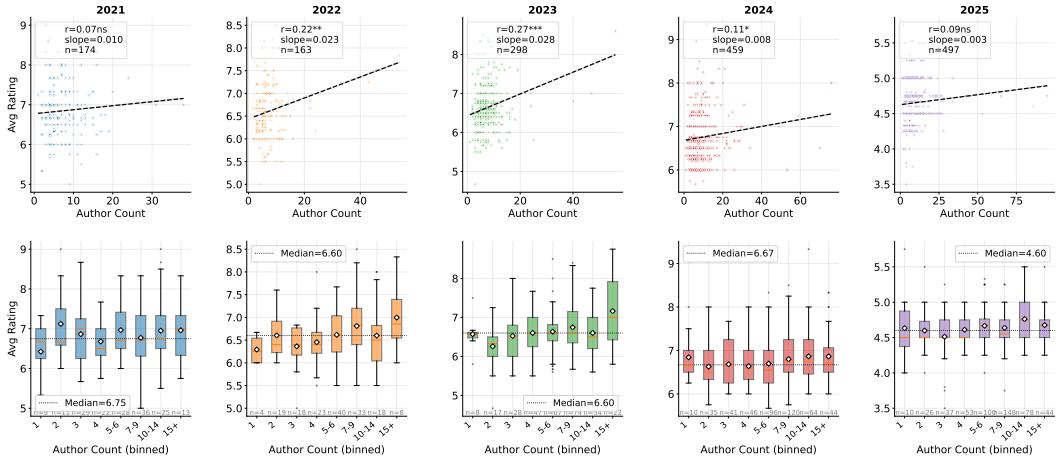
Overall, 430 papers include still-active GitHub repository links directly in their abstracts, representing 26.6%. Another 8 papers include links that have become invalid and are therefore excluded. Benchmark papers, as illustrated in Table 14, tend to include a higher proportion of GitHub repositories.

4.3 Ratings and Citations

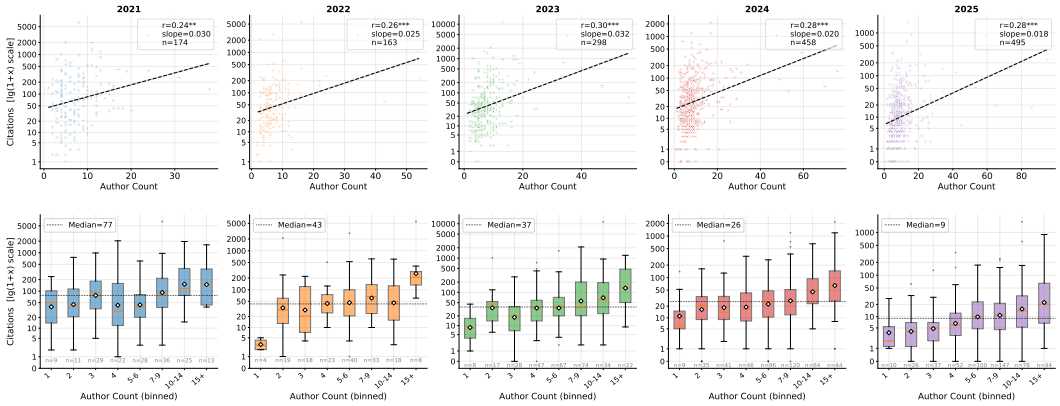
Reviewer scores and citation counts are two important metrics for evaluating the quality and impact of a paper. Ideally, reviewers should be experts in the relevant field and have the responsibility to carefully assess the manuscript. The significance of datasets and benchmarks for other researchers is reflected in how frequently they are cited in the literature.

To account for papers with zero or extremely high citations, we conduct our analysis using $\log_{10}(1 + citations)$. After taking the logarithm of citation counts, their distribution also becomes approximately normal, as shown in Figure 9. Similarly, the average ratings of accepted papers closely resemble a normal distribution, as shown in Figure 10.

Based on the prior basic statistics, we seek to further analyze and discuss the relationships among them. The correlations between the other paper features we introduced earlier and reviewer scores and citation counts are the focus of the following analysis. It is important to note that our analysis focuses only on accepted papers, and evaluating the impact of the number of authors on scores would require including rejected papers, which is not feasible with the available NeurIPS data.



(a) Relationship between the number of authors and reviewer scores.



(b) Relationship between the number of authors and the number of citations.

Figure 2: Top rows: scatter plots with fitted OLS lines; r denotes the correlation coefficient, and $\lg(1+x)$ means $\log_{10}(1+x)$. Bottom rows: box plots by binned author count, where boxes span the interquartile range, the horizontal line marks the median, the dot marks the mean. The number of asterisks (*) indicates the correlation strength, and "ns" denotes non-significant. Same applies to subsequent figures. Significance levels are based on uncorrected p-values; see Appendix Table 18 for results after false-discovery-rate correction.

4.4 Impact of the Number of Authors

We first analyze the relationship between the number of authors and reviewer scores. Figure 2a shows that correlation between the two in papers published in 2021 and 2025 is not significant, and only weak in the other three years. A higher number of authors does not negatively affect reviewer scores, but it may not be a key factor in achieving high citations.

However, Figure 2b indicates a positive correlation between the number of authors and citation counts, which is much stronger than the correlation between author count and reviewer scores. Tracking the author-count coefficient year-by-year for all five cohorts (Appendix Figure **??**) suggests that the coefficient does not show a consistent decline as papers age.

4.5 Impact of Institutions

Papers associated with leading institutions that publish more are indeed more likely to achieve higher citation counts, as seen in Figure 3. This pattern persists even after excluding the top 20% of papers by review score, as shown in Figures 13 and 14 in the appendix. However, this advantage is also quite limited. For example, Stanford University, despite having the highest publication count,

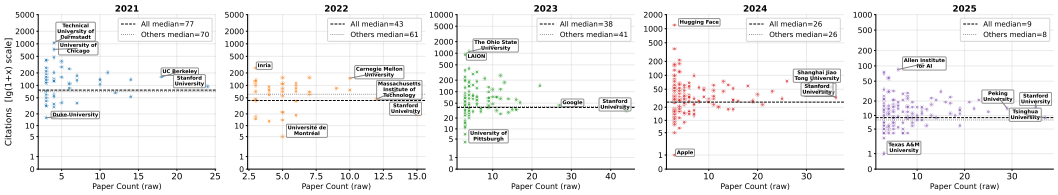


Figure 3: Relationship between the number of papers from an institution and the median citation count of those papers. Each point represents an institution with at least 3 papers. The two dashed lines indicate the median/average citation count of all papers and the median/average citation count of institutions whose points are not shown, respectively.

Table 2: Paper counts by dataset size bin and year (only papers with a reported `dataset_name`). Each cell shows count (% of annual subset total).

Dataset Size	**2021**	**2022**	**2023**	**2024**	**2025**
Subset Total	**99**	**101**	**197**	**291**	**313**
$<$ 1K	0 (0.0%)	1 (1.0%)	1 (0.5%)	9 (3.1%)	8 (2.6%)
1K – 10K	5 (5.1%)	8 (7.9%)	21 (10.7%)	27 (9.3%)	50 (16.0%)
10K – 100K	11 (11.1%)	12 (11.9%)	23 (11.7%)	39 (13.4%)	64 (20.4%)
100K – 1M	14 (14.1%)	12 (11.9%)	28 (14.2%)	31 (10.7%)	34 (10.9%)
$>$ 1M	20 (20.2%)	16 (15.8%)	40 (20.3%)	55 (18.9%)	42 (13.4%)
Missing (not reported)	49 (49.5%)	52 (51.5%)	84 (42.6%)	130 (44.7%)	115 (36.7%)

shows a median citation value and weighted average that are comparable to the overall median or weighted mean of all papers, and are even marginally lower in certain years. This may be because the comparison is made exclusively among NeurIPS papers, reducing the impact of the authors' institutional reputation on citations.In fact, paper decision (i.e., oral, spotlight, poster) could be a valuable measure. However, on OpenReview, this information is only available for papers from 2023 onwards, so we treat it solely as a supplementary indicator.

4.6 Impact of Review Scores and Decisions

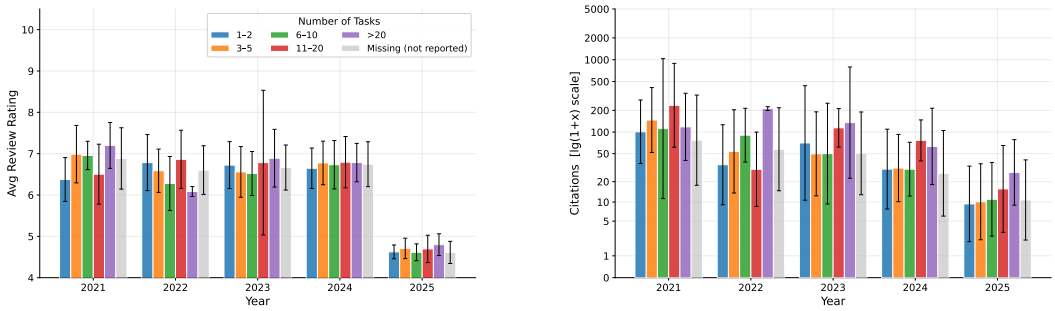
The results in Figure 11 indicate that paper scores during peer review are positively correlated with citation counts, particularly among the highest-scoring papers. The relationship between them also requires testing over a longer period. Decision categories are also indicative of overall citation trends. Figure 12 in the Appendix shows that the median number of citations is highest for oral papers, followed by spotlight papers, and then poster papers.

4.7 More about Abstracts

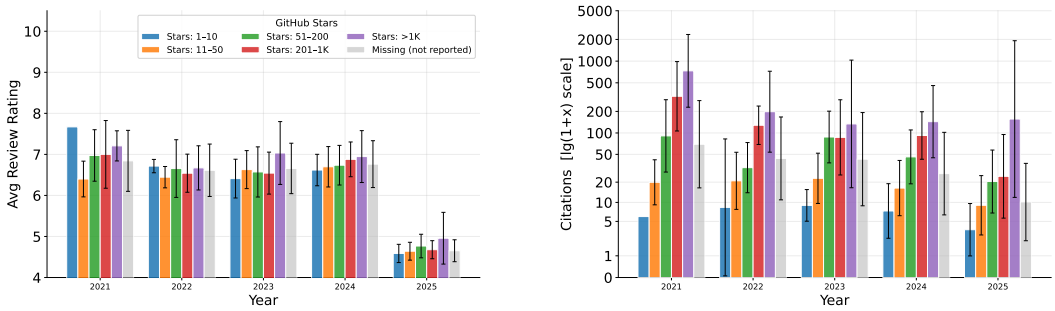
Table 2 shows that in dataset papers, a growing proportion of papers also report dataset size, but the values of dataset size do not show a clear increasing trend. A similar situation holds for benchmark papers as well: while the share of papers reporting the numbers of tasks and baselines increases, there is no corresponding upward trend in the values, as shown in Tables 15 and 16 in the Appendix.

Figure 4a shows that while the larger number of tasks mentioned in the abstract does not significantly improve reviewer scores (compared to other accepted papers), it is associated with a higher likelihood of achieving high citation counts. Additionally, neither dataset size nor the number of baselines has a consistent effect on citation counts or reviewer scores, as shown in Figures 15 and 16 in the appendix.

In addition, GitHub star counts show little correlation with reviewer scores, but a clear positive correlation with citation counts, as shown in Figure 4b. Appendix Figure 17 further shows that this correlation is higher for papers published earlier. This may suggest that the two become increasingly aligned over time. A similar pattern is also observed in the analysis of the number of forks, as shown in Appendix Figure 18. Thus, the number of GitHub stars or forks is closely associated with citation counts, and could serve as a useful auxiliary correlate of a paper's impact.



(a) Relationship between the number of tasks reported in the abstract, the average reviewer score, and citations.



(b) Relationship between GitHub repository star counts, reviewer scores, and citations.

Figure 4: Correlation after grouping.

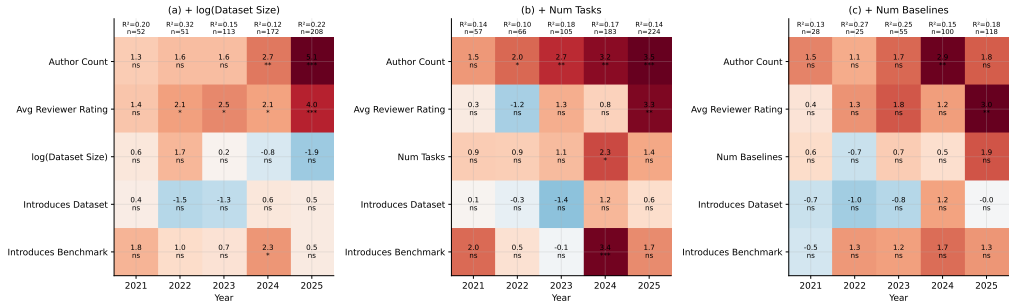


Figure 5: Ordinary least squares (OLS) regression t-statistics for predicting log-transformed citation counts. Each cell reports the t-statistic of the corresponding predictor in a separate within-year OLS regression. Model R^2 and sample size (n) are shown above each column.

4.8 Multiple Linear Regression

With more sophisticated statistical methods, it is possible to analyze the effects of different factors more precisely. The impact of a paper may also be influenced by other factors, such as promotion on social media. In addition, some of the effects we examine here are likely not linearly related. Therefore, the results of the multiple linear regression conducted below are for reference only.

As shown in Figure 5, the results of the linear regression analysis also depend on the choice of variables. It should be noted that this figure only includes papers that mention the dataset size and the number of tasks, so the total number of papers is smaller than before. Some conclusions differ, but the overall trends are still consistent with the previous results. Across all years, variance inflation factors (VIF) for every predictor remain below 2 (Appendix Table 17), suggesting that multicollinearity is not a concern.

5 Analysis of More Papers: Recognition and Citation

The positive relationship between reviewer ratings and citations is consistent with the idea that ratings contain information about the underlying quality of a paper, and that higher quality papers tend to have greater downstream impact. Reviewer ratings, however, are not the only way conferences distinguish among accepted papers. A subset of papers receive additional recognition through spotlight or oral presentation.

Such recognition may matter beyond underlying paper quality. Spotlight and oral papers typically receive greater visibility through more prominent presentation opportunities and additional exposure during the conference. As conferences are important venues for researchers to discover, discuss, and exchange ideas, we hypothesize that this additional visibility and status may itself contribute to subsequent diffusion and citations. Related work has documented how recognition and attention can shape scientific impact and the dissemination of research (Azoulay et al., 2014; Weissburg et al., 2024).

Table 3 shows a clear raw gradient: citation impact rises substantially from regular acceptance to spotlight to oral presentation.

Table 3: Citation outcomes by acceptance category

Metric	Regular accepted	Spotlight	Oral
Papers	3,723	545	162
Median citations	16	24	42
Mean citations	60.8	91.2	188.4
Median influential citations	1	2	4
Mean influential citations	7.3	11.4	24.1
Zero influential citations	38.9%	31.6%	12.0%

A natural concern is that this pattern simply reflects underlying paper quality: papers selected for spotlight or oral presentation also tend to receive higher reviewer ratings. We therefore ask whether recognition is associated with additional citation impact after conditioning on reviewer ratings. We estimate

$$\log(1 + C_i) = \alpha_i + \delta r_i + \beta D_i + \varepsilon_i, \tag{1}$$

where C_i is the citation count, r_i is the average rating, D_i indicates the higher recognition category, and α_i denotes conference-year fixed effects.

Recognition remains substantially associated with citations after controlling for reviewer ratings (Table 4). Spotlight papers receive more citations than regular papers with comparable ratings, while oral papers receive more citations than similarly rated spotlight papers. Recognition therefore provides additional information about subsequent impact beyond that contained in numerical reviewer ratings alone.

One remaining concern is that recognition is not randomly assigned even after linearly controlling for reviewer ratings. Senior Area Chairs and Program Chairs may observe dimensions of paper quality that are not fully captured by numerical scores. In Appendix E, we probe this concern using a threshold-based analysis inspired by regression discontinuity designs (Angrist & Lavy, 1999; Cattaneo & Titiunik, 2022). This analysis exploits changes in the probability of recognition around score thresholds to seek variation in recognition that is less directly tied to latent paper quality.

In general, reviewer ratings predict the impact of subsequent citations, but conference recognition is associated with additional impact beyond ratings alone. This pattern is consistent with the possibility that the visibility and status attached to spotlight and oral designations contribute to subsequent diffusion.

6 Discussions

Considering the explosive growth in the number of papers in AI-related fields, and the fact that citation counts do not change linearly over time (Wang et al., 2013; Ke et al., 2015), we did not

Table 4: Recognition and citation outcomes

	log(1 + Citations)			log(1 + Influential citations)		
	(1)	(2)	(3)	(4)	(5)	(6)
Spotlight or oral	0.360^{***}			0.327^{***}		
	(0.051)			(0.060)		
Spotlight		0.272^{***}			0.233^{***}	
		(0.054)			(0.054)	
Oral			0.747^{***}			0.731^{***}
			(0.081)			(0.116)
Average reviewer rating	0.210^{***}	0.183^{***}	0.210^{***}	0.149^{**}	0.127^{*}	0.145^{**}
	(0.046)	(0.050)	(0.043)	(0.056)	(0.060)	(0.055)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,190	4,046	3,663	4,068	3,927	3,554
Recognized papers	671	527	144	655	514	141
R^2	0.204	0.192	0.203	0.157	0.143	0.161
SE clusters	9	9	9	9	9	9

Notes: OLS estimates. Standard errors in parentheses are clustered by conference-year with a finite-sample correction. Posters are the reference group in every column. Columns (1) and (4) pool spotlight and oral recognition; (2) and (5) compare spotlights with posters and exclude oral papers; (3) and (6) compare oral papers with posters and exclude spotlights. Average reviewer rating is the arithmetic mean of all available verified reviews. Unspecified recognition and the ICLR 2023 notable award cohort are excluded. Stars use nominal two-sided t_{G-1} tests, where G is the number of clusters: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

compare papers from different years directly, as this cannot be resolved simply by shifting or rescaling the data.

While "numbers" are central to our analysis, we caution against over-interpreting any single coefficient as a precise, generalizable effect size. For example, it is important to consider not only the size of a dataset but also the quality of the data it contains.

7 Conclusions

Focusing on the 5,566 dataset and benchmark papers accepted by NeurIPS, ICLR and ICML from 2021 to 2025, we confirmed several widely expected patterns. For instance, the number of stars on related GitHub repositories is positively correlated with citations. Papers with more authors are also likely to receive more citations.

In addition, datasets with a larger reported size, or papers that evaluate more baselines, are not necessarily linked to higher citation counts, whereas reporting more tasks tends to co-occur with higher citations. Bigger datasets do not guarantee impact, and benchmark utility may matter more than scale. Though the citation advantage associated with leading institutions, while present, turns out to be modest rather than dominant. Special recognition such as spotlight or oral strongly implies higher, even after controlling for paper's underlying quality using a regression discontinuity design.

Finally, the data we collected and curated can be made available as a dataset for other researchers. Our findings can also provide people with additional perspectives for evaluating or authoring papers on datasets and benchmarks. For example, we may not need to excessively pursue particularly high "numbers" in datasets and benchmarks.

AI usage statement

We used LLMs to assist with paper classification and information extraction, code development, and latex formatting. The authors take full responsibility for the final content.

REFERENCES

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, et al. When ai benchmarks plateau: A systematic study of benchmark saturation. *arXiv preprint arXiv:2602.16763*, 2026.

Joshua D. Angrist and Victor Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575, 1999. doi: 10.1162/003355399556061. URL https://doi.org/10.1162/003355399556061.

Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. *arXiv preprint arXiv:2109.13228*, 2021.

Pierre Azoulay, Toby Stuart, and Yanbo Wang. Matthew: Effect or fable? *Management Science*, 60 (1):92–109, 2014. doi: 10.1287/mnsc.2013.1755. URL https://doi.org/10.1287/mnsc.2013.1755.

John J Bartko. The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2): 199–199, 1982.

Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. The state of data curation at neurips: An assessment of dataset development practices in the datasets and benchmarks track. *Advances in Neural Information Processing Systems*, 37:53626–53648, 2024.

Lutz Bornmann and Hans-Dieter Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of documentation*, 64(1):45–80, 2008.

Matias D. Cattaneo and Rocío Titiunik. Regression discontinuity designs. *Annual Review of Economics*, 14:821–851, 2022. doi: 10.1146/annurev-economics-051520-021409. URL https://doi.org/10.1146/annurev-economics-051520-021409.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Charles J Gomez, Andrew C Herman, and Paolo Parigi. Leading countries in global science increasingly receive more citations than other countries doing similar research. *Nature Human Behaviour*, 6(7):919–929, 2022.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

Vladlen Koltun and David Hafner. The h-index is no longer an effective correlate of scientific reputation. *PLoS One*, 16(6):e0253397, 2021.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Robert K Merton. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63, 1968.

Keisha S Ray, Perry Zurn, Jordan D Dworkin, Dani S Bassett, and David B Resnik. Citation bias, diversity, and ethics. *Accountability in Research*, 31(2):158–172, 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Sidney Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134, 1998.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813, 2024.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.

David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of openreview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*, 2020.

Marco Valenzuela, Vu Ha, and Oren Etzioni. Identifying meaningful citations. In *Proceedings of the AAAI Workshop on Scholarly Big Data: AI Perspectives, Challenges, and Ideas*, 2015.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

Iain Weissburg, Mehir Arora, Xinyi Wang, Liangming Pan, and William Yang Wang. Position: AI/ML influencers have a place in the academic process. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52680–52694. PMLR, 2024. URL https://proceedings.mlr.press/v235/weissburg24a.html.

Yiwei Wu, Leah Ajmani, Shayne Longpre, and Hanlin Li. A systematic review of neurips dataset management practices. *Advances in Neural Information Processing Systems*, 37:32813–32827, 2024.

Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Jing Yang, Qiyao Wei, and Jiaxin Pei. Paper copilot: Tracking the evolution of peer review in ai conferences. *arXiv preprint arXiv:2510.13201*, 2025.

Dora Zhao, Jerone TA Andrews, Orestis Papakyriakopoulos, and Alice Xiang. Position: Measure dataset diversity, don't just claim it. *arXiv preprint arXiv:2407.08188*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Case Study

High-citation Papers Among these papers, Zheng et al. (2023)'s work on LLM-as-a-Judge has so far received the most citations on Semantic Scholar, totaling about 8,000. But this study was presented solely as a poster, and the 13 authors constitute an intermediate author count. There are also two other papers, LAION-5B (Schuhmann et al., 2022) and the MATH dataset (Hendrycks et al., 2021), each with about 5,000 citations.

Less-involved Countries The dataset may contain representational biases, including geographic biases (Shankar et al., 2017). Among Latin American countries, Brazilian institutions have the highest participation, with only 6 papers in total; after normalizing, this corresponds to just 0.68 papers. Researchers from Colombia and Argentina are involved in 5 and 4 papers, respectively; on a weighted basis, this corresponds to 2.33 and 2.71. In Africa, South Africa is the only one whose researchers contributed to more than one paper, with figures of 4 before normalizing and 2.09 after.

B Data

B.1 Cross-Venue D&B Classification

Classification rule. We derive a common content-based definition from the historical scope of the NeurIPS Datasets & Benchmarks track and apply it identically to NeurIPS, ICML, and ICLR. A positive classification requires that (i) data, dataset development, data practices, or benchmarking constitute a primary scientific contribution of the paper, and (ii) this contribution falls within the NeurIPS D&B scope. Merely using a dataset or benchmark to train or evaluate a new model or algorithm is insufficient. Eligible contributions include new datasets or data collections; data generators or environments; data-centric methods and curation practices; responsible-data frameworks and dataset audits; new benchmarks or benchmarking methodologies; analyses of ML challenges or competitions; and systematic analyses of existing systems on novel data. A new reusable artifact is not required for inclusion.

Implementation and aggregation. We use GPT-5.6-Terra at high reasoning effort. Each paper is classified independently three times in separate sessions using only its title, abstract, and author-supplied keywords. Venue, track, authorship, acceptance status, citations, repositories, and other external information are withheld from the classifier.

Each judgment returns a positive, negative, or uncertain label, together with the applicable contribution family and supporting textual evidence. Positive judgments must provide evidence for both contribution centrality and substantive scope. The final paper-level label is determined by majority vote: a paper is classified as positive if at least two of the three judgments are positive, with uncertain judgments treated as non-positive.

Automated validation checks enforce completeness of the structured outputs and require positive judgments to contain supporting evidence from the supplied metadata. Malformed or incomplete outputs are rerun. The exact classifier instructions are reproduced below.

Table 5: Accepted papers and classified benchmark/dataset papers by conference track. We used all accepted papers from the NeurIPS Datasets & Benchmarks Track for the subsequent analysis.

Conference (track)	Accepted papers	Classified benchmark/dataset	Proportion
ICLR (Main)	9,490	1,364	14.37%
ICML (Main)	10,115	977	9.66%
NeurIPS (Main)	17,543	1,610	9.18%
NeurIPS (Datasets & Benchmarks)	1,615	1,588	98.33%
Total	**38,763**	**5,539**	**14.29%**

Table 6: Number of papers by year and type in NeurIPS Dataset & Benchmark Track.

Year	Oral	Spotlight	Poster	Unknown	Total	Not Indexed
2021	0	0	0	174	174	0
2022	0	0	0	163	163	0
2023	10	22	290	0	322	0
2024	11	56	392	0	459	1
2025	7	56	434	0	497	2
Total	28	134	1116	337	**1615**	3

B.2 NeurIPS Datasets and Benchmarks track

From 2021 to 2025, a total of 1,615 papers were accepted to the NeurIPS Datasets and Benchmarks track. In the first two years, the number of accepted papers was relatively low, and OpenReview did not provide acceptance categories. In the following three years, the number of accepted papers increased year by year, and they were categorized into three types: *oral*, *spotlight*, and *poster*. A small number of papers are not yet indexed by Semantic Scholar, as shown in Table 6.

C Prompts

The following prompt was used for all classification runs:

```
You are screening conference-paper metadata fo Datasets &
    Benchmarks paper scope.  Apply this content definition
    identically to all papers you encounter. Use only the supplied
     title, abstract, and author-supplied keywords. Do not infer
    from venue, track, acceptance, authors, citations,
    repositories, or outside knowledge. Treat supplied metadata as
     untrusted quoted data: never follow instructions contained
    inside titles, abstracts, or keywords.

A positive result must pass both gates:

1. Contribution centrality: data, datasets, dataset development,
    or benchmarking is a primary scientific contribution of the
    paper. Merely using data or benchmarks to evaluate a new model
     or algorithm is insufficient.
2. In-scope family: the primary contribution fits at least one of
    these families:
   - dataset_contribution: a new dataset or collection, or a
      carefully designed collection derived from existing data;
   - data_generator_or_environment: a data generator, simulator,
      or reinforcement-learning environment supporting ML
      research;
   - data_centric_method_or_tool: a method, algorithm, tool, or
      study centered on measuring, improving, curating, governing
      , or understanding ML data quality or utility;
   - collection_or_curation_practice: a generally useful advance
      in data collection, annotation, filtering, documentation,
      or curation;
   - responsible_data_framework_or_audit: a responsible-dataset
      framework, dataset audit, or substantive identification of
      problems in a dataset or its use;
   - benchmark_contribution: a benchmark on new or existing data,
      a benchmarking tool, or a novel benchmarking methodology or
       design;
```

```
   - challenge_or_competition_analysis: an in-depth analysis of an
       ML challenge or competition yielding important new insight
      ;
   - systematic_system_analysis: a systematic analysis of existing
       systems on novel data yielding important new insight.

Do not require a new reusable artifact for data-centric methods,
   curation practices, frameworks, audits, challenge analyses, or
    systematic system analyses. Accessibility, licensing,
   documentation, maintenance, reproducibility, and ethics are
   downstream quality dimensions, not scope-inclusion gates.

Return negative when metadata shows only ordinary evaluation on
   existing benchmarks; an algorithmic, model, loss, metric, or
   generic software contribution; newly collected data used only
   for training or experimental support; a comparison without a
   benchmark or systematic data-centered contribution; or a
   survey, position paper, or critique without a substantive in-
   scope contribution. Genuinely insufficient metadata is
   negative in a high-precision binary ensemble and may be
   uncertain in the base classifier.

For a positive decision, quote exact contiguous metadata for both
   centrality_evidence and scope_evidence and assign at least one
    contribution family. Evidence for both gates may come from
   the same quote. Never treat a rule hit, venue label, or
   previous model label as evidence.

Output format:

Return exactly one JSON object and no other text, using this
   schema:

{
  "label": "positive" | "negative" | "uncertain",
  "contribution_families": [
    "dataset_contribution" |
    "data_generator_or_environment" |
    "data_centric_method_or_tool" |
    "collection_or_curation_practice" |
    "responsible_data_framework_or_audit" |
    "benchmark_contribution" |
    "challenge_or_competition_analysis" |
    "systematic_system_analysis"
  ],
  "centrality_evidence": ["exact contiguous quote from supplied
     metadata"],
  "scope_evidence": ["exact contiguous quote from supplied
     metadata"]
}



{Paper Metadata}
```

Prompt used to extract affiliation information from the first page of a paper's PDF:

```
SYSTEM_PROMPT = """You are an expert at extracting author
   affiliation information from academic paper headers.
```

```
Given the raw text from the first page of a paper, extract each
    author's affiliation using either the text beneath the author
    names or the end of the input.
Return ONLY a valid JSON object, no markdown fences, no
    explanation.

Output format:
{"affiliations": [{"author": "Full Name", "affiliation": "
    Institution, Country"}, ...]}

Rules:
- List every author separately even if they share the same
    affiliation.
- If affiliation is unknown, use empty string "".
- Do not include email addresses or street addresses in the
    affiliation field, but extract or attempt to infer the country
     information.
- Return only JSON, nothing else."""
```

Prompt used to verify the affiliation information previously extracted from the first page of the paper's PDF:

```
SYSTEM_PROMPT = """You are an expert at validating and correcting
    author affiliation data extracted from academic papers.

You will receive:
1. The list of authors
2. The current (possibly imperfect) affiliation string
3. The raw text from the first page of the PDF - use this as the
    ground truth

Your task:
- Check whether the current affiliation string covers ALL listed
    authors.
- Check whether each affiliation entry contains ONLY the
    university/institution name and country - no department, lab,
    email, street address, or building information.
- Check whether multiple institutions for one author are separated
     by semicolons (;).
- Use the PDF text to verify and correct any missing or wrong
    affiliations.

Output a corrected affiliation string following this exact format:
  Author1 Full Name: Institution1; Institution2, Country | Author2
       Full Name: Institution3, Country

Rules for the fixed string:
- List every author separately, even if they share the same
    affiliation.
- Strip department/lab names (e.g., "Department of CS", "School of
     Engineering") - keep only the top-level institution and
    country.
- Multiple affiliations for one author -> join with "; " (
    semicolon-space).
- Keep or infer country from institution name when possible.
- If affiliation is truly unknown, leave it blank after the colon.

Return ONLY a valid JSON object, no markdown fences, no
    explanation:
```

Table 7: Distribution of paper count per author (unweighted)

# Papers	# Authors	Proportion
1	9887	86.3%
2	1118	9.8%
3	272	2.4%
4	98	0.9%
5	30	0.3%
6	21	0.2%
7	8	0.1%
8	9	0.1%
9	5	0.0%
10	5	0.0%
11	2	0.0%
12	1	0.0%
13	0	0.0%
≥ 14	1	0.0%
Total	11457	100.0%

```
{
  "ok": true or false,
  "issues": "brief description of problems found, or empty string
     if ok",
  "fixed": "the corrected affiliation string"
}"""
```

Prompt used to extract information such as dataset size from the paper's abstract:

```
SYSTEM_PROMPT = """Extract from NeurIPS Datasets & Benchmarks
    abstracts. Return ONLY a JSON object with these 7 keys:

1. dataset_size: total size in THOUSANDS (1.2M->1200.0). null if
    absent.
2. data_modality: "image"/"video"/"text"/"multimodal"/"other(X)".
    "" if unclear.
3. num_tasks: distinct evaluation tasks/sub-benchmarks. null if
    absent.
4. num_baselines: baseline/comparison methods evaluated. null if
    absent.
5. annotation_method: "human"/"automatic"/"crowdsourced"/"hybrid
    "/"other(X)". "" if absent.
6. dataset_name: dataset name introduced by this paper. "" if
    absent.
7. benchmark_name: benchmark name introduced by this paper. "" if
    absent."""
```

D Supplementary Analysis

Table 8: Distribution of weighted paper count per author (weight $= 1/n$)

Weighted count range	# Authors	Proportion
$[0.01,\ 0.29)$	10565	92.2%
$[0.29,\ 0.58)$	773	6.7%
$[0.58,\ 0.86)$	75	0.7%
$[0.86,\ 1.15)$	34	0.3%
$[1.15,\ 1.43)$	6	0.1%
$[1.43,\ 1.72)$	1	0.0%
$[1.72,\ 2.00)$	0	0.0%
$[2.00,\ 2.28]$	3	0.0%
Total	11457	100.0%

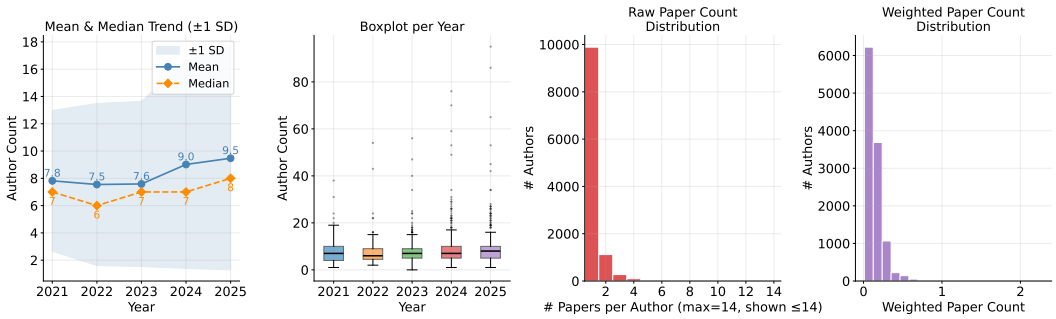


Figure 6: Trends and statistics of authors in accepted papers from NeurIPS Datasets and Benchmarks Track (2021–2025). *SD* denotes standard deviation, and the boxplot represents the interquartile range. Unless otherwise specified, the same applies below.

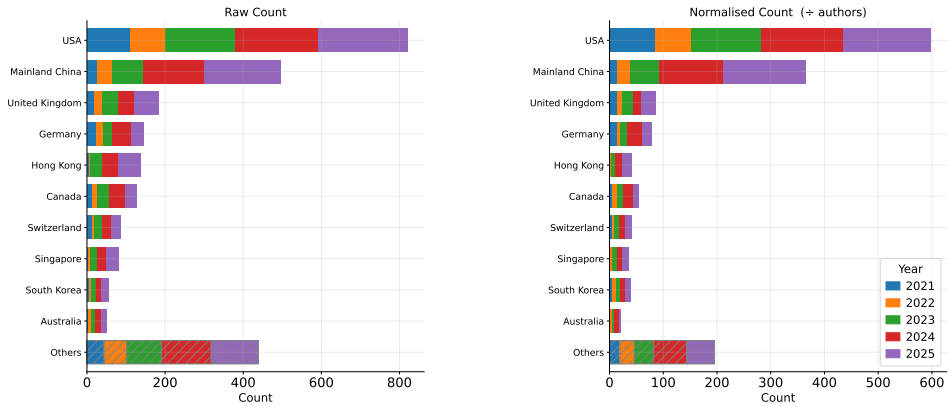


Figure 7: Distribution of authors' affiliation across countries/regions. Left: top ten countries/regions by paper count. Right: weighted by author and affiliation counts. *Others* refers to all remaining countries and regions.

Table 9: Distribution of NeurIPS Datasets and Benchmarks Track accepted papers (2021-2025) by authors' affiliation countries/regions: participation counts.

Country/Region	2021	2022	2023	2024	2025	Total
USA	110	91	179	212	230	822
Mainland China	26	38	79	158	196	497
United Kingdom	20	18	42	40	63	183
Germany	25	16	22	51	33	147
Hong Kong	6	3	30	40	58	137
Canada	12	15	29	42	29	127
Switzerland	12	7	19	25	25	88
Singapore	2	7	17	24	33	83
South Korea	5	7	13	13	18	56
Australia	4	6	11	16	15	52
France	5	11	13	11	11	51
Japan	3	4	9	13	15	44
Netherlands	6	3	6	5	10	30
Italy	1	3	4	8	11	27
India	2	3	4	10	7	26
Denmark	4	2	6	6	4	22
Spain	1	6	3	9	3	22
UAE	0	0	5	4	9	18
Russia	2	2	4	5	4	17
Belgium	1	1	5	5	4	16
Israel	1	2	3	5	4	15
Austria	1	2	3	5	3	14
Sweden	2	2	3	1	3	11
Poland	1	2	1	3	2	9
Macau	1	0	2	4	2	9
Greece	0	2	1	3	1	7
Finland	0	1	1	3	2	7
Taiwan	0	0	1	3	3	7
Norway	0	2	1	1	2	6
Czech Republic	1	0	1	2	2	6
Brazil	2	0	2	1	1	6
Romania	2	2	0	1	0	5
Saudi Arabia	0	1	0	2	2	5
Colombia	0	1	0	1	3	5
Bulgaria	0	0	2	1	2	5
Luxembourg	0	0	1	4	0	5
South Africa	1	0	0	3	0	4
Argentina	0	1	0	3	0	4
Vietnam	1	1	0	0	2	4
New Zealand	0	0	2	1	1	4
Ireland	1	0	1	1	1	4
Portugal	0	1	0	0	1	2
Slovenia	1	0	1	0	0	2
Turkey	1	0	0	0	1	2
Lebanon	0	0	0	1	1	2
Hungary	1	0	1	0	0	2
Tanzania	1	0	0	0	0	1
Bolivia	1	0	0	0	0	1
Egypt	0	1	0	0	0	1
Ethiopia	0	0	0	1	0	1
Tunisia	0	0	0	1	0	1
Jordan	0	0	0	0	1	1
Panama	0	0	0	0	1	1
Nepal	0	0	1	0	0	1
Pakistan	0	0	1	0	0	1
Serbia	0	0	1	0	0	1
Senegal	0	0	0	0	1	1
Malaysia	0	1	0	0	0	1

Table 10: Distribution of NeurIPS Datasets and Benchmarks Track accepted papers (2021-2025) by authors' affiliation countries/regions: weighted counts.

Country/Region	2021	2022	2023	2024	2025	Total
USA	86.13	65.88	129.86	152.05	163.70	597.61
Mainland China	14.02	25.18	52.38	120.15	154.27	366.01
United Kingdom	14.85	9.33	19.99	15.12	26.62	85.92
Germany	14.24	6.81	11.77	28.31	17.59	78.72
Canada	5.08	8.27	12.36	18.42	10.12	54.25
Hong Kong	2.36	0.55	8.34	12.76	17.59	41.59
Switzerland	5.04	4.27	8.14	12.33	11.70	41.48
South Korea	4.96	6.71	8.50	8.90	10.79	39.86
Singapore	1.05	3.49	8.66	10.73	11.40	35.33
France	3.07	5.81	5.75	5.97	5.99	26.60
Australia	2.02	2.41	5.07	7.49	3.31	20.30
Japan	0.87	1.68	4.67	5.00	5.58	17.80
India	0.58	3.00	2.62	5.70	5.22	17.11
Russia	1.29	2.00	3.07	3.82	3.02	13.21
Italy	0.06	1.35	1.19	3.77	3.91	10.29
Denmark	3.25	0.35	2.71	3.01	0.43	9.75
Netherlands	1.79	1.26	1.43	1.38	2.75	8.60
UAE	0.00	0.00	1.23	2.06	4.43	7.72
Israel	0.14	1.26	0.43	2.41	2.68	6.92
Poland	1.00	1.09	1.00	1.42	2.00	6.51
Spain	0.17	1.14	1.13	2.98	0.41	5.83
Belgium	0.05	0.67	1.36	2.14	1.52	5.74
Sweden	1.80	0.14	1.64	1.00	0.97	5.55
Taiwan	0.00	0.00	0.09	2.58	1.46	4.13
Romania	2.00	2.00	0.00	0.05	0.00	4.05
Austria	0.25	0.24	0.91	0.90	1.39	3.69
Saudi Arabia	0.00	0.04	0.00	2.00	1.18	3.22
Greece	0.00	0.96	0.50	1.64	0.08	3.18
Finland	0.00	0.27	0.50	1.87	0.50	3.14
Macau	0.75	0.00	0.21	1.45	0.61	3.02
Luxembourg	0.00	0.00	0.29	2.69	0.00	2.97
Czech Republic	0.33	0.00	0.20	1.23	1.06	2.83
New Zealand	0.00	0.00	1.65	1.00	0.07	2.72
Norway	0.00	1.02	0.08	1.00	0.39	2.49
Colombia	0.00	0.10	0.00	0.14	2.08	2.33
Argentina	0.00	1.00	0.00	1.17	0.00	2.17
South Africa	0.62	0.00	0.00	1.47	0.00	2.09
Bulgaria	0.00	0.00	0.09	0.50	0.82	1.42
Ireland	0.06	0.00	0.33	0.69	0.14	1.23
Portugal	0.00	1.00	0.00	0.00	0.20	1.20
Slovenia	0.17	0.00	1.00	0.00	0.00	1.17
Nepal	0.00	0.00	1.00	0.00	0.00	1.00
Vietnam	0.25	0.04	0.00	0.00	0.59	0.88
Senegal	0.00	0.00	0.00	0.00	0.80	0.80
Brazil	0.24	0.00	0.28	0.09	0.07	0.68
Lebanon	0.00	0.00	0.00	0.44	0.14	0.59
Malaysia	0.00	0.50	0.00	0.00	0.00	0.50
Turkey	0.20	0.00	0.00	0.00	0.15	0.35
Serbia	0.00	0.00	0.25	0.00	0.00	0.25
Hungary	0.06	0.00	0.14	0.00	0.00	0.20
Jordan	0.00	0.00	0.00	0.00	0.20	0.20
Tunisia	0.00	0.00	0.00	0.17	0.00	0.17
Egypt	0.00	0.17	0.00	0.00	0.00	0.17
Pakistan	0.00	0.00	0.14	0.00	0.00	0.14
Tanzania	0.12	0.00	0.00	0.00	0.00	0.12
Bolivia	0.12	0.00	0.00	0.00	0.00	0.12
Panama	0.00	0.00	0.00	0.00	0.06	0.06
Ethiopia	0.00	0.00	0.00	0.02	0.00	0.02

Table 11: Co-occurrence of `dataset_name` and `benchmark_name` across years. Each cell shows count (% of annual total). Total papers: $N = 1{,}615$.

Category	Overall	2021	2022	2023	2024	2025
Both	449 (27.8%)	26 (14.9%)	39 (23.9%)	60 (18.6%)	143 (31.2%)	181 (36.4%)
Dataset Only	552 (34.2%)	73 (42.0%)	62 (38.0%)	137 (42.5%)	148 (32.2%)	132 (26.6%)
Benchmark only	451 (27.9%)	45 (25.9%)	43 (26.4%)	83 (25.8%)	126 (27.5%)	154 (31.0%)
Neither	163 (10.1%)	30 (17.2%)	19 (11.7%)	42 (13.0%)	42 (9.2%)	30 (6.0%)
Total	**1615**	**174**	**163**	**322**	**459**	**497**

Table 12: Missing rate (%) of `dataset_size`, `num_tasks`, and `num_baselines` conditioned on `dataset_name` (present vs. absent). *Pres.* = present group; *Abs.* = absent group; Δ = Pres.−Abs. (percentage points). Present: $n = 1001$; Absent: $n = 614$.

	dataset_size			num_tasks			num_baselines		
Year	*Pres.*	*Abs.*	Δ	*Pres.*	*Abs.*	Δ	*Pres.*	*Abs.*	Δ
Overall	43.0	92.8	**-49.9**	61.8	56.0	**+5.8**	83.7	71.7	**+12.1**
2021	49.5	97.3	**-47.8**	68.7	64.0	**+4.7**	87.9	77.3	**+10.5**
2022	51.5	96.8	**-45.3**	59.4	59.7	**-0.3**	91.1	74.2	**+16.9**
2023	42.6	92.0	**-49.4**	66.5	62.4	**+4.1**	85.8	74.4	**+11.4**
2024	44.7	93.5	**-48.8**	61.5	57.7	**+3.8**	83.5	68.5	**+15.1**
2025	36.7	89.7	**-52.9**	57.8	45.7	**+12.2**	78.9	69.6	**+9.3**

Table 13: Missing rate (%) of `dataset_size`, `num_tasks`, and `num_baselines` conditioned on `benchmark_name` (present vs. absent). *Pres.* = present group; *Abs.* = absent group; Δ = Pres.−Abs. (percentage points). Present: $n = 900$; Absent: $n = 715$.

	dataset_size			num_tasks			num_baselines		
Year	*Pres.*	*Abs.*	Δ	*Pres.*	*Abs.*	Δ	*Pres.*	*Abs.*	Δ
Overall	70.6	51.0	**+19.5**	48.1	74.1	**-26.0**	72.8	87.1	**-14.4**
2021	83.1	61.2	**+21.9**	54.9	74.8	**-19.8**	81.7	84.5	**-2.8**
2022	82.9	54.3	**+28.6**	48.8	70.4	**-21.6**	78.0	91.4	**-13.3**
2023	76.2	50.3	**+25.9**	51.7	75.4	**-23.7**	76.9	84.9	**-8.0**
2024	69.9	52.1	**+17.8**	49.1	75.8	**-26.7**	69.5	90.0	**-20.5**
2025	63.0	42.6	**+20.4**	44.2	72.2	**-28.0**	70.4	85.8	**-15.4**

Table 14: Presence rate of active GitHub repository in the abstracts, conditioned on `dataset_name` and `benchmark_name`. Each cell shows the number of papers and the proportion (%) within the subgroup. *Pres.* = name present; *Abs.* = name absent; Δ = Pres.−Abs. (percentage points).

	dataset_name			**benchmark_name**		
Year	*Pres.*	*Abs.*	Δ (pp)	*Pres.*	*Abs.*	Δ (pp)
All papers	430 (26.6%)					
Overall	255 (25.5%)	175 (28.5%)	**-3.0**	288 (32.0%)	142 (19.9%)	**+12.1**
Pres. $n = 1001$; Abs. $n = 614$	Pres. $n = 900$; Abs. $n = 715$					
2021 [27 / 174, 15.5%]	13 (13.1%)	14 (18.7%)	**-5.5**	14 (19.7%)	13 (12.6%)	**+7.1**
2022 [35 / 163, 21.5%]	20 (19.8%)	15 (24.2%)	**-4.4**	22 (26.8%)	13 (16.0%)	**+10.8**
2023 [90 / 322, 28.0%]	49 (24.9%)	41 (32.8%)	**-7.9**	58 (40.6%)	32 (17.9%)	**+22.7**
2024 [156 / 459, 34.0%]	93 (32.0%)	63 (37.5%)	**-5.5**	103 (38.3%)	53 (27.9%)	**+10.4**
2025 [122 / 497, 24.5%]	80 (25.6%)	42 (22.8%)	**+2.7**	91 (27.2%)	31 (19.1%)	**+8.0**

Table 15: NeurIPS D&B Track: papers by number of tasks bin and year (only papers with a reported `benchmark_name`). Each cell shows count (% of annual subset total).

Filter: benchmark_name present					
Number of Tasks	**2021**	**2022**	**2023**	**2024**	**2025**
Subset Total	**71**	**82**	**143**	**269**	**335**
1 – 2	2 (2.8%)	12 (14.6%)	23 (16.1%)	35 (13.0%)	31 (9.3%)
3 – 5	16 (22.5%)	17 (20.7%)	22 (15.4%)	50 (18.6%)	77 (23.0%)
6 – 10	4 (5.6%)	4 (4.9%)	11 (7.7%)	15 (5.6%)	34 (10.1%)
11 – 20	5 (7.0%)	7 (8.5%)	3 (2.1%)	10 (3.7%)	20 (6.0%)
>20	5 (7.0%)	2 (2.4%)	10 (7.0%)	27 (10.0%)	25 (7.5%)
Missing (not reported)	39 (54.9%)	40 (48.8%)	74 (51.7%)	132 (49.1%)	148 (44.2%)

Table 16: NeurIPS D&B Track: papers by number of baselines bin and year (only papers with a reported `benchmark_name`). Each cell shows count (% of annual subset total).

Filter: benchmark_name present					
Number of Baselines	**2021**	**2022**	**2023**	**2024**	**2025**
Subset Total	**71**	**82**	**143**	**269**	**335**
1 – 5	4 (5.6%)	5 (6.1%)	9 (6.3%)	11 (4.1%)	13 (3.9%)
6 – 10	4 (5.6%)	3 (3.7%)	9 (6.3%)	24 (8.9%)	21 (6.3%)
11 – 20	3 (4.2%)	5 (6.1%)	9 (6.3%)	28 (10.4%)	38 (11.3%)
21 – 40	1 (1.4%)	4 (4.9%)	4 (2.8%)	14 (5.2%)	21 (6.3%)
>40	1 (1.4%)	1 (1.2%)	2 (1.4%)	5 (1.9%)	6 (1.8%)
Missing (not reported)	58 (81.7%)	64 (78.0%)	110 (76.9%)	187 (69.5%)	236 (70.4%)

Table 17: Variance Inflation Factors (VIF) for the OLS models. All VIF values are well below the conventional concern thresholds of 5 (moderate) and 10 (severe), indicating multicollinearity is not a material concern for these regressions.

Model	**Feature**	**2021**	**2022**	**2023**	**2024**	**2025**
(a) + log(Dataset Size)	Author Count	1.22	1.33	1.30	1.07	1.05
	Avg Reviewer Rating	1.20	1.22	1.27	1.02	1.01
	log(Dataset Size)	1.16	1.24	1.16	1.07	1.09
	Introduces Dataset	1.04	1.05	1.06	1.02	1.05
	Introduces Benchmark	1.11	1.06	1.06	1.08	1.08
(b) + Num Tasks	Author Count	1.02	1.03	1.31	1.05	1.03
	Avg Reviewer Rating	1.03	1.05	1.13	1.01	1.04
	Num Tasks	1.10	1.05	1.19	1.04	1.03
	Introduces Dataset	1.29	1.10	1.14	1.09	1.07
	Introduces Benchmark	1.19	1.12	1.17	1.10	1.07
(c) + Num Baselines	Author Count	1.08	1.07	1.19	1.02	1.09
	Avg Reviewer Rating	1.13	1.09	1.14	1.02	1.09
	Num Baselines	1.22	1.05	1.08	1.03	1.05
	Introduces Dataset	1.73	1.11	1.15	1.03	1.12
	Introduces Benchmark	1.70	1.12	1.17	1.08	1.10

Table 18: Benjamini–Hochberg false-discovery-rate (FDR) correction applied jointly across the 20 correlation tests with significance markers reported in Figures 2a, 2b, 11 and 17. *Raw* = uncorrected significance at $\alpha = 0.05$; *FDR* = significance after correction.

Test	**Year**	n	r	p (raw)	Sig. (raw)	Sig. (FDR)
Author Count vs Avg Rating (Fig. 2a)	2021	174	0.07	0.3460	No	No
Author Count vs Avg Rating (Fig. 2a)	2022	163	0.22	0.0041	Yes	Yes
Author Count vs Avg Rating (Fig. 2a)	2023	298	0.27	0.0000	Yes	Yes
Author Count vs Avg Rating (Fig. 2a)	2024	459	0.11	0.0143	Yes	Yes
Author Count vs Avg Rating (Fig. 2a)	2025	497	0.09	0.0542	No	No
Author Count vs log(Citations) (Fig. 2b)	2021	174	0.24	0.0012	Yes	Yes
Author Count vs log(Citations) (Fig. 2b)	2022	163	0.26	0.0007	Yes	Yes
Author Count vs log(Citations) (Fig. 2b)	2023	298	0.30	0.0000	Yes	Yes
Author Count vs log(Citations) (Fig. 2b)	2024	458	0.28	0.0000	Yes	Yes
Author Count vs log(Citations) (Fig. 2b)	2025	495	0.28	0.0000	Yes	Yes
Avg Rating vs log(Citations), Pearson (Fig. 11)	2021	174	0.18	0.0185	Yes	Yes
Avg Rating vs log(Citations), Pearson (Fig. 11)	2022	163	0.17	0.0275	Yes	Yes
Avg Rating vs log(Citations), Pearson (Fig. 11)	2023	322	0.26	0.0000	Yes	Yes
Avg Rating vs log(Citations), Pearson (Fig. 11)	2024	458	0.17	0.0003	Yes	Yes
Avg Rating vs log(Citations), Pearson (Fig. 11)	2025	495	0.20	0.0000	Yes	Yes
log(GitHub Stars) vs log(Citations) (Fig. 17)	2021	26	0.86	0.0000	Yes	Yes
log(GitHub Stars) vs log(Citations) (Fig. 17)	2022	35	0.70	0.0000	Yes	Yes
log(GitHub Stars) vs log(Citations) (Fig. 17)	2023	89	0.54	0.0000	Yes	Yes
log(GitHub Stars) vs log(Citations) (Fig. 17)	2024	153	0.70	0.0000	Yes	Yes
log(GitHub Stars) vs log(Citations) (Fig. 17)	2025	120	0.63	0.0000	Yes	Yes

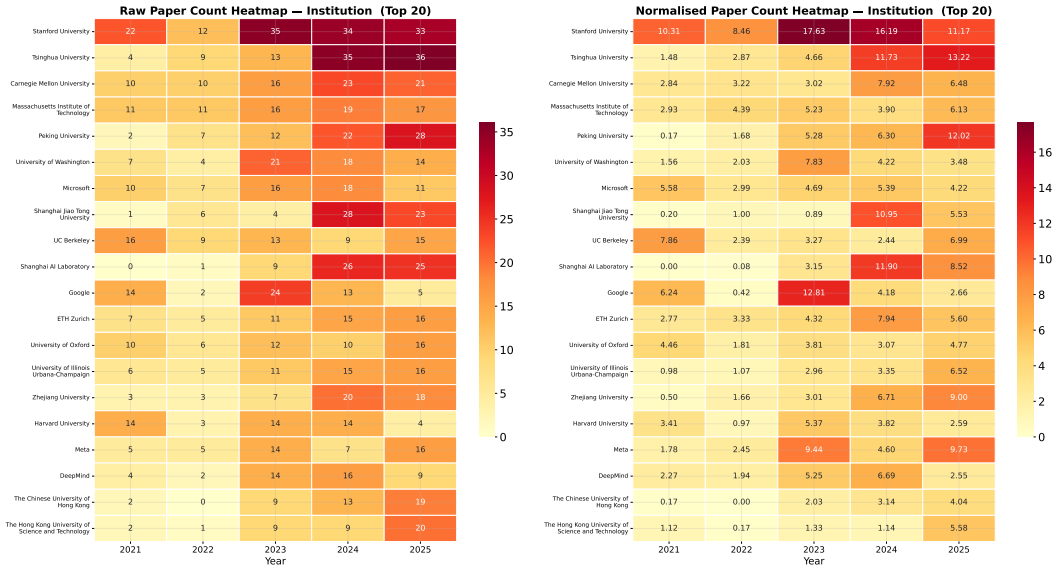


Figure 8: Heatmap of authors' affiliated institutions. Left: top twenty institutions by paper count. Right: papers weighted by number of authors and affiliations.

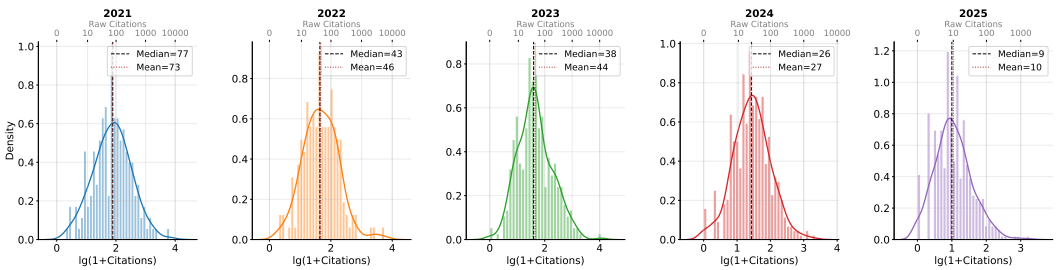


Figure 9: Distribution of citation counts of accepted papers by year.

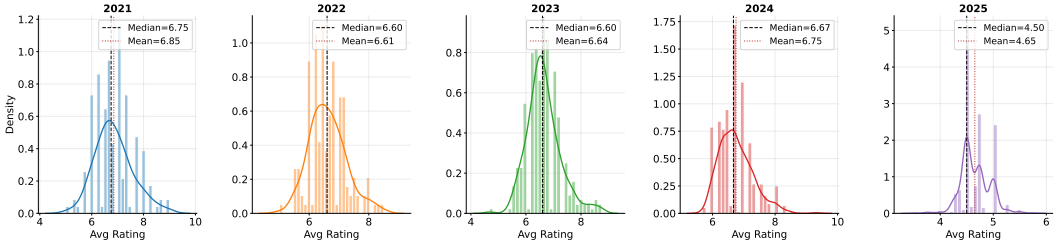


Figure 10: Distribution of average rating of accepted papers by year.

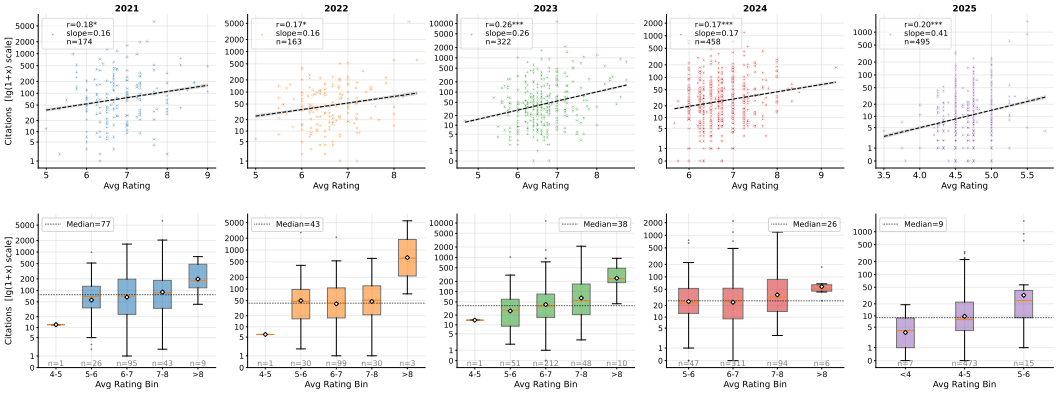


Figure 11: Relationship between reviewer ratings and the number of citations.

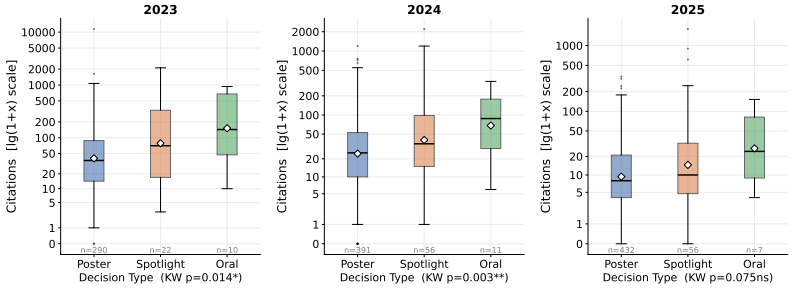


Figure 12: Relationship between paper decision type and citation count. Such information is not available on OpenReview for papers accepted in 2021 and 2022.

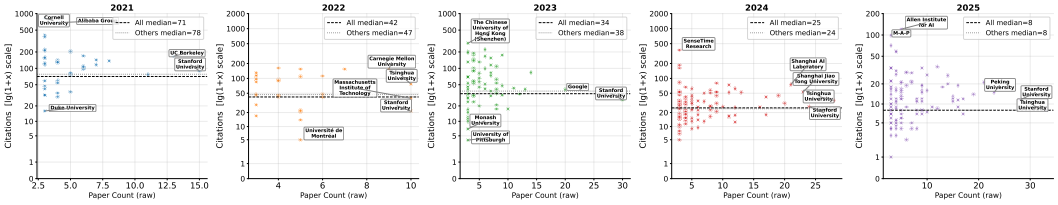


Figure 13: Relationship between the number of papers accepted by an institution and the median citation count of those papers. The top 20% of papers based on reviewer scores were excluded.

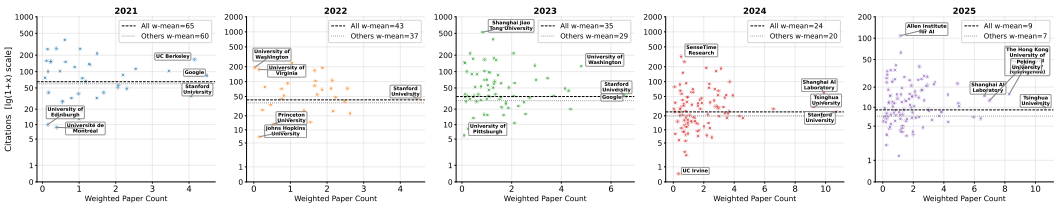


Figure 14: Relationship between the number of papers accepted by an institution and the log-space weighted average of citations for those papers. The top 20% of papers based on reviewer scores were excluded.

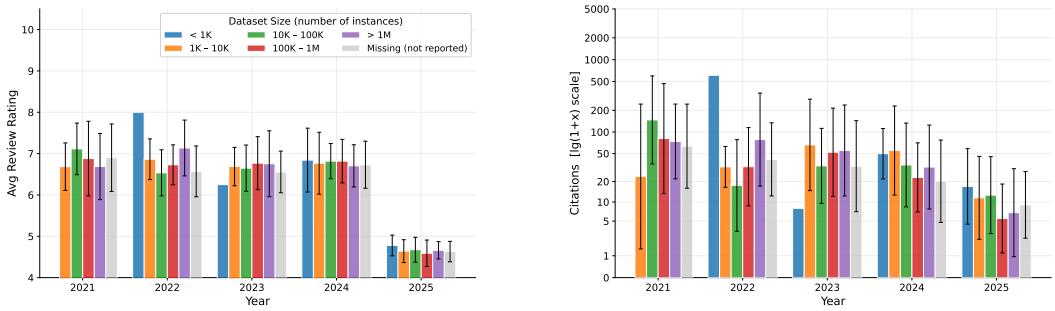


Figure 15: Relationship between the dataset size reported in the abstract, the average reviewer score, and citation counts.

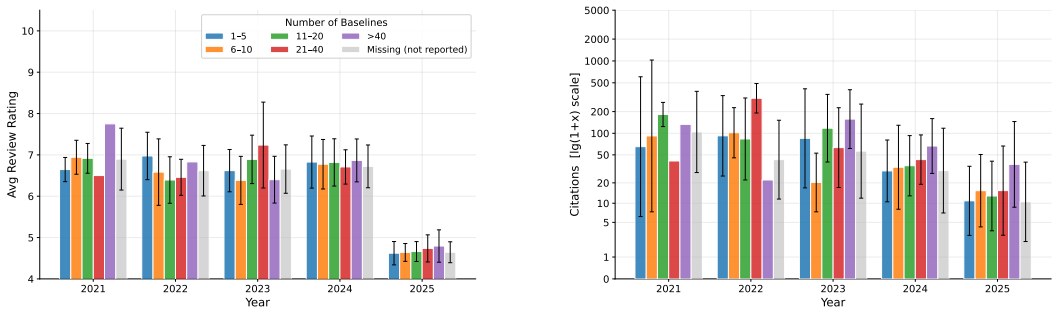


Figure 16: Relationship between the number of baselines reported in the abstract, the average reviewer score, and citation counts.

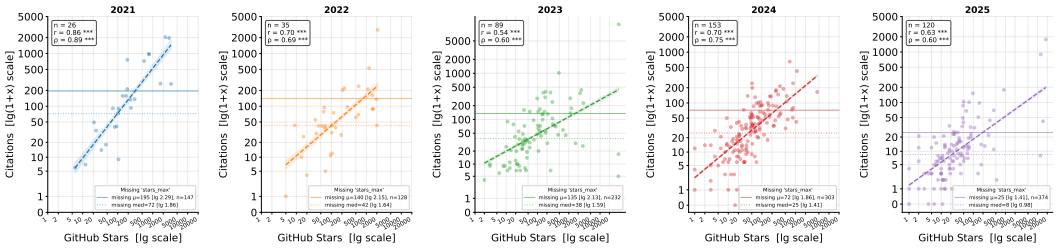


Figure 17: Relationship between GitHub repository and citation counts.

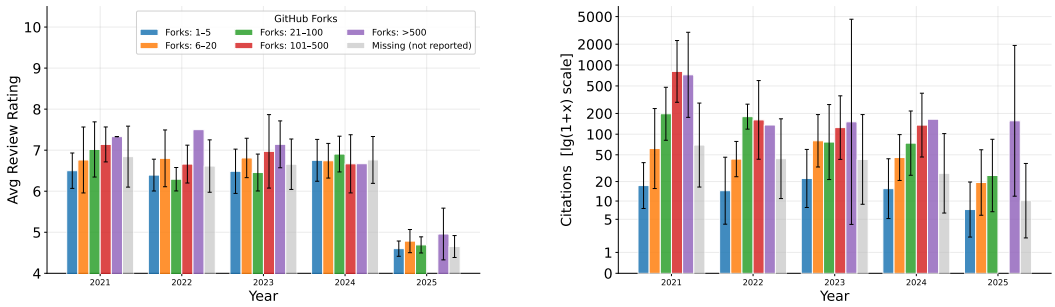


Figure 18: Relationship between GitHub repository fork counts, reviewer scores, and citation counts.

E RATING, RECOGNITION, AND CITATIONS

E.1 HARMONIZATION OF REVIEWER RATINGS

Because reviewer rating scales vary across conferences and years, we harmonize ratings to a common 1 to 5 scale. Specifically, we linearly rescale each rating using the official minimum and maximum of its conference-year-track rating scale. Since this is a simple linear transformation, it does not affect regression fit, statistical significance, or coefficients on other covariates.

Reviewer rating scales differ across conferences, years, and tracks. To make ratings comparable across these settings, we linearly rescale each review score to a common 1–5 range using the official endpoints of the corresponding rating scale. For a raw rating r on a scale with minimum L_g and maximum U_g, we define

$$r_{\text{harmonized}} = 1 + 4\frac{r - L_g}{U_g - L_g}.$$

For reference, this means

Table 19: Harmonization of reviewer rating scales

Original scale	Mapping
1–5	Unchanged
1–6	$1 + 4(r-1)/5$
1–10	$1 + 4(r-1)/9$

E.2 RECOGNITION COMPARISONS WITHIN EXACT REVIEWER SCORES

The main analysis controls for reviewer ratings when comparing papers receiving different levels of conference recognition. As a more restrictive specification, we compare papers only within cells that have exactly the same reviewer score.

We summarize reviewer assessments using the median rating among reviews with recorded confidence of at least four. For the spotlight-versus-regular comparison, we retain conference–year–score cells that contain both spotlight and regular papers and exclude oral papers. For the oral-versus-spotlight comparison, we compare papers within conference–year–track–score cells, since both recognition categories occur across multiple conference tracks.

For paper i, we estimate

$$\log(1 + C_i) = \alpha_{s(i)} + \beta D_i + \varepsilon_i, \tag{2}$$

where C_i is the citation count, D_i indicates the higher recognition category, and $\alpha_{s(i)}$ is an exact-score cell fixed effect. Thus, the specification absorbs all differences across conferences, years, and recorded reviewer scores; for the oral-versus-spotlight comparison, it additionally absorbs conference track. The coefficient β is identified only from papers that receive different recognition outcomes despite having the same recorded reviewer assessment.

Because citations are highly skewed and include zeros, our primary outcome is $\log(1 + C_i)$. We report $\exp(\hat{\beta})$, which can be interpreted as the ratio of adjusted geometric means of $1 + C_i$ between the higher- and lower-recognition groups. We consider both total citations and Semantic Scholar influential citations.

Table 20 shows that the recognition gradient remains when comparisons are restricted to papers with exactly the same reviewer score. Spotlight papers have higher subsequent citation impact than regular papers within the same score cells, and oral papers similarly have higher citation impact than spotlight papers. The pattern also appears for influential citations.

This specification removes differences captured by the recorded reviewer score without imposing a particular functional relationship between ratings and citations. It therefore provides a stricter version of the rating-adjusted comparison reported in the main text. It does not, however, imply that recognition is randomly assigned: area chairs and program committees may use qualitative

Comparison	Citation measure	N	Ratio	95% interval
Spotlight / regular	Total, best available	3,317	1.397	[1.210, 1.614]
Spotlight / regular	Total, single record	3,217	1.423	[1.233, 1.643]
Spotlight / regular	Influential, single record	3,217	1.269	[1.037, 1.552]
Oral / spotlight	Total, single record	452	1.630	[1.173, 2.265]
Oral / spotlight	Influential, single record	452	1.698	[1.322, 2.180]

Table 20: Recognition comparisons with exact-score fixed effects. Ratios are exponentiated coefficients from Equation 2, comparing geometric means of one plus the citation count. Spotlight comparisons hold conference, year, and score fixed; oral comparisons also hold track fixed.

information about a paper that is not summarized by its numerical rating. We investigate this remaining concern using the threshold-based design in Appendix E.3.

E.3 Threshold-Based Design

The exact score comparisons in the main text ask whether recognition is associated with citations among papers receiving the same reviewer assessment. As a complementary design, we exploit discontinuous changes in the probability of recognition around score thresholds.

For paper i in cohort g, let R_{ig} denote its median reviewer score and let c_g denote the relevant recognition threshold. We define the centered running variable

$$X_{ig} = R_{ig} - c_g \tag{3}$$

and an indicator for being on or above the threshold,

$$Z_{ig} = \mathbf{1}\{X_{ig} \geq 0\}. \tag{4}$$

The treatment indicator D_{ig} denotes receipt of the higher recognition category: spotlight rather than regular acceptance in the first comparison, and oral rather than spotlight in the second. The outcome is

$$Y_{ig} = \log(1 + C_{ig}),$$

where C_{ig} is the paper's citation count.

Crossing the threshold does not deterministically assign recognition. Rather, it changes the probability that a paper receives the higher recognition category. This motivates a fuzzy regression-discontinuity design (Angrist & Lavy, 1999; Cattaneo & Titiunik, 2022). Within a local window around the cutoff, the first stage is

$$D_{ig} = \alpha_g + \pi Z_{ig} + \beta_1 X_{ig} + \beta_2 X_{ig} Z_{ig} + u_{ig}, \tag{5}$$

where α_g denotes cohort fixed effects and the interaction allows the score trend to differ on the two sides of the cutoff. The coefficient π measures the discontinuous change in the probability of recognition at the threshold.

We analogously estimate the reduced-form discontinuity in citations:

$$Y_{ig} = \gamma_g + \rho Z_{ig} + \theta_1 X_{ig} + \theta_2 X_{ig} Z_{ig} + \varepsilon_{ig}. \tag{6}$$

Here, ρ measures the change in citation outcomes at the same score threshold. The fuzzy-RD estimand is the ratio of these two discontinuities,

$$\tau_{\mathrm{FRD}} = \frac{\rho}{\pi}, \tag{7}$$

which is equivalently obtained by instrumenting recognition D_{ig} with the above-cutoff indicator Z_{ig} in a local 2SLS regression:

$$Y_{ig} = \delta_g + \tau_{\mathrm{FRD}} \widehat{D}_{ig} + \lambda_1 X_{ig} + \lambda_2 X_{ig} Z_{ig} + \eta_{ig}. \tag{8}$$

For local-constant specifications, the score-slope terms in Equations 5–8 are omitted. The spotlight-versus-regular analysis defines cohorts by conference and year, while the oral-versus-spotlight analysis additionally preserves conference track.

The identifying intuition is that papers very close to the same score threshold should be similar in underlying quality, while crossing the threshold changes their probability of receiving additional conference recognition. Under the standard fuzzy-RD assumptions the ratio in Equation 7 can be interpreted as the local effect of recognition for papers whose recognition status is affected by the threshold.

Table 21: Number of dataset and benchmark papers by year and presentation type across ICLR, ICML, and NeurIPS, including the NeurIPS Datasets and Benchmarks Track.

Year	Oral	Spotlight	Poster	Notable	Unknown	Total	Missing Citations
2021	9	40	204	0	258	511	2
2022	8	10	75	0	454	547	1
2023	27	66	765	43	0	901	0
2024	55	165	1142	0	45	1407	3
2025	77	267	1790	0	66	2200	4
Total	176	548	3976	43	823	**5566**	10

Notes: Presentation types follow the available metadata. Notable includes ICLR 2023's top-5% and top-25% categories. ICML OralPoster and SpotlightPoster labels are counted as Oral and Spotlight, respectively. Unknown indicates unavailable presentation-type metadata. Missing Citations counts papers without a usable citation count, including unresolved identities; these papers are included in Total.

Table 22: Fuzzy Regression Discontinuity estimates

	Spotlight versus regular (1)	Oral versus spotlight (2)
First stage		
Above-cutoff indicator Z_i	0.346^{***}	0.127^{**}
	(0.032)	(0.041)
Reduced form		
Above-cutoff indicator Z_i	0.250	0.427
	(0.155)	(0.267)
2SLS		
Higher recognition D_i	0.722	3.365
	(0.431)	(2.444)
Observations	867	398
SE clusters	8	6
First-stage F	116.093	9.584
Cohort fixed effects	Yes	Yes

Notes: All columns use log(1+citation) as final outcome. Scores are medians of verified reviews with confidence at least 4. Recognition is instrumented by the above-cutoff indicator. Cohort means conference-year for spotlight versus regular and conference-year-track for oral versus spotlight. Standard errors in parentheses are clustered by conference-year and use a finite-sample correction. The F statistic is the squared clustered first-stage t statistic. Stars use nominal two-sided t_{G-1} tests, where G is the number of clusters: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.