Position: AI/ML Influencers Have a Place in the Academic Process

Iain Xie Weissburg [*1] **Mehir Arora** [*2] **Xinyi Wang** [2] **Liangming Pan** [2] **William Yang Wang** [2]

Abstract

As the number of accepted papers at AI and ML conferences reaches into the thousands, it has become unclear how researchers access and read research publications. In this paper, we investigate the role of social media influencers in enhancing the visibility of machine learning research, particularly the citation counts of papers they share. We have compiled a comprehensive dataset of over 8,000 papers, spanning tweets from December 2018 to October 2023, alongside controls precisely matched by 9 key covariates. Our statistical and causal inference analysis reveals a significant increase in citations for papers endorsed by these influencers, with median citation counts 2-3 times higher than those of the control group. Additionally, the study delves into the geographic, gender, and institutional diversity of highlighted authors. Given these findings, we advocate for a responsible approach to curation, encouraging influencers to uphold the journalistic standard that includes showcasing diverse research topics, authors, and institutions.

1. Introduction

In the evolving landscape of artificial intelligence and machine learning (AI/ML), the exponential increase in conference papers, as depicted in Figure 1, alongside rapid technological advancements, has significantly transformed the dissemination of scholarly knowledge. A notable aspect of this transformation is the practice of online preprint sharing, with platforms like ArXiv becoming particularly prominent within the AI/ML community. This phenomenon allows for early access to research, often months before official publication, raising questions about these papers' evolving relevance and reception in traditional academic forums. Most notably, how do people select papers to read in the online age?

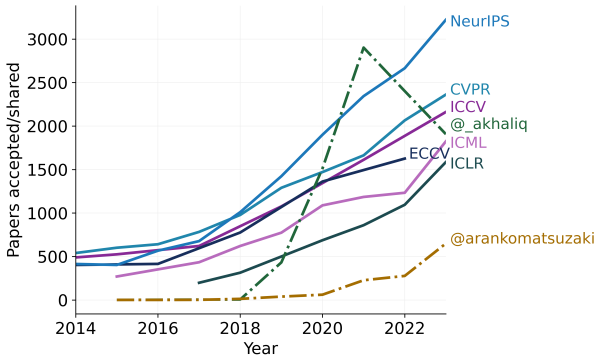


Figure 1. The number of papers accepted to top AI/ML conferences (solid) and shared by influencers (dashed) from 2014-2023 (Li, 2023).

This paper aims to explore the changing dynamics of academic discourse within the AI/ML community, emphasizing the constructive role of social media in addressing the challenges posed by the sheer volume of literature. We study the class of $\mathbb{X}$ (formerly Twitter) influencers. Specifically, we require that influencers (1) share freshly posted preprints from ArXiv, (2) regularly post about papers, (3) do not focus on self-promotion, and (4) have a large audience. AK (@_akhaliq) and Aran Komatsuzaki (@arankomatsuzaki) are the only influencers found to satisfy our requirements.

Acknowledging the vital function of these influencers as curators, we examine the impact of their endorsements on the citation counts of shared papers. However, we also underscore the importance of maintaining a balanced research ecosystem. An over-reliance on a select group of curators may inadvertently skew the research landscape, emphasizing certain topics or perspectives over others. Therefore, we advocate for a responsible approach to curation, encouraging influencers to maintain journalistic standards, showcasing diverse research topics, authors, and institutions.

[*]Equal contribution [1]Department of Electrical and Computer Engineering, University of California, Santa Barbara, USA [2]Department of Computer Science, University of California, Santa Barbara, USA. Correspondence to: Iain Xie Weissburg <ixw@ucsb.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Contributions. To this end, we provide the following three main contributions to the discussion about the increasing impact of academic social media figures in the AI/ML domain:

• **Comprehensive Dataset with Control Samples.** We collect a dataset of shared papers from AK and Komatsuzaki containing key paper information. We select control samples through precise matching, using several paper and author covariates.

• **Thorough Analysis of Citations and Demographics.** Our comprehensive analysis aims to determine if papers these influencers share receive higher citation counts than non-shared papers. We conduct statistical and causal inference analysis and find significant evidence for a direct effect. We then explore the geographic and gender distributions of influencer-shared papers and compare them to reference points from the community.

• **Proposals for Future AI/ML Information Sharing.** Finally, we propose that the academic community engage in a future discussion on evolving the conference system. We want to collaborate with program chairs at major ML conferences to find solutions to the AI/ML research cycle outpacing current conference cycles, making many works outdated by the time of conferences. Second, we propose a future workshop to address the issue by informing the AI/ML community and gathering solutions from community members. Third, we call for a panel of influencers, industry, and academics to discuss how the evolving landscape of AI/ML research can be improved.

2. Related Work

For over a decade now, social media platforms–namely 𝕏–have been studied as an influential means of scholarly communication. Darling et al. (2013) discusses the potential role of Twitter in many stages of the 'life-cycle' of academic authorship and publication. Eysenbach (2011) provides early evidence to support social media sharing as a predictor of higher citation count in medical publications. More recently, studies have shown a statistically significant relationship between Twitter presence and citation counts across fields, especially after the first tweet (Peoples et al., 2016; Vaghjiani et al., 2021). Others suggest that the correlation between tweets and citations is insignificant through randomized trials (Tonia et al., 2016; Branch et al., 2023).

We note that the AI/ML community is much more active on platforms like 𝕏. While many previous works examine fields with a much lower volume of repository publications, the short research cycle and the need for quick dissemination of results, coupled with the rise of arXiv, motivates us to study the problem from the perspectives of modern AI/ML research lens. Additionally, the literature thus far has focused on social media as a forum for multi-user online discussion. In our work, we examine top-down dissemination from singular influencers, with many more followers than previously studied. We also contribute a geographic and gender analysis of influencer-sharing patterns and provide recommendations to conferences, influencers, and the wider community on managing the evolving field of Artificial Intelligence research.

	AK	**Komatsuzaki**
# Unique Papers	9,171	1,273
# with all attributes	8,259 (90%)	1,191 (94%)
# Matched Papers	6,890 (75%)	955 (78%)

Table 1. The number of unique papers tweeted by each influencer, the subset of papers with all desired attributes available through S2 (Table 3), and the papers finally included in our analysis after matching.

AK		**Komatsuzaki**	
Name	Freq.	Name	Freq.
S. Levine	85	L. Zettlemoyer	31
Furu Wei	82	Quoc V. Le	28
Jianfeng Gao	71	Yi Tay	26
L. Zettlemoyer	64	Furu Wei	23
Ziwei Liu	62	M. Dehghani	18

Table 2. The top 5 most common authors shared by each user and the number of papers where they are credited.

3. Data Collection

We model our analysis on retrospective cohort studies, in which a treatment and control group with identical underlying covariates are compared to determine the average treatment effect. In our case, we assume that a paper's citation count is most strongly influenced by elapsed time, quality, topic, and author prominence. While elapsed time is simple to measure, paper quality and topic are difficult to quantify. We use the publication venue and year as a proxy for quality and use a text embedding of the paper's title and abstract to approximate the topic. We use author citations and h-indices to adjust for their prominence.

As such, our data collection process consists of three parts: (1) collecting the Target Set, the papers tweeted by @_akhaliq and @arankomatsuzaki, (2) collecting a large dataset of potential papers to match against, and (3) forming the Control Set by matching papers from (1) to papers from (2) with respect to the year of publication, the publication venue, and a text embedding of title and abstract. We detail each step below.

`title, abstract, year, venue, citationCount, embedding.specter_v2, authors.{name, affiliations, hIndex, citationCount}`

Table 3. The desired attributes for each paper, queried from S2.

`year, venue, open_access, n_authors, first(h-index), max(h-index), first(citations), max(citations), topic_embedding`

Table 4. The covariates used for control matching. `first` = the corresponding value for the paper's first author, `max` = the maximum corresponding value among a paper's authors.

Target Set. The Target Set is collected by searching for document identifiers in the $\mathbb{X}$ feeds of both influencers (identifiers are found in links to `arXiv.org` or `huggingface.co/papers`). We use the Semantic Scholar (S2) API to query every document for desired attributes (Table 3). Notably, some papers are either not available on S2 or are not freely accessible. Hence, we remove any paper that lacks any required attribute. The number of unique papers and papers with all attributes for each influencer is shown in Table 1.

We find that roughly 90% of papers tweeted by either influencer have all attributes available. While we express moderate caution about the excluded minority of papers, as they may be systematically different from those included, we believe the many papers remaining in our analysis are sufficient to draw meaningful conclusions.

Additionally, we find that half of the papers are tweeted within 24 hours of their initial ArXiV release, with 95% of tweets occurring within a week. This verifies that our influencers do indeed tweet freshly posted papers. We also find significant overlap in sharing, with 65% of Komatsuzaki-shared papers also tweeted by AK.

Control Set. To build our Control Set, we first collect a corpus of papers presented at the same venues and in the same years as those in the Target Set. Specifically, for every instance of a paper published in year y at venue v, we query S2 for all papers published in year y at venue v.

We follow the recommendations in (King & Nielsen, 2019) by performing exact matching on our categorical (year, venue, open_access[1]) and binned variables (n_authors, h-index, author cites), and Euclidean distance[2] matching on our continuous variables (topic_embedding). Optimal matching can be reduced to the linear sum assignment problem, for which many efficient algorithms exist (Crouse, 2016). We use the implementation available in SciPy (Virtanen et al., 2020).

[1]Due to the nature of influencer sharing, all sampled papers are open access.

[2]Note that Euclidean distance is equivalent to cosine similarity for normalized vectors.

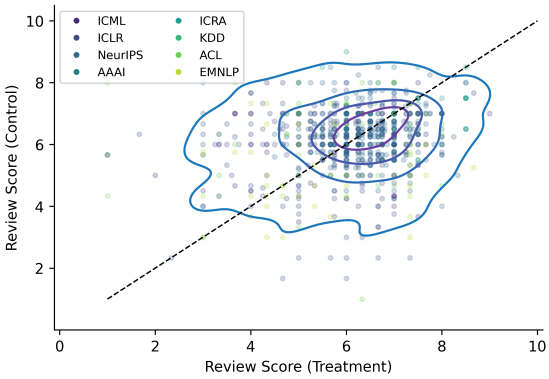


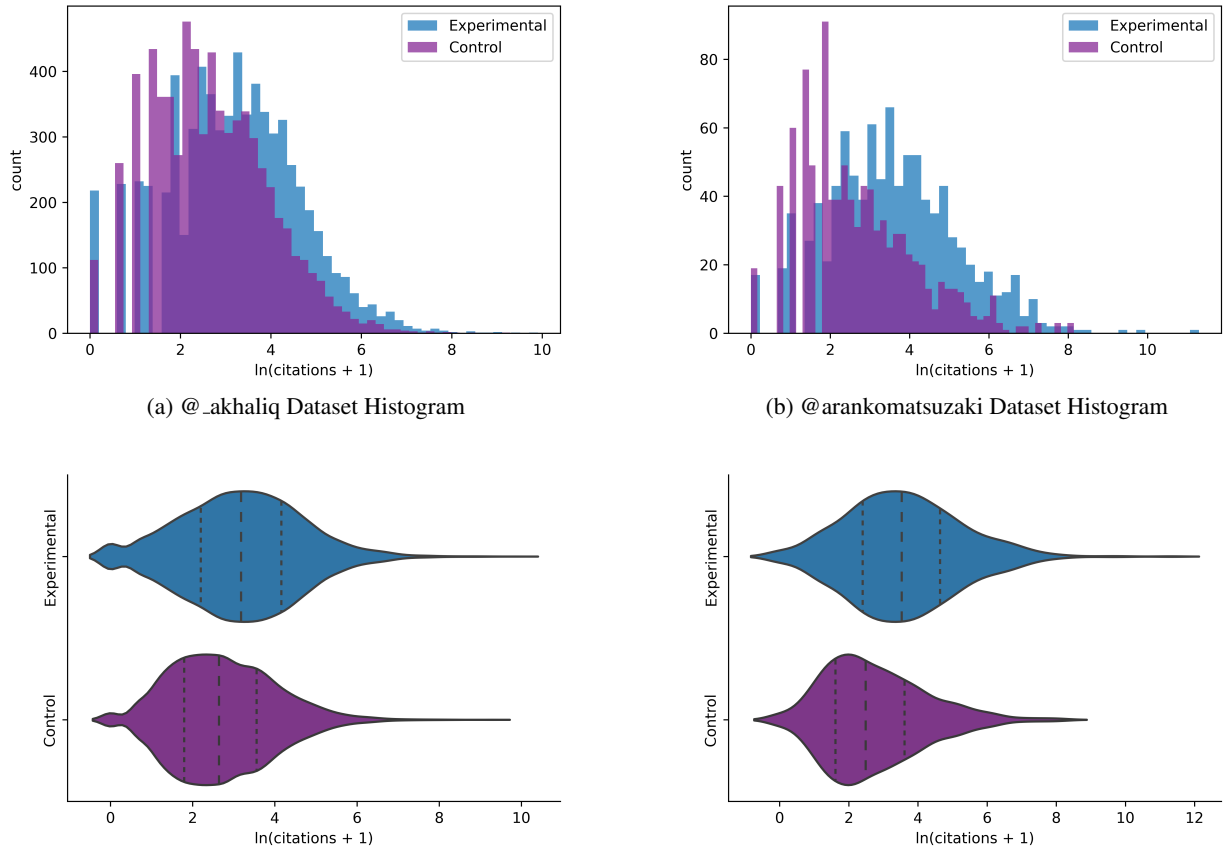
Figure 2. Mean OpenReview scores of tweeted papers vs non-tweeted controls from 8 major ML conferences, the kernel density estimate of the joint distribution of scores, and the identity line (dotted). This shows the quality of papers in both sets are more or less equivalent.

Using this methodology, we match each paper in the Target Set to a paper in the Control Set with respect to the gathered covariates (Table 4). We exclude any paper for which we cannot find an exact match on categorical variables. For matching on topic embeddings, we use SPECTER2 (Singh et al., 2022). Matches between papers were strong, and the distribution of cosine similarities and example matched pairs are detailed in Appendix A.

Review Scores. To determine if we have successfully controlled for quality through our matching, we will look at the review scores of experimental and control pairs from some selected conferences. We extract the review data using OpenReview (2023). Across both paired datasets, we found 939 out of 7222 unique pairs with available scores.

We plot each treatment paper's mean review score against its paired control's mean score (Figure 2). For conferences that do not use numbered review scores, we assign numbers based on those of other conferences (7: Accept, 5: Borderline Accept, 3: Borderline Reject, 2: Reject).

Using the same three significance tests from Table 5, we do not find sufficient evidence to reject the null hypothesis that the control and experimental scores are from the same distribution (p-value > 0.2). Assuming that mean review scores are an accurate measure of paper quality, we conclude that we have effectively controlled for paper quality in our matching.



(a) @_akhaliq Dataset Histogram

(b) @arankomatsuzaki Dataset Histogram

(c) @_akhaliq Dataset Violin Plots

(d) @arankomatsuzaki Dataset Violin Plots

Figure 3. Plots showing the distribution of citations in the two experimental datasets and matched control samples. Citation counts are scaled with the natural logarithm using `numpy.log1p`. Both comparisons show that papers shared by influencers have attained significantly higher citations for all three quartiles than those in the control sets.

4. Analysis

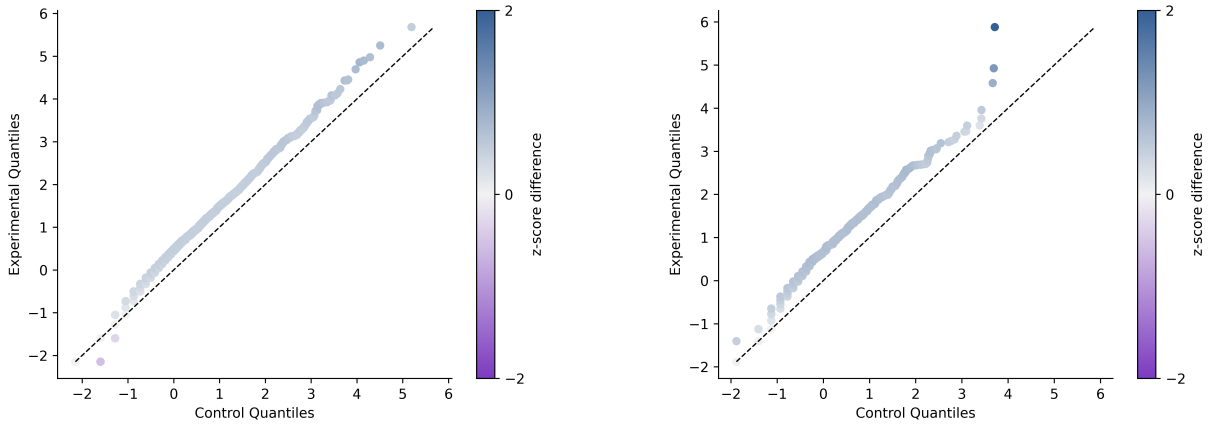
To answer the central question of our work, we compare the impacts of papers shared by AK and Komatsuzaki against our control set. Then, we conduct multivariate analyses by geographic distributions and author attributes in their selected papers.

Contrasting Analysis. For the contrasting analysis, we will test the following hypotheses for correlation:

- **Null:** Influencer-shared papers have the same citation count as others in the same field.
- **Alternative:** Influencer-shared papers have a higher citation count than others in the same field.

We compare our paired target and control sets, as described in Section 3. We find that papers tweeted by AK have a median citation count of 24 (95% CI: 23, 25) versus 14 (95% CI: 13, 15) in the control group, and papers tweeted by Komatsuzaki have a median citation count of 31 (95% CI: 27, 34) versus 12 (95 % CI: 10.5, 13.5). Visually, we can see that both experimental set distributions are skewed toward higher citation counts when compared to their corresponding control sets (Figure 3). In the violin plots (Figures 3c and 3d), the three quartiles and max values are all higher in both of the shared paper distributions compared to the controls. In the 2-Sample Q-Q plots (Figure 4), we can see that the normalized quantiles are consistently higher for the test distributions.

Finally, we establish statistical significance with three tests comparing the distributions of the experimental data with that of the control sets, Epps-Singleton (ES), Kolmogorov-Smirnov (KS), and Mann-Whitney U (MWU), none of which assume normal distribution, which is essential for our data. Table 5 shows the results, all with p-values well below even a stringent $\alpha = 0.001$. From this, we can strongly reject the null hypothesis that the citation distributions for the influencer-shared papers and the control papers are the same.



(a) @_akhaliq Dataset

(b) @arankomatsuzaki Dataset

Figure 4. 2-Sample Q-Q Plots comparing the experiment and control set distributions across every quantile. To build the plot, citation counts are log-scaled, normalized to the control distribution (z-scores), sorted, and paired in order. The dotted line shows an equal distribution; any points above the line show a higher experimental quantile, and vice versa. The plots show that both experimental distributions are consistently higher, especially closer to the median.

Overall, the correlation between influencer tweets and citation count–and not review scores–points to a shift in how the community finds and reads papers. While traditionally, top conference acceptance (i.e. review score) has been a primary indicator of future citation count (Lee, 2018), we have shown that the sharing practices of far-reaching influencers are now a significant indicator of future research impact through citations.

Causal Inference: Setup. Although we have shown a significant correlation between influencer sharing and citations, we have not investigated a causal link. For this, we turn to the model, techniques, and assumptions presented in Elazar et al. (2023). For additional background and details, see Appendix C

We aim to estimate the causal effect of the treatment A, indicating that the paper was shared by influencers[3], on the outcome Y, if the paper is "highly-cited" or "less-cited."

[3]To increase sample size and simplify our analysis, we combine our test sets for this portion of the analysis.

Test	AK	Komatsuzaki
Epps-Singleton	$p < 0.0001$	$p < 0.0001$
Kolmogorov-Smirnov	$p < 0.0001$	$p < 0.0001$
Mann-Whitney U	$p < 0.0001$	$p < 0.0001$

Table 5. 2-sample Statistical significance tests for the difference in distributions of the two experiment and control datasets. These tests suggest a statistically significant difference in the experimental and control distributions for both influencer datasets.

Besides A, there are a number of other factors, known as *confounders*, that can affect Y. We divide these confounders into two sets: *observed* C–which are the same used in our matching (Section 3)–and *unobserved* U–which are difficult to quantify or measure (e.g. novelty, contribution, hype).

To debias the effects of unobserved confounders U, we employ the negative outcome control (NOC) framework (Card & Krueger, 1993; Lipsitch et al., 2010). By finding a *negative control outcome* N that shares the same confounders as Y, but is not causally affected by A, we can attempt to correct for the bias introduced by U. To this end, we select a paper's average review score for N, which we have already shown is not significantly correlated to A (Figure 2) and which we believe share many of the same confounders (e.g., quality, contribution, author experience). However, another NCO can be substituted if deemed more suitable.

To account for fine-grained citation counts and review scores in comparison to the binary nature of Y and N, we define Y_p and N_q. For both, their value is 1 if the corresponding real-valued quality (citation count or average review score) is above the pth (or q^{th}) quantile of the relevant sample distribution. We will use values of $\{50, 75, 90\}$ for both p and q in our analysis.

Finally, we estimate the causal effect of influencer sharing on paper citations with the *average treatment effect of the treated*. Using the difference-in-difference assumption with NOC, we can write our estimate (Elazar et al., 2023):

$$\text{ATET} = \mathbb{E}[Y_{A=1} - N_{A=1}] - \mathbb{E}[Y_{A=0} - N_{A=0}]$$

where ATET $\in [-100\%, 100\%]$ values closer to 0% indi-

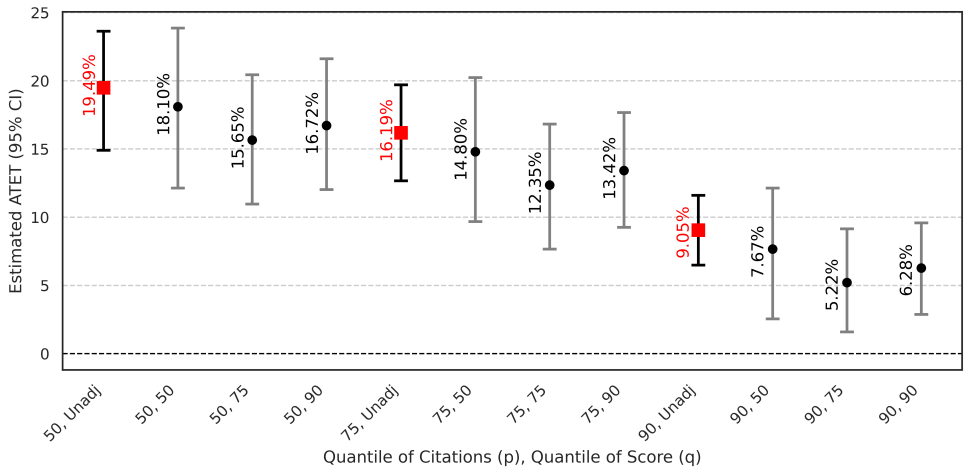


Figure 5. Forest plot of ATET and confidence intervals approximated with negative outcome control (NOC). Larger positive values indicate a stronger positive causal effect of the treatment (influencer sharing) on the outcome (a paper being "highly-cited"). To ensure robust results, we vary the quantile threshold for "highly-cited" and "highly-scored" papers; "Unadj." values (red squares) show the effect estimate before applying NOC. These results, where no confidence intervals contain 0%, indicate a significant positive causal effect of influencer sharing on paper citations.

cate weak to no effect of the treatment, and large absolute values indicate a stronger effect.

Causal Inference: Results. First, we estimate the effect of A on Y without accounting for the effects of U. We estimate ATET (95%-bootstrapping CI) to be 19% (15-24), 16% (13-20), 9% (6-12) for $p = \{50, 75, 90\}$, respectively. This confirms the significant association between influencer sharing and citations we investigated during the contrasting analysis. However, this does not preclude the existence of strong confounders.

For that, we will use NOC to attempt to debias the effects of unobserved confounders, including quality. We record our results in Figure 5 and observe the following: (1) debiased effects are smaller than the unadjusted estimates, indicating our NCO has succeeded in accounting for some unobserved confounders; (2) all of our estimated effects are significant to the 95% level, because none contain 0 in their confidence interval; (3) influencer sharing is most successful at increasing the citations in lower percentiles–papers with very high citation counts experience a smaller (yet still significant) effect compared to those with fewer citations.

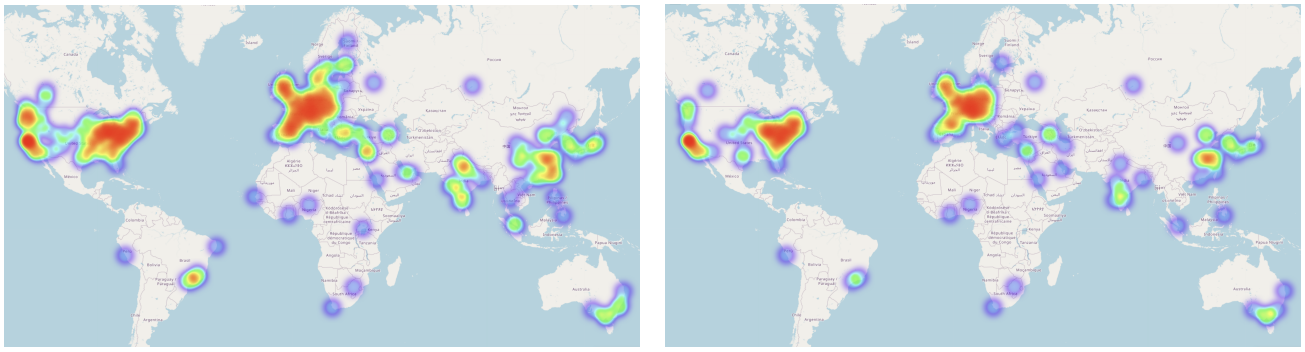
Geographic Distributions. In exploring the evolving landscape of machine learning (ML) paper dissemination, it is essential to consider the implications of a more centralized curation model, particularly as it relates to geographic and gender diversity in scholarly works. Our approach is to present data and observations that highlight trends ***and do not attribute intentional bias to the influencers involved***.

Our analysis begins by examining the geographic distribution in the dissemination of ML papers. Given the American affiliations of AK and Aran Komatsuzaki, we explore whether this translates into a geographic skew in the papers they share. To contextualize our findings, we refer to the geographic distribution of AI repository publications from the Stanford HAI 2023 AI Index Report (Figure 7). We choose to view this data in particular, as our selected influencers share papers from repositories (i.e. ArXiV).

To test this, we first collect the geographic data of the shared papers. First, we use S2 and dblp (2023) to collect the affiliation data of all listed authors from each test set. We then utilize the Nominatim geocoding API to find the approximate latitude and longitude of each affiliation, manually adjusting visibly inaccurate coordinates. From this information, we find the country of each affiliation and then use majority voting to assign each publication a geographic area. At this point, we can see that both influencers share papers from around the world in Figure 6. Finally, we aggregate these countries into the same geographic areas used in the HAI Report and plot using a similar format (Figure 8).

To account for the discrepancy between date ranges in the HAI Report and our influencers' activity, we will limit our analysis to the overlap between them. Additionally, we will focus on the range from 2018 to 2021, because of the low sample size of shared papers pre-2018–only 5 in total.

During these years, Figure 7 indicates a slight decline in the United States' share of AI repository publications following its peak. Concurrently, the European Union and United Kingdom demonstrate a modest uptick in publications after a consistent decline from 2010-2017, while China's share continues to rise.



(a) affiliation heatmap of @_akhaliq-shared authors

(b) affiliation heatmap of @arankomatsuzaki-shared authors

Figure 6. Geographic heatmaps of the unique affiliations of influencer-shared paper authors. High density (red) on the map represents a large number of unique institutions and not a large number of papers, citations, or authors from the given area. Both influencers shared from institutions around the world, with especially large hotspots in the US and Europe.

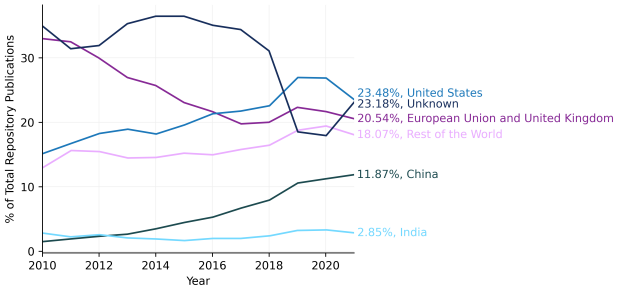


Figure 7. AI repository publications by geographic area and publication year, 2010–2021. Data extracted from Figure 1.1.19 of the Stanford HAI 2023 AI Index Report (Maslej et al., 2023).

In contrast, the sharing patterns of the influencers from 2018 to 2021, shown in Figure 8, demonstrate a notable deviation from these global trends. Specifically, both influencers' data exhibits a consistent focus on U.S.-affiliated papers, with a much smaller, relatively constant portion from other areas. The increase in U.S.-affiliated papers seems to be a change in reporting rather than a change in sharing practices.

We must note, though, that using solely self-reported affiliations can have an inherent bias toward the United States. For example, many researchers affiliated with U.S.-based organizations are assigned to the United States despite working out of another area. Additionally, we must note the prominence of the "Unknown" category in the data of both influencers, where affiliations were not found.

Nevertheless, our results highlight the potential for centralized individuals to shape the perceived narrative of AI research prominence.

Gender Distributions. Beyond geographic diversity, gender diversity is crucial in Computer Science and Engineering, fields historically dominated by men. We extract author names and affiliations as described above. In this section, we filter only the first authors of each paper. For gender classification, we used the AMiner Scholar Gender Prediction API, which categorizes authors as "male," "female," or "UNKNOWN" based on name and affiliation–if available. The API uses a majority vote of the results from three sources: google image search and facial recognition, the Facebook Generated Names List (Tang et al., 2011), and WebGP (Gu et al., 2016).

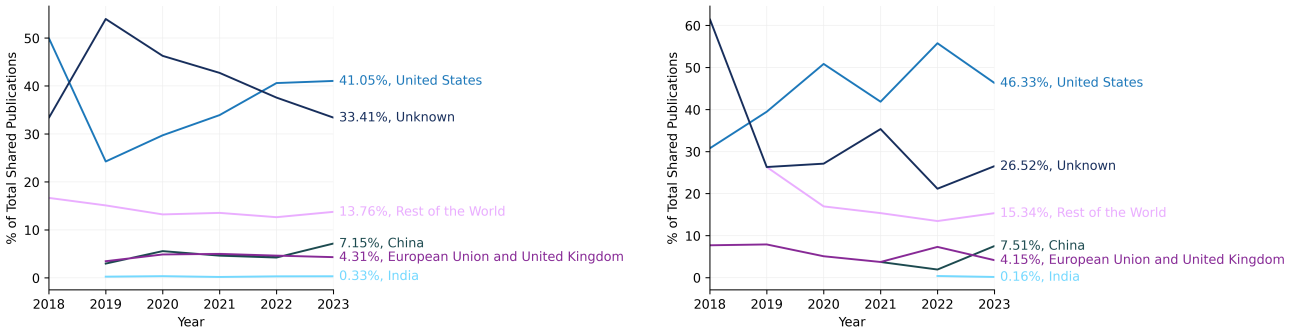
To ground our view of the overall gender distribution in the field, we reference the Taulbee survey's reported gender distribution of US Ph.D. awardees and faculty in CS and related fields from 2021-2022 (Zweben & Bizot, 2023). To match the classifications we have available, we will consider the binary reported genders from the survey.

Our analysis revealed an 80:20 male-to-female ratio among authors with identifiable genders in the @_akhaliq dataset and an 81:19 ratio in the @arankomatsuzaki dataset. These ratios align somewhat with the Taulbee survey's reported 77:23 ratio in computing Ph.D. awardees and deviated slightly more from the 76:24 ratio in faculty. These deviations may stem from a trend toward increasing female representation in the ML space; the survey is recent data, while our influencer data spans several years into the past.

5. Discussion

Our analysis suggests that ML influencers strongly affect paper visibility, indicating a change in how ideas are propagated through the community. We discuss the downstream implications of influencers on the community and make recommendations on how to help improve the paper curation problem and enhance equity in publication visibility.

Influencers and the ML Community. Influencers serve as pivotal curators in the ML landscape. With the explo-



(a) @_akhaliq-shared papers by geographic area per year

(b) @arankomatsuzaki-shared papers by geographic area per year

Figure 8. Influencer-shared publications by geographic area and publication year, 2018-2023, from Semantic Scholar and DBLP affiliation data.

sive growth of machine learning research, the community increasingly relies on social media to keep up-to-date with new developments. Their role in streamlining the dissemination process is akin to that of journalists in news media, making novel ideas and breakthroughs more accessible.

Yet, reliance on a select few narrows the full spectrum of research. We propose maintaining a competitive, open academic space with diverse, highly visible ideas, advocating for influencers to diversify showcased techniques and concepts. This encourages a broad exploration of the ML landscape, mitigating the risk of a homogeneous research narrative.

Enhancing Equity. The move towards influencer-led dissemination opens paths to greater equity in the ML community. Our findings—pointing out the geographic and gender disparities in shared research—underscore the potential for a more inclusive academic dialogue. We clarify that our analysis does not suggest influencers currently exhibit bias; rather, it highlights an opportunity to mitigate existing inequities through proactive online engagement. By promoting diverse global perspectives and addressing gender disparities, influencers can foster a more equitable field. This effort aligns with our commitment to enhance equity without overemphasizing certain demographics, reflecting the concentration of ML research.

To stimulate community reflection and action, we recommend collaborating with conference chairs to address the pace of research, organizing workshops to explore dissemination improvements, and convening panels of influencers, industry, and academia to discuss enhancements in AI/ML research visibility. These steps aim to refine the research dissemination process and uplift the ICML community's engagement with emerging and diverse ideas.

6. Conclusion

Our study delves into the influence of social media influencers on the dissemination and recognition of academic work in AI/ML, specifically examining the impact of AK and Aran Komatsuzaki on platforms like $\mathbb{X}$ and Hugging Face. Our analysis demonstrates that papers endorsed by these influencers receive significantly more citations than those that are not, underscoring the role of influencers in not only amplifying the reach of specific research but also in shaping its visibility in the field. This influence extends beyond merely sharing higher-quality papers, highlighting their ability to promote substantial findings within the community.

We further discuss the dual role of influencers as both catalysts for visibility and curators of content, emphasizing the need for a balance in their influence to ensure a diversity of perspectives in the AI/ML research landscape. The implications of this influence are profound, suggesting a need for the academic community to reassess traditional methods of paper selection and review. Our findings encourage a dialogue among conference organizers and academic institutions to evolve the peer-review process and adapt to these changing norms. Additionally, we encourage future research into these trends in other scientific fields and the mechanisms behind the influence of social media on academic recognition.

Impact Statement

Recognizing the influential role of social media influencers in the academic community, particularly in the fields of AI/ML, it's important to gently encourage them to be aware of their impact. Influencers play a key role in shaping discussions and trends, and there is a growing need for them to consider the diversity and inclusivity of the research they share. This is not to undermine their contributions but to

enhance the richness of academic discourse. Similarly, the academic community might benefit from reevaluating traditional metrics of research impact, embracing both peer-reviewed channels and digital platforms for a more holistic approach to recognizing scholarly work.

In doing so, we hope to spark future research into understanding any unintentional biases influencers might have, such as favoring research from prestigious institutions or specific regions. This is a subtle yet crucial aspect of ensuring a balanced representation in the academic landscape, preventing the overshadowing of significant but less publicized work. While influencers may have personal or commercial interests, we trust in their capacity to acknowledge and manage these to avoid conflicts of interest. This approach would help in preventing echo chambers and promoting a diverse range of perspectives in AI/ML research. Overall, we advocate for a collaborative and mindful effort among influencers and the academic community to foster an equitable and inclusive environment for scholarly discussions.

Acknowledgements

This work was supported in part by the National Science Foundation Research Experience for Undergraduates under CAREER Award Grant No. 2048122.

References

Branch, T. A., Côté, I. M., David, S. R., Drew, J., LaRue, M., Márquez, M. C., Parsons, E. C. M., Rabaiotti, D., Shiffman, D., Steen, D. A., and Wild, A. L. Controlled experiment finds no detectable citation bump from Twitter promotion. *bioRxiv*, 2023. URL https://api.semanticscholar.org/CorpusID:262087567.

Card, D. and Krueger, A. B. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. *NBER Working Paper Series*, 1993. URL https://api.semanticscholar.org/CorpusID:1140202.

Crouse, D. F. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016. doi: 10.1109/TAES.2016.140952.

Darling, E. S., Shiffman, D. S., Côté, I. M., and Drew, J. A. The role of twitter in the life cycle of a scientific publication. *ArXiv*, abs/1305.0435, 2013. URL https://api.semanticscholar.org/CorpusID:7583994.

dblp. dblp Computer Science Bibliography: Monthly Snapshot Release of October 2023. https://dblp.org/xml/release/dblp-2023-10-01.xml.gz, 2023. Accessed: 2023-10-01.

Elazar, Y., Zhang, J., Wadden, D., Zhang, B., and Smith, N. A. Estimating the causal effect of early arxiving on paper acceptance. *ArXiv*, abs/2306.13891, 2023. URL https://api.semanticscholar.org/CorpusID:259251893.

Eysenbach, G. Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. *Journal of Medical Internet Research*, 13, 2011. URL https://api.semanticscholar.org/CorpusID:2157129.

Feder, A., Keith, K. A., Manzoor, E. A., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., Stewart, B. M., Veitch, V., and Yang, D. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2021. URL https://api.semanticscholar.org/CorpusID:237386009.

Gu, X., Yang, H., Tang, J., and Zhang, J. Web User Profiling Using Data Redundancy. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '16, pp. 358–365. IEEE Press, 2016. ISBN 9781509028467.

Imbens, G. and Rubin, D. B. Causal inference for statistics, social, and biomedical sciences: An introduction. 2015. URL https://api.semanticscholar.org/CorpusID:123646167.

King, G. and Nielsen, R. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, 27(4): 435–454, 2019 2019. URL https://doi.org/10.1017/pan.2019.11.

Lee, D. H. Predictive power of conference-related factors on citation rates of conference papers. *Scientometrics*, 118:281–304, 2018. URL https://api.semanticscholar.org/CorpusID:53247921.

Li, X. Conference-Acceptance-Rate, 2023. URL https://github.com/lixin4ever/Conference-Acceptance-Rate. GitHub repository. Accessed: 2023-12-18.

Lipsitch, M., Tchetgen, E. J. T., and Cohen, T. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21:383–388, 2010. URL https://api.semanticscholar.org/CorpusID:13292337.

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., and Perrault, R. Artificial Intelligence Index Report 2023, 2023.

OpenReview. OpenReview API. https://api.openreview.net, 2023. Accessed: 2024-01-29.

Peoples, B. K., Midway, S. R., Sackett, D. K., Lynch, A. J., and Cooney, P. B. Twitter Predicts Citation Rates of Ecological Research. *PLoS ONE*, 11, 2016. URL https://api.semanticscholar.org/CorpusID:13689703.

Sheppard, L., Levy, D., Norris, G. A., Larson, T. V., and Koenig, J. Q. Effects of ambient air pollution on nonelderly asthma hospital admissions in seattle, washington, 1987-1994. *Epidemiology*, 10 1:23–30, 1999. URL https://api.semanticscholar.org/CorpusID:36893348.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://api.semanticscholar.org/CorpusID:254018137.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen, E. J. T. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 31 3:348–361, 2016. URL https://api.semanticscholar.org/CorpusID:207137332.

Tang, C., Ross, K., Saxena, N., and Chen, R. What's in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook. In Xu, J., Yu, G., Zhou, S., and Unland, R. (eds.), *Database Systems for Advanced Applications*, volume 6637 of *Lecture Notes in Computer Science*, Berlin, Heidelberg, 2011. Springer. doi: 10.1007/978-3-642-20244-5_33.

Tonia, T., Oyen, H. V., Berger, A., Schindler, C., and Künzli, N. If I tweet will you cite? The effect of social media exposure of articles on downloads and citations. *International Journal of Public Health*, 61:513–520, 2016. URL https://api.semanticscholar.org/CorpusID:27919192.

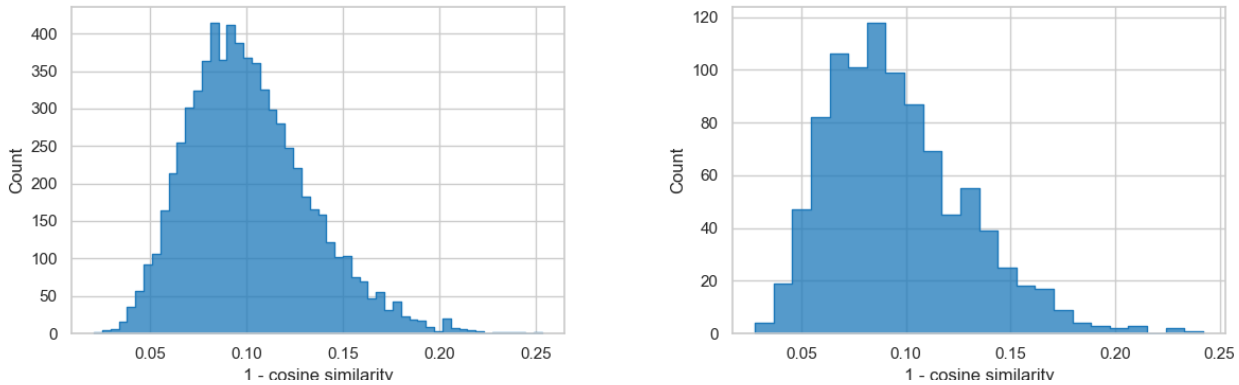
Vaghjiani, N. G., Lal, V., Vahidi, N., Ebadi, A., Carli, M., Sima, A. P., and Coelho, D. H. Social Media and Academic Impact: Do Early Tweets Correlate With Future Citations? *Ear, nose, & throat journal*, pp. 1455613211042113, 2021. URL https://api.semanticscholar.org/CorpusID:237292881.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Zhang, B., Small, D. S., Lasater, K. B., McHugh, M., Silber, J. H., and Rosenbaum, P. R. Matching one sample according to two criteria in observational studies. *Journal of the American Statistical Association*, 118:1140 – 1151, 2021. URL https://api.semanticscholar.org/CorpusID:240560096.

Zweben, S. and Bizot, B. CRA 2022 Taulbee Survey: Record Doctoral Degree Production; More Increases in Undergrad Enrollment Despite Increased Degree Production. *Computing Research News*, May 2023.

A. Target-Control Matching



(a) Cosine Distances of Matches for @_akhaliq

(b) Cosine Distances of Matches for @arankomatsuzaki

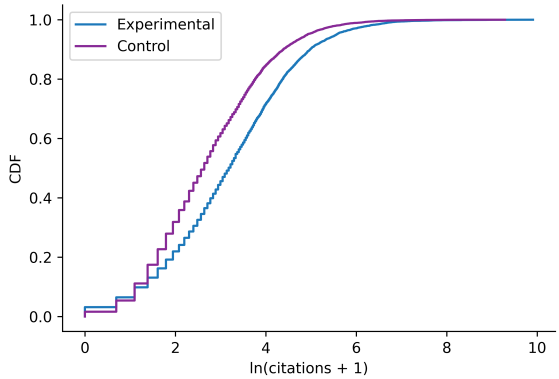
Figure 9. Plots of the cosine distances (1 - cosine similarity) between matched pairs for each of @_akhaliq and @arankomatsuzaki. We note that the cosine distances are low, with the a mode of around 0.07 for both. This indicates that most pairs are very well matched in topic.

Qualitatively, the matched pairs are very similar in topic, almost always covering the same sub-field of research (for example, language model hallucinations). We supply a random sample of matched pairs in table 6. The distribution of scores is shown in table 9, and shows that the matched pairs have high cosine similarity scores. Our results are consistent across many choices of matching schemes, such as when we use more than three quantiles for binned variables, or exclude author characteristics altogether.

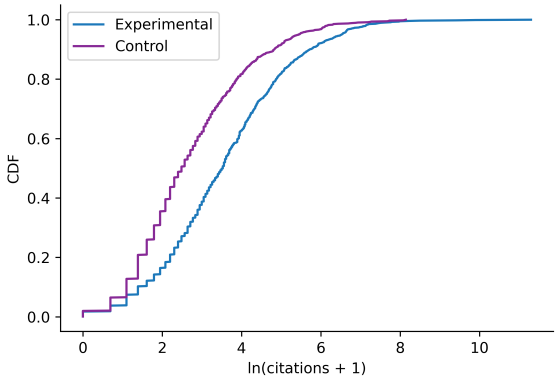
Score: 0.09	**Target Title:** Unifying Vision, Text, and Layout for Universal Document Processing **Target Abstract:** We propose Universal Document Processing (UDOP), a foundation Document AI model which unifies text, image, and layout modalities together with varied task forma... **Control Title:** SwinTextSpotter: Scene Text Spotting via Better Synergy between Text Detection and Text Recognition **Control Abstract:** End-to-end scene text spotting has attracted great attention in recent years due to the success of excavating the intrinsic synergy of the scene text detection ...
Score: 0.09	**Target Title:** reStructured Pre-training **Target Abstract:** In this work, we try to decipher the internal connection of NLP technology development in the past decades, searching for essence, which rewards us with a (pote... **Control Title:** Adapting BigScience Multilingual Model to Unseen Languages **Control Abstract:** We benchmark different strategies of adding new languages (German and Korean) into the BigScience's pretrained multilingual language model with 1.3 billion para...
Score: 0.12	**Target Title:** Quantifying Memorization Across Neural Language Models **Target Abstract:** Large language models (LMs) have been shown to memorize parts of their training data, and when prompted appropriately, they will emit the memorized training dat... **Control Title:** How Much Data Are Augmentations Worth? An Investigation into Scaling Laws, Invariance, and Implicit Regularization **Control Abstract:** Despite the clear performance benefits of data augmentations, little is known about why they are so effective. In this paper, we disentangle several key mechani...
Score: 0.07	**Target Title:** Do Language Models Know When They're Hallucinating References? **Target Abstract:** State-of-the-art language models (LMs) are famous for"hallucinating"references. These fabricated article and book titles lead to harms, obstacles to their use, ... **Control Title:** Retrieving Supporting Evidence for LLMs Generated Answers **Control Abstract:** Current large language models (LLMs) can exhibit near-human levels of performance on many natural language tasks, including open-domain question answering. Unfo...
Score: 0.10	**Target Title:** Singularity: Planet-Scale, Preemptive and Elastic Scheduling of AI Workloads **Target Abstract:** Lowering costs by driving high utilization across deep learning workloads is a crucial lever for cloud providers. We present Singularity, Microsoft's globally d... **Control Title:** Edge Impulse: An MLOps Platform for Tiny Machine Learning **Control Abstract:** Edge Impulse is a cloud-based machine learning operations (MLOps) platform for developing embedded and edge ML (TinyML) systems that can be deployed to a wide r...
Score: 0.09	**Target Title:** A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning **Target Abstract:** We present a large-scale study on unsupervised spatiotemporal representation learning from videos. With a unified perspective on four recent image-based framewo... **Control Title:** Target-Aware Object Discovery and Association for Unsupervised Video Multi-Object Segmentation **Control Abstract:** This paper addresses the task of unsupervised video multi-object segmentation. Current approaches follow a two-stage paradigm: 1) detect object proposals using ...

Table 6. Randomly sampled matched pairs of papers tweeted by Komatsuzaki. Score refers to $1 -$ cosine similarity, so lower scores indicate a closer match. Controls show similar topics to the target individuals across similarity levels, indicating a good match.

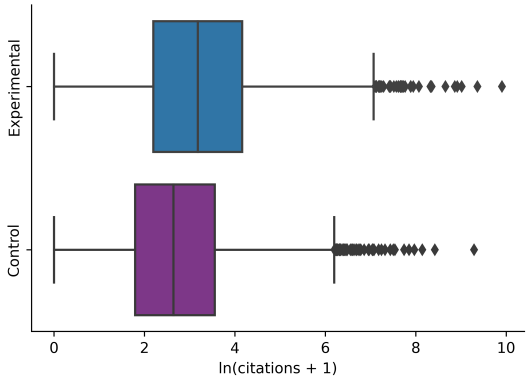
B. Additional Analysis Plots



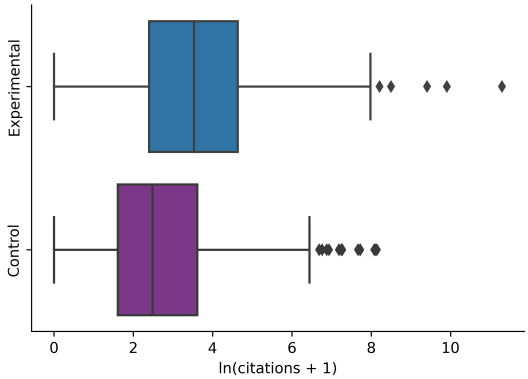
(a) @_akhaliq Dataset Cumulative Distribution Functions



(b) @arankomatsuzaki Dataset Cumulative Distribution Functions



(c) @_akhaliq Dataset Box Plots



(d) @arankomatsuzaki Dataset Box Plots

Figure 10. Additional plots showing the distribution of citations in the two experimental datasets and matched control samples. Citation counts are scaled with the natural logarithm using `numpy.log1p`.

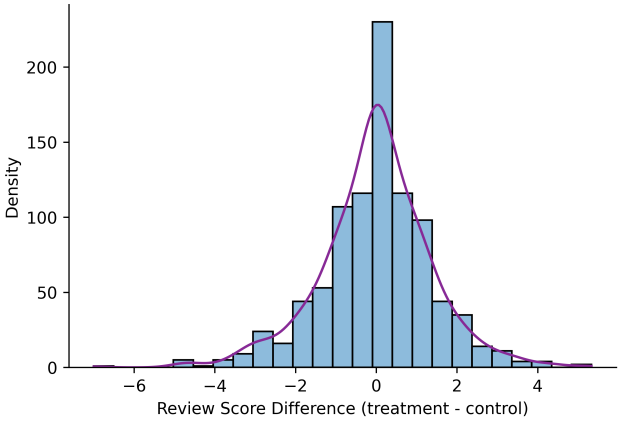


Figure 11. Histogram and kernel distribution estimate of the OpenReview score difference between experiment and control samples in the merged dataset. The mean at 0, and symmetric KDE reinforce that both review scores are from the same distribution.

C. Causal Inference

For our causal inference analysis, we draw primarily from the model, techniques, and assumptions presented in Elazar et al. (2023). In turn, they point interested readers to Imbens & Rubin (2015) and Feder et al. (2021).

Background. In our work, we study the outcome as a binary variable Y, indicating if a paper is "highly-cited" ($Y = 1$) or "less-cited" ($Y = 0$). By introducing our treatment A, we can define two cases for any given paper: Y_1, the outcome if the paper were shared by influencers (when $A = 1$), or Y_0, the outcome if the paper were not shared by influencers (when $A = 0$). In actuality, we can only observe exactly one of these for each paper. Rather, the goal of causal inference is to estimate the change in outcome if the treatment were flipped (Figure 12). Specifically, we hope to answer the question, *does influencer sharing increase the likelihood of a paper being highly-cited?*

To accomplish this, we use the *average treatment effect on treated*:

$$\text{ATET} = \mathbb{E}[Y_1 - Y_0 | A = 1]$$

which is the expected increase in highly-cited papers when shared by influencers, for those already shared.

Negative Outcome Control and Difference-in-Difference. To debias the effects of our unobserved confounders, we will utilize the Negative Outcome Control (NOC) framework (Card & Krueger, 1993; Lipsitch et al., 2010). Specifically, we will use Difference-in-Difference (DiD) to estimate ATET over two time periods $t \in \{0, 1\}$ (Sheppard et al., 1999):

$$\text{ATET} = \mathbb{E}[Y_1(1) - Y_0(1)] - \mathbb{E}[Y_1(0) - Y_0(0)]$$

which can be estimated by a fitting a logistic model and bootstrapping for confidence intervals (Elazar et al., 2023).

Sofer et al. (2016) establishes that NCO and DiD-based estimators are equivalent, meaning ATET can be rewritten as:

$$\text{ATET} = \mathbb{E}[Y_1 - N_1] - \mathbb{E}[Y_0 - N_0]$$

and solved by taking the difference across treatment groups of the difference between Y and N.

Assumptions. To perform our causal estimate, we assume the following conditions hold:

1. **Ignorability:** $\{Y_0, Y_1\} \perp\!\!\!\perp A | (C, U)$
2. **Positivity:** $0 < \mathbb{P}(A = 1 | C = c, U = u) < 1$
3. **Consistency:** $Y_a = Y^{obs}$ *if* $A^{obs} = a$
4. **Negative control:** $N_a = N$ *for* $a = 0, 1$
5. **Additive Equi-confounding:**
   $\mathbb{E}[Y_a(1) - Y_a(0) | U, A = a, C]$
   $= \mathbb{E}[Y_a(1) - Y_a(0) | A = a, C]$ *for* $A = 0, 1$

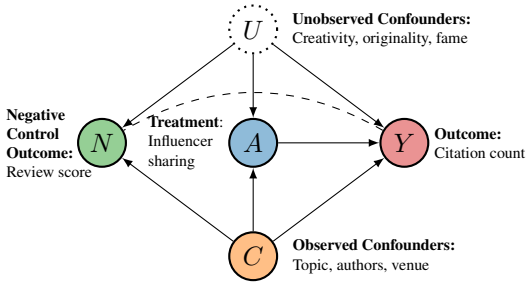


Figure 12. Causal graph of our setup. A and Y are binary treatment and effect variables, respectively: whether a paper was shared by influencers, and whether the paper is highly cited. As we cannot measure the unobserved confounders (e.g. quality), we estimate the effect of sharing using a negative control outcome variable (N). Solid edges represent a directed causal effect, while dashed edges represent an association. Figure and caption modified from Elazar et al. (2023).

In other words, *ignorability* assumes (C, U) are the only confounders with an effect on the outcome; *positivity* means each paper has a non-zero chance of being and not being influencer-shared; *consistency* establishes that the potential outcomes Y_a and observed outcomes Y^{obs} agree for a given treatment level A^{obs}; *negative control* assumes that the NCO is not causally affected by the treatment A; *additive equi-confounding* relies on a linear (or logistic) outcome model for U.

Review Scores as a Negative Control Outcome. The goal of the Negative Outcome Control (NOC) framework (Card & Krueger, 1993; Lipsitch et al., 2010) is to find a Negative Control Outcome (NCO) variable that is affected by the same confounders (C, U) as the outcome Y, but not the treatment. This way, we can de-bias the effects of confounders not captured by the metadata and text embeddings collected from S2. For this, we choose the average conference review score for each paper, collected as described in Section 3. We believe that this sub-sample is representative for the following reasons: (1) influencers often share preprint works before conference acceptance or proceedings are released; (2) our matching accounts for publication venue as a confounder; (3) using bootstrapping on the three tests from Table 5, we find there is not significant evidence to reject that the sub-sample and full dataset citations come from different distributions (p-value > 0.05).

Matching and Causal Inference. Elazar et al. (2023) use tripartite matching on a number of covariates (Zhang et al., 2021), and only sampled from one venue. To control the topic covariate, they use SPECTER embeddings to build 20 topic clusters and minimize distance for numerical variables

(author citations, institution rank, etc.)

Instead, we group the treated and control samples based on a mix of exact matching (venue, year, open access) and quantile binning (number of authors, h-index, and author cites). From these groups, we match samples by minimizing the L_2 difference of text embeddings. This way, we can most accurately capture increased nuance between "viral" topics, techniques, and models.

Additionally, we use first-author information, rather than both minimum and average, to increase our group sizes and improve the quality of topic matches while maintaining descriptive author details.

D. Limitations

While our study provides valuable insights into the role of influencers in the machine learning (ML) community, there are several limitations that must be acknowledged:

Binary Gender Analysis: Our gender prediction utilizes the AMiner Gender Prediction API, which combines the Facebook Generated Names List (Tang et al., 2011) with precision and recall of 95% and 81%, and MagicFG (Gu et al., 2016) with precision and recall of 93% and 94% respectively. Exact numbers for the combined API are not published by the creators but can reasonably be assumed to be higher than each individual method. We will ensure that these details are more accessible in the final paper.

Additionally, the gender prediction API only outputs binary genders (male and female), limiting our ability to capture the full spectrum of gender identities. Consequently, our findings may not accurately represent the diversity of gender identities within the ML community.

Accuracy with Non-Western Names: The gender prediction API may exhibit lower accuracy when analyzing non-Western names (Gu et al., 2016). As a result, the study's conclusions regarding gender disparity might not fully capture the global diversity of our samples.

Self-Reported Affiliations: The affiliations in our dataset are self-reported, which can lead to inconsistencies such as misspellings or missing location information. This limitation affects the accuracy of our geographical analysis, as the data may not accurately reflect the true affiliations or locations of the authors. Consequently, our observations regarding geographical concentration and diversity must be interpreted with caution. Additionally, this limitation posed as a barrier to investigating institutions as an observed confounder for causal inference. We invite future studies with more complete affiliation data to investigate this.

Lack of Randomized Control Trial: The study did not include a randomized control trial, often considered the gold standard for establishing causality. Without this, we cannot conclusively determine whether the patterns observed are directly attributable to the influencers' activities or are coincidental. Though we attempt to remedy this through causal inference, such techniques are contingent upon the validity of our assumptions.