The Double Bind: Revisiting Preprinting and Peer Review Two Years After the Removal of the ACL Anonymity Period

A Pranav[†], **Shane Storks**[‡], **Anne Lauscher**[†]
[†]University of Hamburg, Germany [‡]University of Michigan, USA
{pranav.agrawal, anne.lauscher}@uni-hamburg.de, sstorks@umich.edu

Abstract

ACL removed the anonymity period for conference submissions in February 2024, allowing unrestricted preprinting during review. To examine how preprints and author recognition affect outcomes across institutional hierarchies, we track preprinting trends for 47k publications, survey 75 NLP researchers, interview 14 community members, and analyze 1.9k peer reviews. We observe that more elite institutions post preprints more frequently (52% vs. 36% by 2025). Most participants agree that preprinting gives these institutions an advantage in peer review, and indeed, reviewer knowledge of authors inflates scores at elite institutions ($d = 0.43$, $p < 0.001$) but not elsewhere, also lowering review quality. Nonetheless, the anonymity period was found largely ineffective; instead, underrepresented researchers emphasize struggles with visibility, review quality, and external structural barriers. To counteract these inequities, we make recommendations for review quality improvement and increasing investment in diversity initiatives that center the perspectives of affected communities.

1 Introduction

Peer review governs scientific publishing and career progression, but bias at various stages favors some authors over others (Smith et al., 2023). When reviewers learn author identities, they systematically favor established researchers and prestigious institutions (Tomkins et al., 2017; O'Connor et al., 2017; Kern-Goldberger et al., 2022). Publicly posted preprints compromise anonymity: over a third of reviewers search for preprints of papers they review (Rastogi et al., 2022), and papers with popular preprints receive higher scores and more acceptances (Bharadhwaj et al., 2020). These biases raise questions about whose work gets published and whose gets overlooked.

To protect double-blind review, ACL introduced an **anonymity period** in 2018 (ACL Admin Wiki, 2025b,a), which prohibited posting or discussing preprints from one month before submission until decision notification. Critics argued this disadvantaged early-career researchers, whose career advancement could be hindered by not being able to promote their work, or if similar work appeared online while theirs remained embargoed. After community discussion and a survey, ACL repealed the anonymity period in February 2024 (ARR, 2024), removing all restrictions on preprinting.

P4: *Anonymity didn't do much to protect underrepresented groups. We should focus on addressing D&I issues rather than implementing anonymity policies.*
P14: *Unless a paper is becoming very popular on social media, I don't think it has very large effects on how the review goes.*
P1: *The anonymity policy of a year ago was the worst of two options.*

Table 1: Quotes from interviewees on anonymity period effectiveness and alternative approaches.

The anonymity period was intended to protect underrepresented researchers from implicit bias. Whether it succeeded, and who benefits from its removal, remains unclear. Prior work establishes that bias exists in aggregate (Tomkins et al., 2017; Bharadhwaj et al., 2020; Rogers et al., 2023), but does not examine how preprint visibility creates differential advantages, nor does it capture marginalized researcher experiences. This paper uses mixed methods to address these gaps and examine whether the anonymity period achieved its intended protection. Specifically, our study examines two research questions:

RQ1: How do researchers across career stages, institutions, and regions perceive preprint culture and its equity implications? Debates about preprint policies rarely include early-career researchers or those from underrepresented institu-

Findings of the Association for Computational Linguistics: ACL 2026, pages 4566–4584

tions and regions, despite these groups being most affected. We survey 75 NLP researchers and interview 14 community members from a range of positionalities about their experiences with preprinting, self-promotion, and review fairness (§4).

RQ2: Do preprints and author recognition affect review outcomes differently across institutional and geographic tiers? If these signals benefit some researchers more than others, current practices compound existing inequities. We first establish that preprint adoption concentrates at higher-tier institutions and countries, creating baseline disparities before any reviewer bias enters (§3). We then analyze ARR reviews stratified by tier, preprint timing, and reviewer-reported author knowledge, measuring both scores and review quality (§5). We find that preprint presence and author recognition differentially affect outcomes by tier. Notably, reviewer knowledge of authors increases scores while decreasing review quality, suggesting that favourable treatment comes at the cost of rigorous feedback.

Finally, in §6 we synthesise our qualitative and quantitative findings into policy recommendations, centering the perspectives of researchers from underrepresented institutions and regions.

2 Background

We first review ACL's anonymity policies, then related work around preprinting and peer review.

ACL anonymity policy history. ACL conferences and TACL have long enforced double-blind peer review. Early online calls for papers made no mention of non-anonymized versions of papers appearing online (NAACL, 2001; HLT/EMNLP, 2005; ACL-IJCNLP, 2009). As arXiv gained popularity in computer science (Sutton and Gong, 2017), venues developed varied preprint policies: some requested authors refrain from posting submissions (NAACL, 2013; EACL, 2014), while others explicitly permitted preprints (EMNLP, 2015; ACL, 2016). ACL formalized the anonymity period in 2018 to protect review integrity and limit bias from reviewers recognizing authors or institutions (ACL Admin Wiki, 2025b,a). However, as the pace of NLP research accelerated, critics argued that this policy disadvantaged early-career researchers while established researchers with less pressure to publish quickly faced fewer downsides of preprinting. This resulted in removal of the anonymity period in February 2024 (ARR, 2024). Appendix A provides the full text for both changes.

Related work. Maintaining anonymity in double-blind review is difficult: while most reviewers cannot correctly guess author identities (Le Goues et al., 2018), prolific authors can be identified with up to 87% accuracy using text analysis (Caragea et al., 2019), and over a third of reviewers search for preprints of papers they review (Rastogi et al., 2022). Comparing single- and double-blind review, studies find significant bias toward well-known authors and institutions (Tomkins et al., 2017; O'Connor et al., 2017). Preprints amplify these effects: papers with popular preprints are more likely to be accepted, with less confident reviewers particularly susceptible to favoring well-known authors (Bharadhwaj et al., 2020). Beyond institutional prestige, substantial evidence links reviewer knowledge of author identity to gender bias (Kern-Goldberger et al., 2022). In the ACL community, Rogers et al. (2023) found that reviewers gave slightly higher scores to papers they thought they knew the authors of, and disproportionately recommended papers with preprints for best paper awards. This paper examines these biases in more detail, in the coming sections, we investigate how preprint culture interacts with institutional and geographic hierarchies to shape review scores, reviewer effort, and acceptance outcomes in NLP venues.

3 Statistics and Trends in Preprinting

First, we establish who posts preprints and when, and how this changed after the removal of the anonymity period. If preprinting behavior itself concentrates among certain researchers, this creates baseline inequities before any reviewer bias enters the picture. We analyze recent preprinting trends across institutional and geographic hierarchies, providing context for interpreting the review outcome disparities we document later.

3.1 Data Collection

To explore this, we collect 46,923 papers published in ACL venues from 2019–2025.[1] To examine trends by institution and country, we canonicalize author affiliations and identify institution countries through manual verification and automated matching. As a proxy for how elite authors' institutions are, we rank them by total publication count during

[1]Collected from ACL Anthology (aclanthology.org).

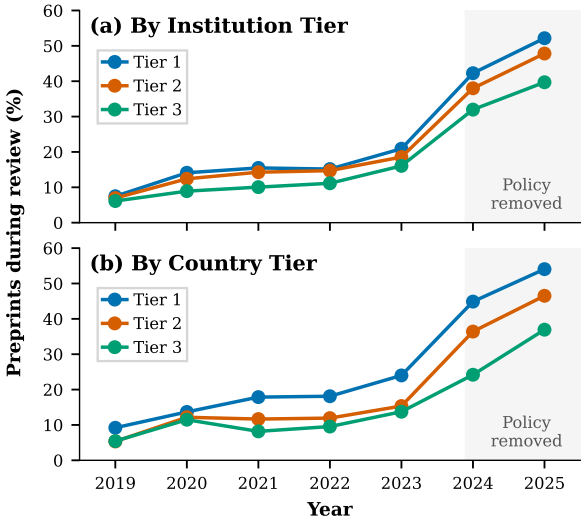

Figure 1: Preprints present during review by institutional and country tiers, 2019-2025. Anonymity period removal in February 2024 accelerated adoption, with Tier 1 institutions and countries showing highest rates.

this period, then stratify them into three **Institution Tiers**: Tier 1 (top 50), Tier 2 (ranks 51–250), and Tier 3 (251+). While publication count is imperfect as a measure of eliteness (e.g., it does not account for institution size), it is the most accessible and salient measure that can be directly applied to both academic and industry institutions.[2] To supplement this ranking with a notion of regional eliteness, we similarly stratify three **Country Tiers** by publication counts: Tier 1 (top 25%), Tier 2 (25–50%), and Tier 3 (bottom 50%). To determine whether a preprint of each paper was available during review, we search arXiv for matching titles and authors, then checking whether the preprint was published before the corresponding submission deadline. Papers with preprints posted before submission are classified as "Preprints Present During Review," while those without are "Preprints Absent During Review."

3.2 Findings

Figure 1 shows preprint timing rates from 2019–2025 based on institution and country tiers.

More prominent institutions show substantially greater rates of preprints present during review. Tier 1 institutions increased from 7.5% preprinting in 2019 to 52.2% in 2025, while Tier 3 increased from 7.9% to 35.8%, widening their gap from near-zero to 16.4%. Tier 2 tracked closer to Tier 1 (44.1% by 2025), suggesting early preprinting concentrates at higher-ranked institutions.

Countries with higher publication volumes also preprint more before and during review. Tier 1 countries increased from 9.2% preprinting in 2019 to 54.1% in 2025, while Tier 3 reached 36.9%, widening their gap from 3.8% to 17.2%. Tier 2 countries reached 46.5% by 2025, following similar patterns as institutions, i.e., where researchers from higher-publishing countries have preprints available during review at substantially greater rates.

Policy change accelerated early preprinting, but the community remains divided on whether to preprint before decisions. The anonymity period removal was followed by sharp increases across all tiers. Overall preprinting jumped from 19.9% in 2023 to 39.3% in 2024, reaching 47.8% in 2025. However, even at peak adoption, only 52.2% of Tier 1 papers had preprints during review in 2025. Tier 3 showed lower adoption at 35.8%. The community remains divided, with roughly half of papers posted during review and half not. This division persists across all tiers, though Tier 1 researchers preprint during review 15-17% more frequently than Tier 3.

These trends show that elite institutions and countries are adopting early preprinting at substantially higher rates, and the anonymity period removal widened the gap. Next, we will explore whether review outcomes compound advantages from this baseline disparity.

4 Community Surveys and Interviews

To gauge what those in the ACL community think about preprinting and peer review, we conducted surveys and interviews.[3] We first describe our methodology, then present unified insights.

4.1 Methodology

Survey. We designed an anonymous online survey examining NLP researchers' experiences with ACL's anonymity policy removal. Questions covered reviewer and author experiences, beliefs about institutional advantages, career impacts, and promotion strategies using Likert scales and free-response formats. We distributed the survey through community mailing lists (ACL Member Portal, Corpora-list), social media, and affinity

[2] Additionally, Szluka et al. (2023) find that publication counts correlate strongly with university ranking positions across major ranking systems (Spearman $\rho > 0.78$).

[3] Details, questions, and extended results in Appendix B.

		Institution			Country		
Finding	**All**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
Preprinting without restriction							
No impact	46.7	45.8	**50.0**	40.0	**47.9**	**50.0**	33.3
Impacts acceptance	20.0	20.8	5.0	**35.0**	**43.8**	35.7	**66.7**
Helps me	28.0	**33.3**	30.0	30.0	**33.3**	7.1	**33.3**
Hurts me	10.7	8.3	10.0	**15.0**	8.3	7.1	**25.0**
Preprints benefit elite institutions							
Strongly agree	33.3	29.2	25.0	**50.0**	29.2	**42.9**	41.7
Any agreement	76.0	70.8	80.0	**85.0**	79.2	64.3	**83.3**
Any disagreement	8.0	**12.5**	10.0	0.0	**8.3**	7.1	0.0
Field pace concerns about preprinting							
Too fast	58.7	**66.7**	55.0	50.0	**62.5**	57.1	50.0
Like sharing sooner	50.7	**58.3**	55.0	40.0	**56.2**	42.9	33.3
Quality declining	26.7	25.0	**40.0**	20.0	**33.3**	21.4	8.3
Overshadow my work	40.0	29.2	**50.0**	**50.0**	37.5	**50.0**	41.7
Self-promotion behavior							
No promotion	21.3	12.5	15.0	**40.0**	10.4	21.4	**58.3**
Before submission	17.3	**25.0**	15.0	15.0	**18.8**	14.3	16.7
After acceptance	56.0	54.2	**70.0**	40.0	**66.7**	64.3	16.7

Table 2: Survey responses by institution and country tier ($n = 75$). Bold indicates highest value within each stratification. Highlighted rows show consensus patterns.

group channels from September to November 2025. We received 75 responses spanning career stages (50.7% graduate students, 18.7% postdocs, 25.3% faculty, 9.3% industry or independent researchers), institution tiers (32% Tier 1, 27% Tier 2, 27% Tier 3), and country tiers (64% Tier 1, 18.7% Tier 2, 16% Tier 3). 59% of respondents participate in affinity groups for underrepresented communities.

Interviews. We recruited 14 participants from survey respondents, prioritizing diversity across geography and career stage. Participants came from North America (4), Europe (5), South Asia (2), Africa (2), and Latin America (1), and included PhD students (5), postdocs (2), and faculty (6). Semi-structured interviews occurred from September to November 2025, lasting 30–60 minutes each. Questions covered publishing contexts, anonymity policy impact, social media's role, and recommendations for supporting underrepresented researchers. With participant consent, we recorded and transcribed all interviews. Two authors iteratively discussed the transcripts and converged on recurring themes around anonymity, preprinting practices, and perceived inequities.

4.2 Findings

We summarize key survey results in Table 2, integrating them with interview responses below.

NLP researchers found the anonymity period ineffective. Nearly half (46.7%) of survey respondents believed that preprinting does not impact acceptance decisions for their papers; only 20.0% thought it did. This disconnect between recognizing systemic inequity and perceiving personal impact suggests the anonymity period was largely seen as ineffective. Perceived impact varied by institution tier: 35.0% of Tier 3 respondents reported that preprinting affected their acceptance outcomes, compared to lower rates at Tier 1 and 2 institutions. Similarly, while 28.0% overall claimed preprinting increased their acceptance chances, respondents from Tier 2 and 3 institutions and affinity group members were more likely to report the opposite.

Qualitatively, participants were mostly indifferent to or supportive of the anonymity period removal, highlighting its shortcomings. One early-career faculty (Tier 1 institution) argued it was *"the worst of the options,"* and another early-career faculty (Tier 3 country) noted that *"This policy didn't do much,"*, instead advocating for truly double-blind review or alternatives to level the playing field, such as mentorship programs for underrepresented researchers. One graduate student (Tier 1 institution) observed that *"at a better resourced institution, you have a better chance of working around the anonymity [period] policy"* by completing and preprinting papers one month before submission deadlines. Another graduate student (Tier 3 institution) found the policy reinforced *"academic isolation"* of under-resourced researchers, as *"visibility is crucial for establishing collaborations, drawing attention to our research, and building our network."* Similarly, an early-career faculty (Tier 3 country) remarked that *"This policy slowed down the progress of junior researchers"* by limiting visibility necessary for career advancement. One participant working on endangered languages described implicit bias tied to their research area:

> *I do sometimes feel like the reviews I get are people think I'm unintelligent... because they think I'm from that part of the world, because I'm writing about those kinds of languages.*

Survey data supports this pattern: researchers in areas like Machine Translation, Multilinguality, and Low-resource Methods report that preprints from others get more attention than their own published work at nearly twice the rate of those in areas like Safety/Alignment, Semantics, and NLP Applications (54% vs. 30%). Altogether, this show how the anonymity period failed to protect underrep-

resented and under-resourced researchers from inequities in peer review.

NLP researchers find preprint culture inequitable, but underrepresented groups feel this more strongly. Given the sharp increase in preprinting after the removal of the anonymity period, we explore community opinions about preprinting and anonymity. Among survey respondents, 90.7% reported encountering papers where they knew the authors or institution during review. This creates clear potential for bias, and 76.0% agreed that publicly sharing preprints benefits authors at well-resourced institutions more (33.3% strongly, 42.7% somewhat); only 8.0% disagreed. Perception of this advantage varies by position: respondents from Tier 3 institutions were most likely to strongly agree (50.0%), while those from Tier 1 institutions were most likely to disagree (12.5%) and least likely to strongly agree (29.2%). Similarly, 41.3% of affinity group members strongly agreed, compared to 20.7% of non-members.

In qualitative responses, several participants stood firmly against preprinting, stressing that *"anonymity is what protects underrepresented people"* and that *"the lack of anonymity perpetuates inequality."* One mid-career faculty member (Tier 1 institution) reported that the rise of preprinting made them lose faith in peer review entirely. Others were more optimistic, noting that reviewers in open venues can be held accountable and that public visibility builds credibility. A mid-career faculty from Latin America (Tier 3 institution) offered nuance: their positionality helped on *"Latin American topics,"* but invited skepticism on technical work because *"this is not the kind of work that some reviewers would expect a Latin American to be good at."* These findings show that researchers from underrepresented backgrounds perceive structural bias more acutely than those who benefit from institutional advantages. Policy discussions must center the voices of marginalised researchers.

The NLP community is divided on preprints leading to fast science. Survey responses on acceleration were divided: 50.7% value sharing work sooner, while 57.3% believe NLP research moves too fast. Tier 1 institutions showed the starkest tension: 66.7% were concerned about field pace, yet 58.3% valued early sharing. One mid-career faculty (Tier 1 country) expressed this tension: *"I am in favor of slower science... not optimizing for flag planting."* Another survey respondent (Tier 1 institution) went further: *"It makes me not want to publish. Why not just go straight to arXiv and skip peer review altogether?"* One respondent (Tier 2 institution) also noted that *"reviewers expect authors to comment on every new paper on arXiv,"* which violates ACL citation policy (ACL Admin Wiki, 2025b).

Underrepresented researchers find self-promotion on social media futile. Diving deeper into the problem of visibility, when asked how they promote papers, 56.0% of all respondents promote only after acceptance, 21.3% do not actively promote at all, and only 17.3% promote before submission. However, promotion behavior varies substantially by country development tier: 58.3% of researchers in Tier 3 countries do not actively promote their work, compared to 10.4% in Tier 1 countries (nearly a sixfold difference). Additionally, researchers in Tier 1 countries are far more likely to promote their work before or after acceptance (85.5% total) compared to those in Tier 3 countries (33.4% total). Pre-submission promotion also varies by research area: researchers in areas like Safety/Alignment, Semantics, and NLP Applications promote at three times the rate of those in Machine Translation, Multilinguality, and Low-resource Methods (22% vs. 7%). Therefore, researchers from underrepresented countries engage in far less self-promotion than others.

Participants described a *"rich-gets-reach"* dynamic where researchers at well-known institutions, or those whose papers are shared by visible colleagues, gain disproportionate preprint visibility. As one participant in West Africa explained, this creates *"a race for visibility, where those who already have a platform gain an even greater advantage."* A participant from North Africa added that fear of trolls, hostility, or discomfort with social media pose additional barriers to self-promotion, which yields little benefit for researchers at less-known institutions with small networks. One participant from Latin America emphasized the importance of grassroots, community-oriented local networks to mutually amplify the work of underrepresented researchers.

Underrepresented researchers recommend structural D&I changes. Qualitative responses revealed barriers beyond policy for researchers

outside Europe and North America. Several participants from the Global South emphasized the importance of journal publications for academic promotion, especially Scopus-indexed journals, which conflicts with NLP's conference-centric culture and increasing pace. As one participant from India explained, *"professors discouraged conference submissions in favor of journal papers for their promotion purposes."* One participant from Latin America recommended that *"ARR should pursue Scopus indexing so outputs 'count' in journal-first systems."* Further, one participant (tier 1 institution) highlighted that researchers at more prolific institutions can consult colleagues who have submitted work to ARR before, possibly making it easier to navigate procedures with implicit norms, e.g., for review response, reviewer issue reporting, resubmission, and commitment to conferences.

Compounding this, participants from both the Global South and North identified conference costs as a major barrier for researchers outside well-resourced labs. One participant from North Africa weighted this as more significant than anonymity policies, noting that *"conference costs (fees + travel) are a major barrier. There are big labs in Africa, but not in NLP."* Another participant from Pakistan highlighted a deeper tension. While acknowledging that for underrepresented researchers, *"their life prospects will be significantly improved"* by entering the NLP community, they questioned the ethics of inclusion itself:

> *"Do we want to pull them into that? Machine learning, and NLP is basically tools for surveillance... what kind of violence are we inflicting on them if we invite them to come into the community?"*

Any actions to mitigate inequities must account for these constraints beyond peer review policy.

5 Experiments on ARR Data

Our qualitative findings suggest that preprint visibility and author recognition systematically advantage researchers at elite institutions. To test whether these patterns appear in actual review outcomes, we analyze peer review data across stratified tiers.

5.1 Methodology

Data. We combine peer reviews from the ACL Rolling Review (ARR) dataset (Dycke et al., 2022) and NLPeer (Dycke et al., 2023), yielding 1,923 reviews spanning late 2023 (EMNLP cycle) through mid-2025 (ACL cycle). The dataset includes only accepted papers. This limits generalizability but provides a control: since all papers cleared the quality threshold, score differences reflect reviewer behavior rather than paper merit. For each paper, we determine preprint presence by searching arXiv for matching titles and checking posting dates against submission deadlines. Papers are classified as "preprint present" if a non-anonymous version was publicly available during review, and "preprint absent" otherwise. Reviewer knowledge of author identity is self-reported in review metadata.

Review quality metrics. Beyond reviews scores, we measure what makes reviews useful to authors using metrics from RevUtil (Sadallah et al., 2025): **Actionability** (concrete improvement suggestions), **Verifiability** (evidence supporting claims), and **Helpfulness** (overall improvement value). RevUtil uses an instruction-tuned Llama 3.1 8B (Grattafiori et al., 2024) trained on synthetically annotated reviews bootstrapped from human labels and validated with human labels, achieving comparable agreement to GPT-4o (OpenAI et al., 2024). We detect low-effort reviewing using LazyReviewPlus (Purkayastha et al., 2026), which annotates various categories of low-effort feedback in a similar dataset of ARR reviews, including non-specific comments, prescriptive demands, and model comparison requests. Specifically, we use an instruction-tuned Phi-4 model (Abdin et al., 2024) aligned with these annotations by Purkayastha et al.. From this model, we focus on **unclear comments** (claiming aspects of a paper are unclear without specific, actionable guidance) and **"compare to model X"** requests (demanding comparison to specific models without rationale), as these showed significant differences across tiers. Results for additional categories appear in Appendix D. We also report **review length** (word count) as a basic measure of review style and effort.

Statistical analysis. We conduct two types of comparisons. **Within-tier comparisons** test whether preprint presence or author recognition affects outcomes within each tier. For continuous outcomes (scores, quality metrics, review length), we use Welch's t-tests given near-normal score distributions and report Cohen's d as a measure of effect size (Cohen, 1988), where $d > 0.2$ in-

Tier	Overall	Length	Actionability	Verifiability	Helpfulness	Unclear%	Compare%
Panel A: Preprint Timing (present / absent during review)							
Inst. Tier 1	3.25/3.05*	368/399	3.56/3.71	4.30/4.43	4.13/4.32	49/58	13/15
Inst. Tier 2	3.32/3.08	358/392	3.72/3.76	4.34/4.07	**4.40**/4.17	42/46	**4/24***
Inst. Tier 3	3.29/3.20	390/378	3.75/3.57	**4.50/4.64**	**4.25**/4.17	**37**/54	**7/4**
Count. Tier 1	3.28/3.06*	370/392	3.62/3.73	4.34/4.38	4.20/4.22	44/55	11/17
Count. Tier 2	3.60/3.35	**551**/410	3.75/3.30	**5.00/4.60**	4.50/**4.70**	32/38	5/13
Count. Tier 3	**3.06**/3.25	**333**/382	3.67/3.43	4.33/4.43	4.19/4.29	38/**25**	8/0
Panel B: Reviewer Knowledge of Author (unknown / known)							
Inst. Tier 1	3.21/3.47*	368/430	3.64/3.57	4.35/4.41	4.20/4.16	51/42	13/13
Inst. Tier 2	3.21/3.26	355/374	3.72/**3.23**	4.32/**3.84**	**4.26**/4.23	47/47	11/**6**
Inst. Tier 3	3.12/3.22	**387**/396	3.72/**3.92**	4.35/4.50	4.19/**4.38**	49/**28***	11/11
Count. Tier 1	3.21/3.40*	368/409	3.68/3.55	4.35/4.32	4.20/4.19	50/39	13/13
Count. Tier 2	3.27/3.50	400/**562**	3.64/**4.00**	4.40/**5.00**	**4.45/4.00**	**37**/33	9/0
Count. Tier 3	**3.02**/3.12	366/422	3.64/3.80	4.29/4.00	4.16/**4.80**	45/50	9/0

Table 3: **Review outcomes by tier.** Each cell shows paired values: *present/absent* during review (Panel A) or *unknown/known* author (Panel B). Cell shading indicates effect size: $|d| \geq 0.5$, $0.3 \leq |d| < 0.5$, $0.2 \leq |d| < 0.3$. * = $p < 0.05$ within-pair. **Bold** = significant difference from Tier 1 within condition ($p < 0.05$). Overall = Overall Assessment (1–5); Length = word count; Actionability/Verifiability/Helpfulness (0–5); Unclear = % reviews with vague comments; Lazy Exp. = % reviews requesting additional experiments without justification.

dicates a small effect, $d > 0.5$ a medium effect, and $d > 0.8$ a large effect. For categorical outcomes (lazy pattern prevalence), we use χ^2 tests. We report p-values as measures of statistical significance: $p < 0.001$ indicates a *highly significant* effect, $p < 0.01$ indicates a *very significant* effect, $p < 0.05$ indicates a *significant* effect, and $p < 0.10$ indicates a *marginally significant* effect. **Cross-tier comparisons** test whether effects differ by institutional or geographic status. We fit ordinary least squares regression models with *interaction terms* (e.g., Tier × Preprint Presence) to detect whether preprint visibility benefits some tiers more than others. We report β coefficients indicating the magnitude of interaction effects and test whether these interactions are statistically significant. All p-values are false discovery rate-corrected (Benjamini and Hochberg, 1995). Table 3 presents the results; full regression outputs appear in Appendix E.

5.2 Findings

Papers with preprints from elite instituitions present during review receive significantly higher scores. Preprint presence during review has significantly higher scores for institution Tier 1 and country Tier 1 papers, confirming the widespread concerns of the NLP community. For institution Tier 1, papers with preprints present score 3.25 on overall score versus 3.05 when absent (Cohen's $d = 0.31, p < 0.01$); country Tier 1 shows an even larger gap of 3.28 versus 3.06 ($d = 0.34, p < 0.001$). Elite institutions gain *significantly* more from preprint visibility than lower-tier institutions ($\beta = 0.15, p = 0.02$).

Reviews for papers with preprints present contain more terse but specific feedback. For Tier 1 institutions, papers with preprints present during review receive shorter reviews (368 vs. 399 words, $d = -0.16$, $p < 0.10$) and less actionable feedback ($d = -0.15$, $p < 0.05$). However, these reviews contain fewer unclear comments ($d = -0.19$, $p < 0.10$). This pattern suggests that preprint presence signals a level of quality to reviewers, reducing both effort and vague criticism.

Author recognition inflates scores significantly, but only at elite institutions. When reviewers report knowing the author, Tier 1 institution papers receive a 0.26-point boost in overall assessment (3.47 vs. 3.21) ($d = 0.43, p < 0.001$). This effect does *not* appear at other tiers: Tier 2 and Tier 3 show *no significant* score differences when authors are known (all $p > 0.50$). Regression models with Tier × Knowledge interactions confirm this concentration *significantly* ($\beta = 0.19, p = 0.018$).

Reviewers demand expensive model comparisons significantly more often when preprints are absent. Papers from mid-tier institutions face sharply increased demands for comparisons to expensive models when preprints are absent during

review, in line with common complaints reported in Section 4. For institution Tier 2, "compare to model X" requests jump from 4% when preprints are present to 24% when absent ($d = 0.71, p < 0.01$). Country Tier 1 shows a similar pattern: 11% with preprints versus 17% without ($p < 0.05$).

Authors from low-tier institutions get higher quality reviews if reviewers know authors. Unlike elite institutions where recognition inflates scores but not quality, lower tiers show a different pattern. When reviewers know authors from institution Tier 3, actionability rises from 3.72 to 3.92 ($p < 0.05$), unclear comments drop *significantly* from 49% to 28% ($p < 0.05$), and extra experiment demands drop from 12% to 0% ($d = -0.37, p < 0.05$). In country Tier 3, helpfulness similarly rises from 4.16 to 4.80 when reviewers know authors ($d = 0.64, p < 0.05$).

Country Tier 3 shows reversed score patterns: papers without preprints receive higher scores. Unlike Tiers 1 and 2, papers from Tier 3 countries receive higher overall assessment scores when reviewers lack preprint access (3.25 vs. 3.06, $d = 0.19$). Though not statistically significant ($p = 0.83$), this pattern aligns with concerns expressed by Tier 3 researchers that posting preprints may disadvantage them (§4). Papers without preprints also receive longer reviews (382 vs. 333 words) and fewer unclear comments (25% vs. 38%).

6 Discussion and Recommendations

The NLP community largely agrees that early preprinting advantages elite institutions when anonymity is compromised, and our analysis empirically confirms this assumption: When reviewers identified Tier 1 authors, their overall scores increased *significantly*, ($p < 0.001$), while Tier 2 and Tier 3 showed no such effect (all $p > 0.50$). Meanwhile, preprint visibility degraded review quality for elite institutions: Tier 1 papers received *marginally* worse reviews when preprints were present. Anonymity thus shields well-resourced researchers from weaker reviews rather than protecting less-resourced researchers from unfair biases. Our qualitative findings point to the same conclusion. The majority of survey and interview participants believed that the anonymity period's removal had no impact on their careers. Both quantitative and qualitative evidence thus converge: **the anonymity period policy did not protect underrepresented researchers**. We recommend two policy directions that directly address the compounding inequities of preprint culture – improving review systems opting for higher quality and investing more in diversity and inclusion measures.

R1: Improving review quality. Participants across tiers reported declining review quality, describing reviews as "random," fixated on minor details, or appearing LLM-generated. Some problems disproportionately affect under-resourced and underrepresented researchers: reviewers request large-scale experiments or comparisons to closed models despite ARR guidelines discouraging such requests (ARR, 2025), and researchers in niche areas receive mismatched reviewers. Although ARR has made efforts to increase reviewer pools and refine guidelines, including policies for recognition of great reviews and reporting low-quality reviews (ARR, 2025), authors can only report violations after the fact and hope area chairs intervene. Recent works explore improving review aggregation and increasing reviewer participation (Kuznetsov et al., 2024) and improving review matching (Thorn Jakobsen and Rogers, 2022). Mentoring junior reviewers leads to better engagement and higher review quality, as shown at ICML 2020 (Stelmakh et al., 2020). Consistently implementing outstanding reviewer awards, possibly with discounted registration, may also incentivize reviewing quality (Shah, 2022).

R2: Increasing D&I investment. While the anonymity period was partly intended to protect underrepresented researchers, participants consistently emphasized that direct D&I investment matters more than anonymity restrictions. Participants cited several barriers beyond anonymity: conference affordability, journal-centric academic contexts common in the Global South, visibility gaps for researchers at lesser-known institutions, and lack of procedural knowledge about ARR. We make several recommendations to address these barriers. To make conferences more affordable for researchers with fewer resources, first, ACL conferences could adopt tiered pricing models like the ACM (ACM, 2025). Second, more targeted efforts for gaining D&I sponsorship by companies could be made – the number of D&I sponsors has dwindled from eight in ACL (2022) to one in ACL (2024) and ACL (2025). Third, to address visibility gaps, ACL could amplify D&I awardees through its social media channels and invest in more grassroots organizations like RAF

(2024). Fourth, ACL venues could create more publication opportunities for affinity groups; NAACL (2022) remains the only recent venue to invite affinity workshops. Fifth, they could focus more on creating research tracks aligned with underrepresented researchers' interests, as NAACL (2024) did with its theme selection. Sixth, ACL could pursue formal Scopus indexing for additional conference proceedings (only consistently done for RANLP (2025) (Scopus, 2025a,b)), enabling researchers in journal-prioritized contexts to benefit from publishing at ACL conferences. Seventh, for authors lacking institutional networks, ACL could establish mentoring programs with experienced researchers throughout the ARR process; volunteer mentors could be incentivized through reduced reviewing loads. Finally, D&I committees could be involved in conference planning before decisions about budgets, venues, and accessibility are finalized (ACL, 2025), and ACL could expand board and committee representation from researchers at non-elite institutions and underrepresented countries.

Conclusion. ACL introduced the anonymity period in 2018 in good faith, responding to documented review biases favoring well-resourced researchers. NLP has since grown rapidly, and preprinting has become standard practice, making multi-month embargoes impractical. Our mixed-methods study finds that anonymity did reduce some of the preprint and self-promotion advantages available to well-resourced researchers. However, this study also finds no evidence that the anonymity period achieved its intended goal of protecting underrepresented researchers; the biases persisted regardless. **We call on ACL to center marginalized researchers in future policy, with concrete action on publication norms, review reforms that protect under-resourced authors, and increased investment in diversity and inclusion.**

Limitations

This study has several limitations. Our quantitative analysis uses ARR data that includes only authors and reviewers who consented to share their reviews, and authors who received poor reviews or reviewers who lacked confidence in their assessments may have opted out. The analysis also depends on authors self-reporting preprint existence and reviewers self-reporting preprint knowledge, both subject to reporting bias. We restricted the sample to accepted papers to control for paper quality and isolate reviewer perception of work that cleared an acceptance threshold (Section 5.1). This narrows the range of outcomes we observe and rules out analysis of acceptance rates or factors behind rejection, both of which matter for understanding inequities in ACL peer review. Characterizing rejected papers systematically is also difficult: ARR does not publish acceptance decisions, and a paper may receive ARR reviews without being committed to an ACL venue. Our review quality metrics rely on LLMs as judges, which are less reliable than expert human annotation. Sadallah et al. (2025) and Purkayastha et al. (2026) evaluated the specific models we use against human annotations on much of the same ARR data: the RevUtil model matches or exceeds GPT-4o (OpenAI et al., 2024) on human agreement, with 0.3 to 0.5 quadratic-weighted Cohen's Kappa (κ^2; Cohen, 1968), and LazyReviewPlus gets 0.51 as the $F_{0.5}$ score on predicting human-labeled review issues. Absolute scores from these models should be read with caution, but relative comparisons across groups remain informative. Finally, we did not recruit early-career researchers for interviews: the anonymity period was repealed before they entered the field, so they have no first-hand experience with the policy change we study.

Ethical Considerations

This research received ethical approval from the Ethics Committee of the University of Hamburg Business School and from the University of Michigan eResearch (ID HUM00283678). Survey and interview participants were not compensated and provided informed consent prior to participation. Interview participants additionally consented to the use of direct anonymous quotes. We do not include identifiable information about participants; all quotes are attributed using pseudonyms. Further details on consent procedures are provided in Appendix B. The peer review data used in our quantitative analysis comes from publicly available datasets where authors and reviewers consented to share their data. Consent to share data may itself correlate with review outcomes, which we discuss in Limitations.

Acknowledgments

We thank our anonymous reviewers for their feedback. We thank the participants of our study for sharing their perspectives with us. The work of Anne Lauscher and Pranav A is funded under the

Excellence Strategy of the German Federal Government and States.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

ACL. 2025. Diversity and inclusion chair handbook.

Organizing Committee of ACL. 2016. Call for papers: ACL 2016.

Organizing Committee of ACL. 2022. ACL 2022 sponsors.

Organizing Committee of ACL. 2024. Sponsors.

Organizing Committee of ACL. 2025. Sponsors.

ACL Admin Wiki. 2025a. ACL author guidelines.

ACL Admin Wiki. 2025b. ACL policies for review and citation.

Organizing Committee of ACL-IJCNLP. 2009. ACL-IJCNLP 2009 2nd call for papers.

ACM. 2025. Policy on geographic apc waivers and discounts.

ARR. 2024. Update to anonymity policy.

ARR. 2025. ARR reviewer guidelines.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300.

Homanga Bharadhwaj, Dylan Turpin, Animesh Garg, and Ashton Anderson. 2020. De-anonymization of authors through arXiv submissions during double-blind review. *Preprint*, arXiv:2007.00177.

Cornelia Caragea, Ana Uban, and Liviu P. Dinu. 2019. The myth of double-blind review revisited: ACL vs. EMNLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2317–2327, Hong Kong, China. Association for Computational Linguistics.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

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

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Yes-yes-yes: Proactive data collection for ACL rolling review and beyond. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 300–318, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Organizing Committee of EACL. 2014. Call for papers.

Organizing Committee of EMNLP. 2015. Call for papers for EMNLP 2015.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Organizing Committee of HLT/EMNLP. 2005. HLT/EMNLP 2005 call for papers.

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

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, and 5 others. 2024. What can natural language processing do for peer review? *Preprint*, arXiv:2405.06563.

C. Le Goues, Y. Brun, S. Apel, E. Berger, S. Khurshid, and Y. Smaragdakis. 2018. Effectiveness of anonymization in double-blind review. *Commun. ACM*, 61(6):30–33.

NAACL. 2022. Affinity group workshops.

NAACL. 2024. NAACL 2024 theme track: Languages of Latin America.

Organizing Committee of NAACL. 2001. NAACL 2001 call for papers.

Organizing Committee of NAACL. 2013. Call for papers for NAACL HLT 2013.

E. E. O'Connor, M. Cousar, J. A. Lentini, M. Castillo, K. Halm, and T. A. Zeffiro. 2017. Efficacy of Double-Blind Peer Review in an Imaging Subspecialty Journal. *AJNR. American journal of neuroradiology*, 38(2):230–235.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Alek-

sander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

Sukannya Purkayastha, Qile Wan, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. 2026. Reviewing the reviewer: Elevating peer review quality through llm-guided feedback. In *Findings of the Association for Computational Linguistics: ACL 2026*.

RAF. 2024. NAACL Regional Americas Fund.

Organizing Committee of RANLP. 2025. Final call for papers, RANLP 2025.

Charvi Rastogi, Ivan Stelmakh, Xinwei Shen, Marina Meila, Federico Echenique, Shuchi Chawla, and Nihar B. Shah. 2022. To arXiv or not to arXiv: A study quantifying pros and cons of posting preprints online. *Preprint*, arXiv:2203.17259.

Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Program chairs' report on peer review at ACL 2023. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages xl–lxxv, Toronto, Canada. Association for Computational Linguistics.

Abdelrahman Sadallah, Tim Baumgärtner, Iryna Gurevych, and Ted Briscoe. 2025. The good, the bad and the constructive: Automatically measuring peer review's utility for authors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28979–29009, Suzhou, China. Association for Computational Linguistics.

Scopus. 2025a. International conference recent advances in natural language processing, RANLP.

Scopus. 2025b. Proceedings of the annual meeting of the Association for Computational Linguistics.

Nihar B Shah. 2022. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87.

Olivia M Smith, Kayla L Davis, Riley B Pizza, Robin Waterman, Kara C Dobson, Brianna Foster, Julie C Jarvey, Leonard N Jones, Wendy Leuenberger, Nan Nourn, and 1 others. 2023. Peer review perpetuates barriers for historically excluded groups. *Nature Ecology & Evolution*, 7(4):512–523.

Ivan Stelmakh, Nihar B. Shah, Aarti Singh, and Hal Daumé III. 2020. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. *AAAI-21*.

Charles Sutton and Linan Gong. 2017. Popularity of arXiv.org within computer science. *Preprint*, arXiv:1710.05225.

Péter Szluka, Edit Csajbók, and Balázs Győrffy. 2023. Relationship between bibliometric indicators and university ranking positions. *Scientific Reports*, 13(1):14193.

Terne Thorn Jakobsen and Anna Rogers. 2022. What factors should paper-reviewer assignments rely on? community perspectives on issues and ideals in conference peer-review. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4810–4823, Seattle, United States. Association for Computational Linguistics.

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

Transactions of the Association for Computational Linguistics. 2024. Submissions.

A Anonymity Policy Language

A.1 Introduction of Anonymity Period

ACL Admin Wiki (2025a) lists the following detailed policy for the anonymity period, first introduced in 2018 (ACL Admin Wiki, 2025b):

> The following rules and guidelines are meant to protect the integrity of double-blind review and ensure that submissions are reviewed fairly. The rules make reference to the anonymity period, which runs from 1 month before the submission deadline up to the date when your paper is either accepted, rejected, or withdrawn.
>
> - You *may not* make a non-anonymized version of your paper available online to the general community (for example, via a preprint server) *during* the anonymity period. By a version of a paper we understand another paper having essentially the same scientific content but possibly differing in minor details (including title and structure) and/or in length (e.g., an abstract is a version of the paper that it summarizes).
> - If you have posted a non-anonymized version of your paper online before the start of the anonymity period, you *may* submit an anonymized version to the conference. The submitted version must not refer to the non-anonymized version, and you must inform the program chair(s) that a non-anonymized version exists. You may not update the non-anonymized version during the anonymity period, and we ask you not to advertise it on social media or take other actions that would further compromise double-blind reviewing during the anonymity period.
> - Note that, while you are not prohibited from making a non-anonymous version available online before the start of the anonymity period, this does make double-blind reviewing more difficult to maintain, and we therefore encourage you to wait until the end of the anonymity period if possible. Alternatively, you may consider submitting your work to the Computational Linguistics journal, which does not require anonymization and has a track for "short" (i.e., conference-length) papers.
>
> The notion of *preprint* is understood broadly to refer to any non-refereed paper posted online, including but not limited to preprint servers such as arXiv. Note that the rule applies only to preprints that authors post themselves, so it does not apply to (say) non-refereed proceedings volumes. The restriction on updating is to prevent authors from circumventing these rules by "flag planting" with a placeholder version over 1 month in advance.

A.2 Ending of Anonymity Period

ARR (2024) updates the anonymity policy for ARR (which mediates most ACL conference venues) by repealing the anonymity period, with Transactions of ACL (TACL) following suit (Transactions of the Association for Computational Linguistics, 2024):

> The ACL has adopted a new anonymity policy effective for all future submissions, including to ARR. This new policy replaces the old policy that prohibited authors from posting or advertising non-anonymous preprints during a period starting one month before the submission or commitment deadline, and continuing while the submission is under review.
>
> The new policy takes effect in ARR beginning with the Feb. 15 submission and commitment cycles. Under the new policy, submissions will remain anonymous during peer review, but authors are free to post and discuss non-anonymous preprints at any time. To protect anonymity during peer review, ARR will take measures to prioritize reviews by reviewers who are not aware of the

author identities. Authors are reminded that widely sharing the work will make it harder to recruit reviewers. (Venues will also institute awards for unpublicized work. These awards are decided by program committees, not ARR.)

B Survey and Interview Details

In this appendix, we first present the list of questions and informed consent information for the survey and interview aspects of our study. We then present an extended version of Table 2 with stratification by membership in affinity groups and information about participants' experiences as reviewers.

B.1 Survey

Our survey was promoted through the field-wide Corpora List (https://list.elra.info/mailman3/hyperkitty/list/corpora@list.elra.info/), and mailing lists and social channels for various affinity groups and EquiCL (https://equicl.github.io/).

Informed consent. Before participants began the survey, we provided the following details to them (some information omitted for anonymity):

- We are ... conducting a study on the impact of ACL's anonymity period policy on the computational linguistics research community.
- **Who should participate:** This survey is intended for researchers who publish in and/or review for ACL, NAACL, EACL, EMNLP, and other *CL conferences.
- **Time commitment:** This survey will take approximately 10-12 minutes to complete.
- **Data privacy:** All responses are completely anonymous and will be analyzed in aggregate. No personally identifiable information will be collected or reported. Your participation is voluntary, and you may skip any questions you prefer not to answer.
- **Purpose:** The data collected will be used exclusively for academic research to understand how current anonymity policies affect different members of our community, with the goal of informing future policy discussions.
- **Questions?** If you have any questions about this survey, or if you want to change/delete the data, please contact [us].
- When you submit this form, it will not automatically collect your details like name and email address unless you provide it yourself.

Participants were required to indicate that "I have read the above information and agree to participate in this survey. I understand that I can withdraw my data anytime by contacting the authors."

Survey questions. Our survey included the following questions (extra details about form structure are italicized):

1. **What is your current career position?** *(Check all that apply)*
   - Undergraduate student
   - Graduate student (Master's)
   - Graduate student (PhD)
   - Postdoctoral researcher
   - Faculty - Early career (Assistant Professor or equivalent)
   - Faculty - Mid career (Associate Professor or equivalent)
   - Faculty - Senior (Full Professor or equivalent)
   - Industry researcher (Junior vs Senior)
   - Government/Non-profit researcher
   - Independent researcher
   - Other *(specify in blank)*
2. **What is your primary country of work/study?**
3. **What is your country of origin?** Write "mixed" if the answer is too complex.
4. **What institution are you primarily affiliated with?**
5. **Which communities do you actively participate in?** (Check all that apply)
   - Black in AI
   - Disability in AI
   - Ethio NLP
   - Indigenous AI

		Institution			Country			Affinity Group	
Finding	**All**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**Yes**	**No**
Perceived impact of preprinting on own papers									
No impact on reviews/decisions	46.1	45.8	**50.0**	38.1	**47.9**	**50.0**	33.3	**48.9**	41.4
Impacts acceptance	36.8	33.3	40.0	**47.6**	35.4	21.4	**58.3**	**44.7**	24.1
Preprinting helps me	28.9	**33.3**	30.0	**33.3**	**33.3**	7.1	**33.3**	**34.0**	20.7
Preprinting hurts me	11.8	8.3	10.0	**19.0**	8.3	7.1	**25.0**	**17.0**	3.4
Preprints benefit well-resourced institutions									
Strongly agree	34.2	29.2	25.0	**52.4**	29.2	**42.9**	41.7	**42.6**	20.7
Any agreement	76.3	70.8	80.0	**85.7**	79.2	64.3	**83.3**	76.6	75.9
Any disagreement	7.9	**12.5**	10.0	0.0	**8.3**	7.1	0.0	**8.5**	6.9
Reviewer experiences with deanonymization									
Anonymity compromised ($\geq$1 time)	57.9	58.3	**65.0**	47.6	56.2	**64.3**	58.3	51.1	**69.0**
Anonymity never compromised	26.3	**29.2**	25.0	28.6	**31.2**	14.3	25.0	**31.9**	17.2
Learned author identity	52.6	45.8	**65.0**	38.1	54.2	42.9	**58.3**	46.8	**62.1**
Learned institution	46.1	**50.0**	**50.0**	33.3	45.8	**50.0**	41.7	38.3	**58.6**
Learned social media praise	14.5	**25.0**	10.0	14.3	14.6	7.1	**16.7**	**19.1**	6.9
Learned social media criticism	9.2	8.3	**15.0**	9.5	6.2	7.1	**16.7**	**10.6**	6.9
Never served as reviewer	14.5	8.3	10.0	**23.8**	10.4	**21.4**	16.7	**17.0**	10.3
Field pace and preprint concerns									
Everything moving too fast	59.2	**66.7**	55.0	52.4	**62.5**	57.1	50.0	55.3	**65.5**
Like sharing work sooner	51.3	**58.3**	55.0	42.9	**56.2**	42.9	33.3	51.1	51.7
Preprint quality declining	26.3	25.0	**40.0**	19.0	**33.3**	21.4	8.3	**29.8**	20.7
Others' preprints overshadow my work	40.8	29.2	**50.0**	**52.4**	37.5	**50.0**	41.7	**46.8**	31.0
Self-promotion behavior									
No promotion	21.1	12.5	15.0	**38.1**	10.4	21.4	**58.3**	**23.4**	17.2
Promote before submission	18.4	**25.0**	15.0	19.0	**18.8**	14.3	16.7	**21.3**	13.8
Promote after acceptance	56.6	54.2	**70.0**	38.1	**66.7**	64.3	16.7	53.2	**62.1**
Number of respondents (n)	76	24	20	21	48	14	12	47	29

Table 4: Survey responses by institution tier, country tier, and affinity group membership. All values are percentages. Bold indicates highest value within each stratification (institution, country, or affinity group). Highlighted rows show consensus patterns across groups. Institution tier responses sum to 65 due to 11 respondents without institutional affiliation data. Country tier responses sum to 74 due to 2 missing values.

- LatinX in AI
- Masakhane/AfricaNLP
- Muslims in ML
- North Africans in NLP
- Queer in AI
- SomosNLP
- WiNLP (Widening NLP)
- Women in ML
- Other *(specify in blank)*

6. **What are your primary NLP research areas?** (Check up to 3) [List of standard NLP areas]
   - Safety and Alignment in LLMs
   - AI/LLM Agents
   - Human-AI Interaction/Cooperation
   - Retrieval-Augmented Language Models
   - Mathematical, Symbolic, and Logical Reasoning in NLP
   - Computational Social Science, Cultural Analytics, and NLP for Social Good
   - Code Models
   - Interpretability, Model Editing, Transparency, and Explainability
   - LLM Efficiency
   - Generalizability and Transfer
   - Dialogue and Interactive Systems
   - Discourse, Pragmatics, and Reasoning
   - Low-resource Methods for NLP

- Ethics, Bias, and Fairness
- Natural Language Generation
- Information Extraction and Retrieval
- Linguistic theories, Cognitive Modeling and Psycholinguistics
- Machine Translation
- Multilinguality and Language Diversity
- Multimodality and Language Grounding to Vision, Robotics and Beyond
- Neurosymbolic approaches to NLP
- Phonology, Morphology and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech Processing and Spoken Language Understanding
- Summarization
- Hierarchical Structure Prediction, Syntax, and Parsing
- NLP Applications
- Other *(specify in blank)*

7. **As a reviewer, have you encountered situations where paper anonymity was compromised through preprints or social media?** *(Select one)*
   - Yes, multiple times (5+)
   - Yes, a few times (2-4)
   - Yes, once
   - No, never
   - I haven't served as a reviewer
8. **As a reviewer, what types of prior knowledge have preprints for papers you reviewed given you?** *(Check all that apply)*
   - Knowing the authors of the paper
   - Knowing the institution of the paper
   - Knowing the criticisms of the paper from social media
   - Knowing the praise of the paper from social media
   - I haven't served as a reviewer
   - Other *(specify in blank)*
9. **How has removal of ACL's anonymity period policy influenced your career?** *(Check all that apply)*
   - I like that I can share my work sooner (and not get scooped)
   - I feel like pre-prints from other researchers get more attention than my published work
   - I feel like the quality of the pre-prints has gone down
   - Everything is moving too fast
   - I think (my/other people's) pre-prints have (hurt/helped) the acceptance decisions on my papers under review
   - Other *(specify in blank)*
10. **How strongly do you agree with this statement?** Publicly sharing preprints (e.g., on arXiv / social media) prior to the review process benefits authors at well-resourced institutions more than it benefits authors at not well-resourced institutions (e.g., in Global South, lesser-known institutions). *(Select one)*
    - Strongly agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Strongly disagree
11. **How do you think ACL's anonymity period policy has influenced the reviews, scores, and acceptance decisions for your papers?** *(Check all that apply)*
    - Preprinting my paper makes it more likely to be reviewed positively and accepted

- Preprinting my paper makes it more likely to be reviewed negatively and rejected
- Preprinting my paper has no impact on its reviews and decisions
- Not preprinting my paper makes it more likely to be reviewed positively and accepted
- Not preprinting my paper makes it more likely to be reviewed negatively and rejected
- Not preprinting my paper has no impact on its reviews and decisions
- Other *(specify in blank)*

12. **How do you typically promote your papers?** *(Select one)*
    - Posting on social media / disseminating your work (via workshops and seminars) before submission
    - Posting on social media / disseminating your work (via workshops and seminars) only after acceptance
    - I don't actively promote my work
    - Other *(specify in blank)*
13. **Thank you for filling out the survey. We are interested in interviewing some survey respondents. If you are willing to be contacted in this regard, please enter your name and email.**

B.2 Interview

Interview participants were selected from survey respondents, and through direct contact of known community leaders and vocal community members.

Informed consent. Before each interview began, participants were informed about the following (in writing and/or conversation):

- The purpose of the study
- The study's IRB approval status
- What interview recordings would be used for (i.e., direct anonymous quotes only)
- The semi-structured format of the interview, and expectation that any questions asked were optional to answer

Participants were asked to confirm their consent vocally, and given an opportunity to ask questions.

Interview questions. In our interviews, we asked participants the following questions:

1. How does publishing and promotion work in your country? Do ACL publications often count for promotion?
2. The anonymity period policy (which restricted preprints during review until February 2024) has been debated extensively. In your view, did this policy ultimately help or harm researchers?
3. Now that researchers can post preprints anytime, do you think this change disproportionately benefits authors from prestigious institutions?
4. What role do you think social media plays in paper visibility and review outcomes?
5. What concrete changes could ARR make to better support underrepresented researchers?
6. Is there anything else about peer review, publishing culture, or ACL policies you'd like to discuss?

All questions were additionally sent to participants in writing for clarity. Interviews were conducted online through video call, except for one participant who preferred to be interviewed by email. At times, conversations went beyond the above questions, and some quotes in the paper may come from these cases.

B.3 Extended Results

Table 4 extends the results in Table 2 to include some aspects briefly mentioned in Section 4. Specifically, it includes stratification by membership in affinity groups, and additional questions about participants' experiences as reviewers in ACL venues.

C Details regarding Tier Processing

C.1 Data Preprocessing

We collect papers published in ACL venues from 2019–2025 using the `acl-anthology` library. We retrieve paper IDs, titles, author strings, and PDF

links. We run GROBID to convert PDFs into TEI XML format. We parse TEI files to extract paper identifiers, titles, authors, and institutional affiliations by reading `orgName` elements of type `institution`. When multiple affiliations appear, we retain all, deduplicate repeated strings, and join them into a semicolon-separated field per paper.

We canonicalize author affiliations and identify institution countries through manual verification and automated matching. We rank institutions by total publication count during this period as a proxy for prominence. We stratify institutions into three tiers: Tier 1 (top 50 by publication count), Tier 2 (ranks 51–250), and Tier 3 (ranks 251+). We stratify countries into three tiers by publication count: Tier 1 (top 25%), Tier 2 (25–50%), and Tier 3 (bottom 50%).

We obtain arXiv posting dates by querying arXiv with title- and author-based matching. We compare each posting date to the corresponding conference submission deadline. Papers with preprints posted before submission are classified as "Preprints Present During Review." Papers without preprints or with preprints posted after submission are classified as "Preprints Absent During Review."

C.2 Conferences and Years Included

Our dataset includes reviews from the following ACL venues and years:

- ACL: 2019, 2020, 2021, 2022, 2023, 2024, 2025
- COLING: 2020, 2022, 2024
- CoNLL: 2019, 2020, 2021
- EACL: 2021, 2023, 2024
- EMNLP: 2019, 2020, 2021, 2022, 2023, 2024, 2025
- NAACL: 2019, 2021, 2022, 2023, 2025

C.3 Tier Classification Examples

Institutions and countries are classified into three tiers based on publication volume in NLP venues. This classification does not correlate with institutional prestige or economic development.

Country Tier Examples:

- **Tier 1:** United States, China, United Kingdom, Germany, Canada, India, South Korea, Japan, Hong Kong, Singapore
- **Tier 2:** Qatar, Iran, Bangladesh, Saudi Arabia, Poland, Greece, Romania, Norway, Turkey, Brazil
- **Tier 3:** Malta, Luxembourg, Philippines, Nigeria, Peru, Lebanon, Malaysia, Sri Lanka, Uruguay

Institution Tier Examples:

- **Tier 1:** Carnegie Mellon University, University of California, Tsinghua University, Peking University, University of Washington, University of Edinburgh, University of Cambridge, Stanford University, Johns Hopkins University, National University of Singapore
- **Tier 2:** Singapore Management University, Cardiff University, Korea Advanced Institute of Science and Technology, University of Illinois at Chicago, King's College London, George Mason University, KU Leuven, Hamad Bin Khalifa University, University of Hong Kong
- **Tier 3:** Grenoble INP, Le Mans University, National Yang Ming Chiao Tung University, Syracuse University, University of Athens, University of Sussex, Zayed University, Bangladesh University of Engineering and Technology, Boise State University, Middle East Technical University

D LazyReviewPlus Analysis Details

D.1 LazyReviewPlus Methodology

The LazyReviewPlus (Purkayastha et al., 2026) model we used in this work (referred to as LazyReviewPlus for brevity) detects low-effort reviewing patterns that compromise review quality. The tool uses an instruction-tuned Phi-4 model aligned with human annotations to identify four categories of lazy review patterns:

1. **Unclear comments**: Vague criticism lacking actionable guidance (e.g., "the writing needs improvement" without specifying what or how).
2. **"Should do X instead"**: Prescriptive demands for alternative approaches without justifying why the current approach is inadequate.
3. **"Compare to model X"**: Requests for comparison to specific models (often expensive

Tier	Any Lazy%	Unclear%	Should Do%	Compare%
INSTITUTION TIER				
1	66/70	42/51	16/19	13/13
2	72/72	47/47	17/18	6/11
3	**50**/67	**28**/49*	22/14	11/11
COUNTRY TIER				
1	66/70	39/50	18/19	13/13
2	67/69	33/37	**0**/19	0/9
3	67/66	50/45	17/17	0/9

Table 5: **Effect of author knowledge on lazy patterns.** Each cell shows paired values: *known/unknown* author. Cell shading indicates effect size: $|d| \geq 0.5$, $0.3 \leq |d| < 0.5$, $0.2 \leq |d| < 0.3$. * = $p < 0.05$ within-pair. **Bold** = notable pattern reversal or large difference. Any Lazy = % reviews with any lazy pattern; Unclear = % with unclear comments; Should Do = % with "should do instead" requests; Compare = % with "compare to model" requests.

Tier	Any Lazy%	Unclear%	Should Do%	Compare%
INSTITUTION TIER				
1	69/73	49/58	17/20	13/15
2	68/68	42/46	18/22	**4/24***
3	57/67	**37**/54	13/21	7/4
COUNTRY TIER				
1	67/72	44/55	18/19	11/17
2	63/63	32/38	**11/38***	5/13
3	63/63	38/25	**8/38***	8/0

Table 6: **Effect of preprint timing on lazy patterns.** Each cell shows paired values: *before/after* submission deadline. Cell shading indicates effect size: $|d| \geq 0.5$, $0.3 \leq |d| < 0.5$, $0.2 \leq |d| < 0.3$. * = $p < 0.05$ within-pair. **Bold** = notable pattern reversal or large difference. Any Lazy = % reviews with any lazy pattern; Unclear = % with unclear comments; Should Do = % with "should do instead" requests; Compare = % with "compare to model" requests.

closed-source systems) without clear rationale for why such comparison is necessary.

4. **Extra experiment demands**: Requests for additional experiments without explaining how they would strengthen the paper's claims.

We apply LazyReviewPlus to all 1,267 reviews in our dataset with sufficient text content. For each review, we record the count of each lazy pattern type and a binary indicator for whether any lazy pattern is present. Across all reviews, the mean lazy pattern count is 0.93 (SD = 0.78), with 69% of reviews containing at least one lazy pattern.

D.2 Effect of Author Knowledge on Lazy Patterns

Table 5 presents the effect of reviewer knowledge of author identity on lazy patterns. We report the percentage of reviews containing each pattern type, comparing cases where reviewers reported knowing the author versus not knowing. Effect sizes (Cohen's d) indicate the magnitude of differences.

When reviewers know authors from Tier 3 institutions, lazy patterns decrease substantially. The prevalence of any lazy pattern drops from 67% to 50% ($d = -0.36$), unclear comments fall from 49% to 28% ($d = -0.43$, $p < 0.05$), and extra experiment demands disappear entirely (0% vs. 12%, $d = -0.37$). This pattern suggests reviewers invest more effort in constructive feedback when they recognize authors from less prominent institutions. Tier 1 and Tier 2 institutions show minimal differences when authors are known.

Country-tier patterns are less consistent due to smaller sample sizes in Tiers 2 and 3. Country Tier 2 shows notably fewer "should do instead" demands when authors are known (0% vs. 19%, $d = -0.49$).

D.3 Effect of Preprint Timing on Lazy Patterns

Table 6 presents the effect of preprint timing on lazy patterns. We compare reviews for papers with preprints posted before versus after the submission deadline.

Papers without preprints present during review receive more lazy feedback across multiple dimensions. The most striking finding concerns "compare to model" requests for institution Tier 2: these jump from 4% when preprints are present to 24% when absent ($d = 0.71$, $p < 0.01$). This pattern suggests reviewers demand expensive model comparisons more often when they cannot verify institutional prestige through preprints.

For lower-tier institutions and countries, preprint absence correlates with increased "should do instead" demands. Country Tier 2 shows a jump from 11% to 38% ($d = 0.66$, $p < 0.05$), and Country Tier 3 shows a similar pattern (8% to 38%, $d = 0.83$, $p < 0.05$). Institution Tier 3 papers without preprints present receive more unclear comments (54% vs. 37%, $d = 0.35$).

These patterns indicate that preprint visibility reduces certain types of lazy reviewing, but the effect varies by institutional context. Elite institutions benefit from reduced unclear comments, while mid-

Outcome	Interaction	β	SE	p
Institution Tier × Preprint Presence				
Overall Score	Tier 1 × Present	0.15	0.06	.020*
Helpfulness	Tier 1 × Present	−0.19	0.08	.024*
Review Length	Tier 1 × Present	−31.2	18.4	.091
Institution Tier × Author Recognition				
Overall Score	Tier 1 × Known	0.19	0.08	.018*
Actionability	Tier 2 × Known	−0.49	0.21	.021*
Unclear %	Tier 3 × Known	−0.21	0.09	.023*
Country Tier × Preprint Presence				
Overall Score	Tier 1 × Present	0.17	0.05	.001**
Compare Model %	Tier 1 × Present	−0.06	0.03	.042*

Table 7: Interaction model coefficients. $^{*}p < 0.05$; $^{**}p < 0.01$. Positive β for scores = larger benefit at that tier. Negative β for quality metrics = treatment reduces quality at that tier.

tier institutions face sharply increased demands for expensive comparisons when preprints are absent.

E Quantitative Analysis: Interaction Models

This appendix provides statistical details for the regression models referenced in Section 5. The main text reports within-tier comparisons (e.g., preprint present vs. absent for Tier 1 institutions). Here we test whether the *magnitude* of these effects differs across tiers.

E.1 Model Specification

We fit OLS regression models with interaction terms to test whether preprint visibility and author recognition affect different tiers differently. The general form is:

$$Y = \beta_0 + \beta_1 \cdot \text{Tier} + \beta_2 \cdot \text{Treatment} + \beta_3 \cdot (\text{Tier} \times \text{Treatment}) + \epsilon \quad (1)$$

where Y is the outcome (score, quality metric, or lazy pattern rate), Tier is a categorical variable (Tier 1 as reference), and Treatment indicates preprint presence or author recognition. A significant β_3 indicates that the treatment effect varies by tier. All p-values are FDR-corrected within outcome families.

E.2 Results

Table 7 presents the key interaction effects. Positive coefficients for score outcomes indicate that the treatment benefit is larger at the specified tier. Negative coefficients for quality metrics indicate that treatment reduces quality more at that tier.

E.3 Interpretation

Three findings emerge from the interaction models.

Score benefits concentrate at elite tiers. The Tier 1 × Present interaction for overall score is positive and significant for both institution ($\beta = 0.15$, $p = .020$) and country ($\beta = 0.17$, $p = .001$) classifications. The Tier 1 × Known interaction shows the same pattern ($\beta = 0.19$, $p = .018$). Preprint visibility and author recognition do not uniformly raise scores.

Quality tradeoffs differ by tier. When reviewers can identify elite authors (Tier 1), scores rise but review helpfulness drops ($\beta = -0.19$, $p = .024$). For Tier 2, author recognition correlates with lower actionability ($\beta = -0.49$, $p = .021$). For Tier 3, author recognition correlates with fewer unclear comments ($\beta = -0.21$, $p = .023$).

Resource-intensive demands shift with preprint visibility. The Tier 1 × Present interaction for "compare to model" requests is negative ($\beta = -0.06$, $p = .042$). When preprints are present, reviewers request fewer model comparisons from Tier 1 authors. When preprints are absent, such requests increase.