Understanding Disagreement in Peer Review

Büşra Asan* Matr.Nr. 7049398
Halil Faruk Karagöz* Matr.Nr. 7058774
Koray Ulusan* Matr.Nr. 5788277
Yusuf Nar* Matr.Nr. 7000476

Abstract

The peer-review process in major machine learning conferences has long been scrutinized for its subjectiveness, inconsistency and delayed review submissions. In this study, we examine review score discrepancies — instances where a single paper receives highly conflicting evaluations. Concentrating on ICLR 2023 submissions, we introduce a discrepancy metric to quantify disagreements and investigate their prevalence. Furthermore, we explore the interplay between conflicting reviewer scores and confidence levels, and examine how discrepancy relates to subject popularity and future impact. Our analysis aims to uncover underlying structural challenges within the review process and offers insights into potential biases that can hinder the evaluation of innovative work.

1 Introduction

The debate around peer review at major machine learning conferences is not new by any means. Despite its important role in determining the acceptance of research papers to conferences and journals, academics have been concerned with its arbitrariness and consistency.

Cortés *et al.* revealed significant inconsistencies by randomly assigning papers to multiple committees and comparing the committee decisions [1]. They also argue that while peer review is effective at filtering out poor papers [2], it often fails to recognize promising ones—a shortcoming further underscored by Brezis *et al.*, who demonstrates that reviewers might penalize innovative work with high potential impact [3]. While these studies highlighted the system's susceptibility to erratic outcomes through randomized controlled experiments, our work takes a complementary approach by analyzing the issue through the lens of "discrepancy" in review scores, instances where a single paper receives divergent evaluations (*e.g.*, 3 and 8 on a 10-point scale) from different reviewers.

We argue that disagreements to this extent are imprints of conceivable conflicts in expert opinion, which appear abnormal and may point to structural problems within the evaluation process; therefore, they should be meticulously investigated. We first investigate the prevalence of high-discrepancy papers among ICLR 2023 submissions. Subsequently, we explore the interplay between conflicting review scores and the corresponding confidence levels of reviewers. Afterwards, we analyze how discrepancy correlates with subject popularity and future impact. Finally, we assess whether late submissions are linked to a higher likelihood of reviewer score discrepancies.

2 Dataset

We focus on the second most impactful machine learning conference ICLR, according to [4] with an h5-index of 304. Unlike the top conference NeurIPS, rejected papers are not withdrawn in the API; thus we are able to keep track of the reviewing process.

Our dataset is curated using OpenReview API [5], extracting features including but not limited to *review text, paper title, keyword, submission date, confidence* and *reviewer score*. Furthermore, we integrate citation counts per paper using the titles from the dataset we constructed by scraping with Scholarly API [6]. Our dataset comprises 3796 paper submissions with a total of 14383 reviews.

*Equal Contribution.

Pre-processing includes typecasting, extracting numerical scores from textual values, and investigating aggregated statistics such as mean and standard deviation. For data cleaning, we remove `NaN` values, exclude reviews that were submitted after 40 days of the submission period, and solve conflicting entries (*e.g.* ethics flags including both "no" and "yes" using the explanations).

We further assign each paper a discrepancy score using the Equation 1. Additionally, we prepare a research topic dictionary by collecting author-provided keywords, removing generic terms (*e.g.* "image", "neural network"), and grouping semantically similar keywords (*e.g.* "NLP" and "language") into a single broader research category. Each paper is then assigned a topic based on keyword matching with this dictionary. We calculate topic popularity from relative frequencies and use these scores to assign a popularity score to each paper in Section 3.3 Our dataset and codes are publicly accessible on https://github.com/halilfarukkaragoz/data-literacy-group16.

3 Methods & Results

We define the *discrepancy* of a paper's reviews as the difference between the highest and lowest review scores it receives, i.e. for a given paper P_i, the discrepancy is given by:

$$\text{Discrepancy}(P_i) = \max R_i - \min R_i \tag{1}$$

where R_i is the set of P_i's review scores. We classify papers with a discrepancy of at least 4 as *high discrepancy* and the remaining papers as *low discrepancy*. Our suggested metric provides a clear and interpretable measure of reviewer disagreement as the discrepancy of at least 4 often ensures that at least one reviewer supports acceptance while another favors rejection with their scores separated by a significant margin. Thus, our metric not only identifies cases of disagreement but also excludes borderline conflicts where scores are close, which makes it more robust for detecting truly contentious papers. We find 631 papers in ICLR 2023 fall into high discrepancy category, constituting 16.62% of the total. We believe this indicates a notable level of expert disagreement. Furthermore, Figure 1 illustrates that high discrepancy papers have a higher rejection rate. As the samples in each group may not follow identical Bernoulli distributions, we use permutation test ($n = 10000$) to see the significance of this difference. Corresponding p-value is less than 10^{-6}, which suggests that difference in the acceptance probabilities between the two groups is significant.

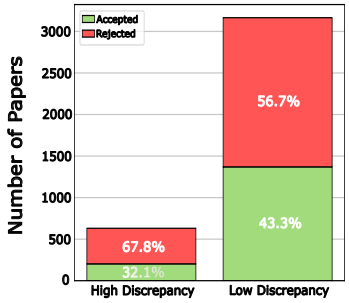

Figure 1: Histogram showing the accepted/rejected rates for high and low discrepancy papers.

3.1 Confidence Scores

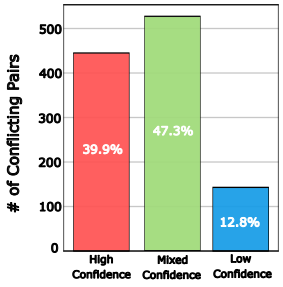

Figure 2: Histogram comparing confidence groups among conflicting reviewer pairs.

Reviewers provide a confidence score ranging from 1 to 5 indicating their confidence in their reviews. When we classify confidence levels of 4 and 5 as *high confidence* and levels of 1, 2 and 3 as *low-confidence* a natural question arises: how many papers have at least one pair of reviewers who are both highly confident yet whose scores are highly discrepant? In fact, 316 out of the 631 high discrepancy papers fall under this category, constituting 50.08% of high discrepancy papers and 8% of the total papers. Furthermore, if two reviewers are assigned to the same paper and have discrepant scores, we classify them as *conflicting reviewers*.

We categorize conflicting reviewers into three mutually exclusive groups: (1) both reviewers are highly confident, (2) one reviewer is highly confident while the other has low confidence, and (3) both reviewers have low confidence.

Figure 2 illustrates the distribution of conflicting reviewers into those categories, where High Confidence, Mixed Confidence and Low confidence indicates the groups 1, 2 and 3 respectively. Our dataset has 1,115 conflicting reviewers in total, 445 of which (40%) involve both reviewers being highly confident. Thus, it is not uncommon for two experts to strongly disagree while both feeling

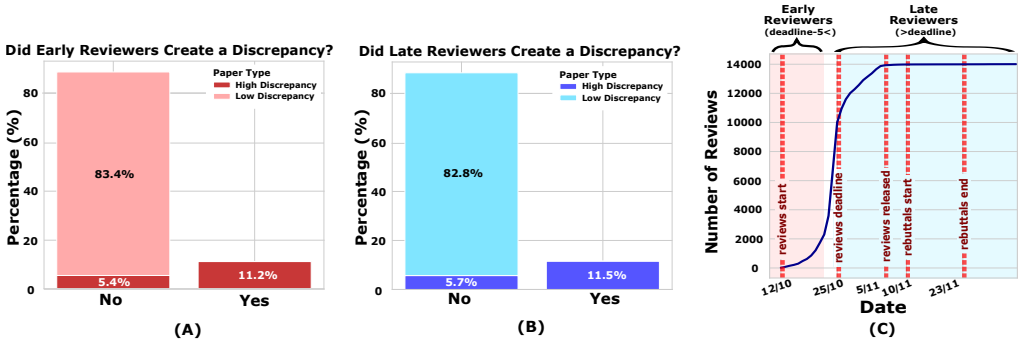

Figure 3: Figures (A) and (B) show the percentage of early and late reviewers who submit conflicting scores relative to previously submitted reviews of the same paper. The amount of discrepancies created in both time periods are almost same in percentage. (C) shows cumulative submissions over time, with key milestone dates marked. Early and late reviewers' submission times are indicated.

sure of their positions according to our data. This result further emphasizes that disagreements are not only common but also strongly defended by a noticeable number of confident reviewers.

3.2 Submission Time

Using the review's creation date, we calculate the elapsed time relative to the submission deadline. We hypothesize whether a short period between the submission and the review deadline increases the likelihood of discrepancies between reviewers' scores.

We observe that 11.2% of the early reviewers in Figure 3 (A) gave the score that created discrepancy among other scores. We define early reviewers as those who submitted their evaluations at least five days before the deadline, a threshold chosen because 59% of the reviews were submitted in the last five days. Number of submitted reviews throughout the reviewing process can be seen in Figure 3 (C). According to our hypothesis, we expect late reviewers to contribute more frequently to discrepancies. However, we see that late reviewers also have a similar chance of creating a discrepancy as shown in Figure 3 (B).

We also conduct a Pearson correlation test between the submission date distributions of high discrepancy and low discrepancy papers by fitting a Gaussian to our histogram plot of number of submission per day. Corresponding p-value is 0.999, which indicates that there is no statistically significant relationship between papers having discrepant scores and late submission date.

3.3 Topic Popularity Analysis

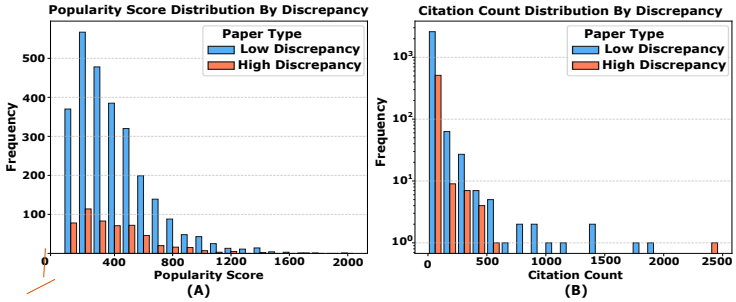

Figure 4: Distribution plots comparing papers with high and low review discrepancies. (A): Popularity Score distribution showing frequency counts. (B): Citation Count distribution on a logarithmic scale.

The motivation for this analysis stems from the concern that "trendy" topics might introduce reviewer bias, leading to greater scoring inconsistencies. To explore this possibility, we define the **Popularity Score**, which sums each paper's distinctive topic frequencies in the corpus. Additionally, we use **citation counts** as a post-publication metric to capture scholarly impact.

Figure 4 illustrates the distributions of Popularity Score and Citation Count for papers with high and low peer-review discrepancies. To assess statistical differences, we apply the non-parametric Mann-Whitney U-Test [7], as our data are non-Gaussian. The resulting p-values (0.9211 for Popularity Score and 0.3493 for Citation Count) indicate no significant difference between the two groups. Thus, based on the p-values obtained (0.9211 for Popularity Score and 0.3493 for Citation Count), we

find no statistically significant evidence that popularity—whether measured by topic prevalence or citation impact—affects scoring inconsistencies.

We categorize papers based on their citation count, using it as a proxy for impact. We define three balanced categories: High Citation (more than 50), Moderate Citation (between 20 and 50), and Low Citation (less than 20). To examine how different features vary across these categories, we constructed a radar plot as shown in Figure 5. We observe that discrepancy is noticeably smaller in certain topics, such as Detection, Federated Learning and Diffusion. On the other hand it is slightly above average in Generative AI. Further study is required to see if this is merely a noise in the data or a meaningful pattern. Additionally, interesting interactions emerge between some of our features. For instance, theory papers tend to receive less confident reviews yet have a higher acceptance rate, whereas contrastive learning papers receive more confident reviews than average. Interestingly, Graph Neural Networks and Reinforcement Learning were popular topics in ICLR2023 but their popularity not translate into higher citation counts. Significance of these observations remains uncertain and necessitates a thorough qualitative analysis, which is beyond the scope of this paper.

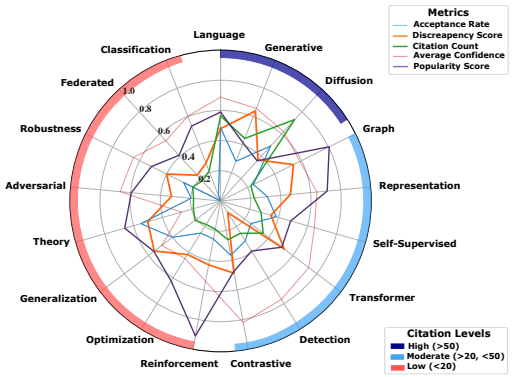

Figure 5: Radar plot showing acceptance rate, discrepancy score, citation count, confidence, and popularity across topics, categorized by citation levels. Metrics are normalized for comparison.

4 Discussion

Notable number of papers exhibit conflicting reviews with confident yet discrepant scores, suggesting that strong disagreements among experts are common. To better understand these inconsistencies, we analyzed their relationship with submission time, topic popularity, and citation count.

Our results were not enough to show significant links between review discrepancy and aforementioned factors. Early and late submitted reviews contribute uniformly to the variation in the discrepancy. Likewise, high discrepancy papers do not demonstrate a substantially different citation count and popularity score distributions when compared with the low discrepancy group.

It is important to note that our analysis is based on a relatively small subset of ML conference publications, and our findings may be strongly influenced by the case-specific characteristics of ICLR 2023. For instance, we found that significant portion of reviews were submitted after the deadline, which implies that the deadline was not strictly enforced. This may have introduced bias in reviewer behavior regarding deadline pressure, potentially concealing the expected effect. Additionally, we utilized citation counts as a proxy for the impact measure of papers. However, ICLR 2023 papers were published in May 2023 and their citation counts may not yet accurately reflect their true influence. Furthermore, discrepancy metric might oversimplify complex disagreements by disregarding intermediate ratings and review text nuances.

Future work could extend this analysis to time series comparisons to determine whether the discrepancy rate has increased in recent years alongside rising submission numbers. Additionally, similar investigations could be conducted at conferences such as NeurIPS and CVPR, and further variables—such as review length and area chair–reviewer conflicts could be explored.

Statement of Contribution

Following CRediT System [8] our work distributuion is: Büşra Asan analyzed the effect of time to review deadline, worked on conceptualization and project management. Halil Faruk Karagöz worked on data curation and investigated popularity and citation effect. Koray Ulusan worked on data curation, resources, correlation analysis and data pipeline. Yusuf Nar conducted pre-analysis on the dataset, highlighting a notable number of papers with confident yet discrepant reviews. All members of the group contributed to writing, proofreading, formal analysis, visualization, and software.

References

[1] Corinna Cortes and Neil D. Lawrence. The nips experiment. https://github.com/lawrennd/neurips2014/blob/master/notebooks/The%20NIPS%20Experiment.ipynb, 2014. GitHub repository.

[2] Corinna Cortes and Neil D. Lawrence. Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment. *arXiv*, September 2021.

[3] Elise S. Brezis and Aliaksandr Birukou. Arbitrariness in the peer review process. *Scientometrics*, 123(1):393–411, April 2020.

[4] Google. Google scholar, 2023. Accessed: 2025-02-02.

[5] OpenAPI definition | OpenReview, January 2025. [Online; accessed 21. Jan. 2025].

[6] Zubair Baig et al. Scholarly: A python package for retrieving google scholar data, 2023. Accessed: 2025-02-02.

[7] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60, 1947.

[8] Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2):151–155, April 2015.