

Manuscript Structure, Communicative Features, and Peer-Review Outcomes: A Large-Scale Observational Study of OpenReview Data

RENLONG JIE

jierenlong@nwpu.edu.cn

Northwestern Polytechnical University

Chen Chu

Yunnan University of Finance And Economics

Ruibin Geng

Northwestern Polytechnical University

Danyang Jia

Northwestern Polytechnical University

Tingyu Lu

Northwestern Polytechnical University

Guiwen Guan

Northwestern Polytechnical University

Zhen Wang

Northwestern Polytechnical University

Research Article

Keywords: peer review, research evaluation, scientometrics, manuscript structure, OpenReview

Posted Date: June 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9861817/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Manuscript Structure, Communicative Features, and Peer-Review Outcomes: A Large-Scale Observational Study of OpenReview Data

Renlong Jie^{1*}, Chen Chu², Ruibin Geng¹, Danyang Jia³,
Tingyu Lu¹, Guiwen Guan³, Zhen Wang^{3*}

^{1*}School of Management, Northwestern Polytechnical University, Xi'an,
710072, Shaanxi, China.

²School of Artificial Intelligence, Optics and Electronics (iOPEN),
Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, China.

³School of Statistics and Mathematics, Yunnan University of Finance
and Economics, Kunming, 650221, Yunnan, China.

*Corresponding author(s). E-mail(s): jierenlong@nwpu.edu.cn;
w-zhen@nwpu.edu.cn;

Contributing authors: chuchenynufe@hotmail.com;
rbgeng@nwpu.edu.cn; jiadanyang93@outlook.com;
lutinyu@nwpu.edu.cn; guanguiwen@bjmu.edu.cn;

Abstract

This paper provides a large-scale observational mapping of how manuscript-level structural and communicative features are associated with peer-review ratings and decisions in OpenReview-based AI conferences. We examine 29,279 submissions and accepted papers from major artificial intelligence conferences (ICLR, NeurIPS, and ICML), focusing on automatically extracted document-level cues such as title length, sentence length, formulas, figures, tables and references. Because decision availability differs across venues and years, acceptance models are estimated primarily on the full-decision ICLR sample, while cross-venue accepted-paper data are used only for descriptive comparisons. Using interpretable regression models and covariate-balancing sensitivity analyses, we document robust adjusted associations between several communicative-structural features and review outcomes: larger author teams, shorter titles, shorter average sentence length and more main-body mathematical exposition are consistently associated with more favorable ratings and acceptance. These patterns remain

after adjusting for a SciBERT-based semantic-content proxy derived from titles and abstracts. The results are observational regularities rather than causal or prescriptive claims, and should not be read as evidence of reviewer bias. The study contributes to scientometric research by clarifying both the value and the boundaries of using OpenReview data for document-level peer-review analysis.

Keywords: peer review, research evaluation, scientometrics, manuscript structure, OpenReview

1 Introduction

Peer review is one of the central institutions through which scientific communities allocate recognition, publication opportunities and research visibility. In high-volume evaluation settings, reviewers are required to assess complex manuscripts under substantial time and attention constraints. While prior scientometric research has examined reviewer behavior, review scores, acceptance patterns and institutional or demographic disparities, less is known about how the observable communicative and structural features of manuscripts are associated with evaluation outcomes. Major artificial intelligence conferences provide a useful empirical setting for studying this question. Their review processes are competitive, large-scale and increasingly transparent through platforms such as OpenReview, making it possible to examine manuscript-level features and review outcomes across thousands of submissions (Yao et al. 2023; Stelmakh et al. 2021; Shah 2022).

Ideally, acceptance of a research paper should be based solely on substantive merit. However, information processing theories suggest that reviewers operating under high cognitive load and limited attention may rely on heuristic cues (Ayal et al. 2011; Chaiken and Ledgerwood 2012). In this context, metadata and presentation-related features may shape how reviewers perceive, process, and evaluate scientific contributions, especially under time constraints. This does not necessarily imply that reviewers are irrational or that presentation features are inherently unfair; rather, it suggests that scholarly evaluation is partly mediated by communication quality. Furthermore, competing objectives among authors, conferences, and reviewers introduce inherent noise into the evaluation process (Zhang et al. 2022; Aczel et al. 2025). Therefore, examining the association between manuscript characteristics (including both author metadata and communicative-structural features) and peer review outcomes is essential for understanding the regularities and potential fairness concerns in scientific evaluation.

The emergence of the OpenReview platform in the AI community provides a unique opportunity to investigate these dynamics. By adopting open peer review models, top-tier conferences such as NeurIPS and ICLR have significantly improved process transparency (Wang et al. 2023; Wolfram et al. 2020). This platform allows for the tracking of reviews, rebuttals, and decision-making processes. Recent studies have used these data to analyze the double-blind process (Wang et al. 2023), explore the feasibility of LLM-augmented reviews (Drori and Te’eni 2024), and outline tasks for

automating parts of the review pipeline (Staudinger et al. 2024). Currently, machine learning approaches have been applied to predict acceptance rates and assess paper quality (Skorikov and Momen 2020; Joshi et al. 2021; Bao et al. 2021; Jin et al. 2024). Kankanhalli (2024) also highlight the potential of generative AI to automate routine tasks, while cautioning against ethical risks. Despite these advances, most of the existing studies focus on content analysis or citation networks (Yan and Ding 2011; Huang et al. 2022). There remains a lack of interpretable statistical analysis on how presentation features and metadata are associated with reviewer ratings and acceptance decisions.

We conceptualize manuscripts not only as containers of scientific content but also as communicative objects. Their titles, sentence structure, formulas, figures, tables and references organize claims, signal evidential support and shape the accessibility of technical arguments. These features do not determine scientific merit, but they may be associated with how reviewers encounter and evaluate a submission.

We treat large AI conferences not as the object of prediction, but as an empirical setting to study a broader scientometric question: how communicative-structural features of manuscripts are associated with evaluative outcomes under high-volume peer review. Our focus on NeurIPS, ICLR, and ICML is motivated by their high submission volumes, transparent OpenReview data, and central role in the AI research community. Unlike previous “black-box” prediction models, we employ interpretable statistical methods to quantify the association between review outcomes and specific manuscript characteristics.

Specifically, our **research objective** is to answer the following questions:

- **RQ1:** How are author-team characteristics associated with peer review scores and acceptance outcomes?
- **RQ2:** How are communicative-structural features of manuscripts, including title length, sentence length, formulas, figures, tables and references, associated with peer review outcomes?
- **RQ3:** To what extent do these associations persist after adjusting for observable venue, year, topic and semantic-content proxies?

We conceptualize manuscripts as communicative objects in peer review. Scientific contributions are not evaluated through substantive content alone, but through structured representations of that content: titles, sentences, figures, tables, formulas, references, and the placement of evidence within the main text. In high-volume peer review settings, these communicative-structural features may be associated with how reviewers locate, process, and assess claims under limited attention. By answering these questions, this research provides large-scale observational evidence that manuscript metadata and presentation-related features are systematically associated with peer review outcomes, and that some associations remain stable after adjusting for venue/year, metadata, and an abstract-level semantic-content proxy.

This study makes three contributions. First, we construct a large-scale paper-level dataset linking OpenReview outcomes with manuscript-level structural features,

with explicit distinctions among full-decision samples, rating samples, and accepted-only cross-venue samples. Second, we document robust adjusted associations between communicative-structural features and reviewer ratings or decisions in samples where such outcomes are observable, using covariate-balancing sensitivity checks to assess persistence after adjusting for observable covariates. Third, we clarify the scope and limitations of using OpenReview data for scientometric analysis of peer review, including the incomplete decision availability across venues and the proxy nature of Grobid-extracted and SciBERT-derived variables.

Conceptually, we distinguish among three channels through which observable submission characteristics may be associated with peer review outcomes. The first is a *capacity channel*, whereby larger author teams may have more complementary expertise, greater resources, and more opportunities for internal quality control. The second is a *communication channel*, whereby concise titles, shorter sentences, and well-organized main-text exposition reduce reviewers’ processing costs and improve perceived clarity. The third is a *visible-evidence channel*, whereby figures, tables, and formulas placed in the main body make empirical or theoretical support more salient during review. These channels do not necessarily represent unfairness; they may partly reflect legitimate components of scholarly communication. However, they also raise fairness concerns if presentation polish receives disproportionate weight relative to substantive novelty or correctness.

2 Related Work

2.1 Peer review as a research evaluation mechanism

Peer review serves as the primary gatekeeping mechanism for scientific quality assurance, yet its reliability, validity, and susceptibility to systemic biases remain active subjects of inquiry (Shah 2022; Aczel et al. 2025; Bornmann 2011). The rapid growth of AI conference submissions has intensified concerns about reviewer overload and the consistency of evaluative judgments (Stelmakh et al. 2021). Studies have documented persistent issues including low inter-reviewer agreement (Shah 2022; Bornmann et al. 2010), demographic disparities in acceptance rates (Tran et al. 2020; Ganapathy et al. 2022), and the potential for institutional and geographic signals to be associated with evaluation outcomes (Stelmakh et al. 2019; Manzoor and Shah 2021). Double-blind review policies, adopted by venues such as NeurIPS and ICLR, aim to reduce identity-based biases, though their effectiveness varies by context (Wang et al. 2023). Huisman and Smits (2017) examined the duration and quality of the peer review process from the author’s perspective, finding significant variation across journals and disciplines. Horbach and Halfman (2019) compared the ability of different peer review procedures to flag problematic publications, concluding that no single procedure is uniformly superior. Ross-Hellauer (2017) provided a systematic review of open peer review models, highlighting the diversity of practices and their implications for transparency. Lee et al. (2023) surveyed the state of peer review research, identifying key gaps in our understanding of review quality and bias. These studies underscore that peer review is not a purely meritocratic filter but a socio-technical process mediated by human cognition, institutional structures, and communication norms.

2.2 Manuscript communication and structural cues

A growing body of work examines how the presentation and communication of research shape its evaluation. Information processing theories suggest that reviewers operating under cognitive load may rely on heuristic cues (Petty and Cacioppo 1986; Chaiken and Ledgerwood 2012). In this context, manuscript features such as title conciseness, sentence length, and the placement of formulas, figures, and tables may function as communicative signals that shape reviewer perceptions. The Elaboration Likelihood Model (Petty and Cacioppo 1986) posits that when evaluators lack the motivation or capacity to process all arguments centrally, peripheral cues such as presentation fluency and visible evidence may influence judgments. We use this framework as an interpretive lens rather than as a directly tested psychological mechanism. Similarly, Signaling Theory (Spence 1978; Connelly et al. 2011) suggests that under information asymmetry, observable features (e.g., technical formalism prominently displayed in the main text) can serve as signals of unobserved quality, provided they are costly to produce and visible to evaluators.

Empirical studies in scientometrics have linked manuscript characteristics to citation and reception patterns. Skorikov and Momen (2020) found that surface features such as title length and figure count carry predictive signal for acceptance decisions. Huang et al. (2022) showed that scholars’ career stage shapes their research topic choices, which in turn affects how their work is received. Hasan and Breunig (2021) examined article length and citation outcomes across disciplines, finding that longer articles tend to receive more citations, though the relationship varies by field. Jamali and Nikzad (2011) investigated the relationship between title type and downloads/citations, revealing that certain title structures are associated with higher download and citation rates. Guo et al. (2018) explored whether the relationship between title length and citations follows a “succinct effect” (shorter titles attract more citations) or an “informative effect” (longer titles provide more information and thus more citations), finding evidence for both mechanisms depending on disciplinary context. Nair and Gibbert (2016) reviewed title attributes across management science and found that title structure significantly predicts citation counts. Hudson (2016) analyzed titles submitted to the UK REF and documented disciplinary differences in title characteristics and their association with impact. Lewison and Hartley (2005) examined the evolution of title length and colon usage across scientific fields over two decades, revealing field-specific conventions and temporal trends.

These findings suggest that communicative-structural features are not merely cosmetic but may mediate the relationship between substantive content and evaluative outcomes. Our study builds on this foundation by systematically quantifying how presentation-related features are associated with peer review outcomes after adjusting for abstract-level content signals, offering evidence on the role of communication quality in high-stakes expert evaluation. Unlike prior work that focused primarily on post-publication citation outcomes, we examine the association between manuscript features and peer-review ratings and decisions, providing insights into the evaluation stage of the scientific communication process.

2.3 Computational analysis of peer review outcomes

The availability of large-scale peer review datasets, such as PeerRead (Kang et al. 2018) and OpenReview archives, has enabled computational studies of the review process. Early work focused on predictive modeling: Skorikov and Momen (2020) used Random Forests with surface features to predict acceptance, while subsequent studies incorporated semantic and sentiment signals (Wang and Wan 2018; Ghosal et al. 2019). More recently, Transformer-based models have been applied to rating prediction (Rimkus et al. 2023; Kang et al. 2022; Xue et al. 2023). These studies demonstrate that review outcomes contain predictable structure, but they typically prioritize accuracy over interpretability and do not examine which associations persist after controlling for abstract-level semantic and framing signals.

A complementary stream of research has developed methods for detecting and quantifying bias. Stelmakh et al. (2019) proposed statistical tests for bias in review ratings, and Manzoor and Shah (2021) measured textual bias by comparing review language before and after double-blind policy implementation. Tran et al. (2020) and Ganapathy et al. (2022) documented demographic and institutional disparities in acceptance rates. At the intersection of scientometrics and review analytics, Ausloos et al. (2019) examined correlations between submission and acceptance patterns across peer review journals, highlighting systematic structural factors that shape publication outcomes. Squazzoni and Gandelli (2012) used agent-based modeling to study how competitive selection and reviewer reciprocity affect peer review quality and bias. Teixeira da Silva et al. (2021) examined peer review during the COVID-19 pandemic, noting how time pressure and manuscript volume affected reviewer behavior and the risk of publishing flawed research. However, most of these studies stop at documenting raw disparities without assessing whether they persist after adjusting for observable covariates and content proxies. Our research addresses this gap by combining interpretable regression models with propensity-score-based robustness analyses. We move beyond simple prediction and raw disparity documentation to examine whether presentation-related associations persist after balancing observable covariates and accounting for a text-based content proxy.

3 Methodology

This study documents the associations between manuscript-level structural features and peer-review outcomes using observational data from OpenReview. Because decision availability differs across venues and years, acceptance models are estimated primarily on the full-decision ICLR sample, while cross-venue accepted-paper data are used only for descriptive comparisons. We apply regression and covariate-balancing sensitivity analyses, with all estimates interpreted as adjusted associations rather than causal effects.

3.1 Scope of inference and sample construction

The availability of review outcomes varies across venues and years. To avoid conflating different data-generating mechanisms, we distinguish three analytic samples with distinct inferential scopes, summarized in Table 1.

Table 1 Scope of inference and sample construction.

Venues/years	Includes rejected?	Used for	Inference allowed
ICLR 2024/2025	Yes	Acceptance models	Associations with decision outcomes within ICLR
ICLR/NeurIPS where ratings available	Partially	Rating regressions	Associations with review scores among observable reviewed papers
ICML/NeurIPS/ICLR accepted papers	No	Descriptive comparison only	Accepted-paper patterns only, not acceptance mechanisms

The *full-decision sample* (ICLR 2024/2025) contains submissions for which both accepted and rejected outcomes are observable. This is the only sample for which acceptance-related associations can be estimated, because the complete decision process is observed. The *rating sample* includes submissions from ICLR and NeurIPS with publicly available reviewer ratings and is used for analyses of review scores; rejected submissions in this sample are partially available depending on venue and year. The *accepted-paper cross-venue sample* includes accepted papers from all three venues and is used only to characterize presentation conventions among successful papers. We do not treat accepted-only observations as representative of the full submission pool, and we avoid using them to estimate acceptance mechanisms.

3.2 Empirical setting and review process

The manuscripts examined in this study were submitted to three leading machine learning conferences (ICLR, NeurIPS, and ICML) between 2023 and 2025. The review processes of these conferences are similarly structured, primarily administered through the OpenReview platform to promote transparency. Each adheres to a double-blind review policy and follows a rigorous multi-stage procedure. The typical workflow includes: (1) Submission and Initial Check, ensuring compliance with formatting and anonymity guidelines; (2) Reviewer Assignment, conducted via a bidding system combined with algorithmic matching; (3) Peer Review and Discussion, during which multiple reviewers evaluate submissions in several structured dimensions, such as soundness, contribution, presentation, confidence, and overall rating. This is followed

by an author rebuttal period, allowing authors to respond to reviewer comments. Subsequently, interactive discussions among reviewers and area chairs take place to clarify perspectives and reach consensus; (4) Meta-Review and Decision, where a senior area chair synthesizes all inputs to make a final recommendation.

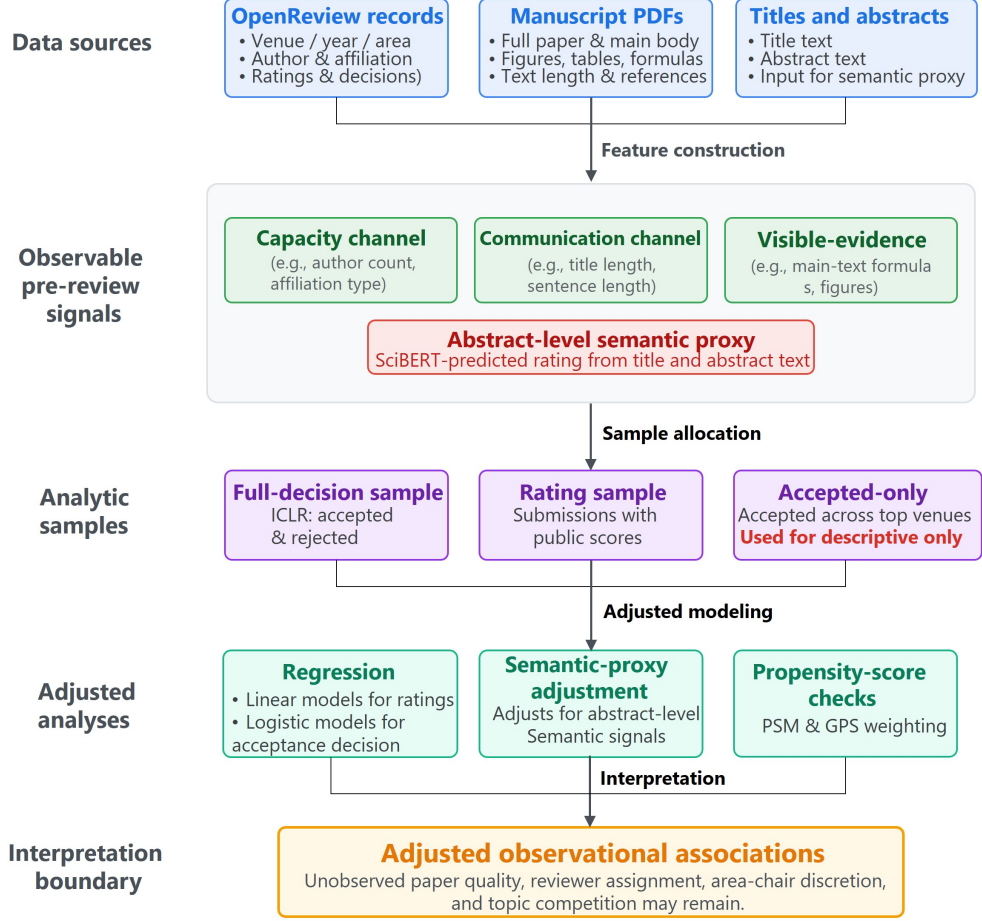


Fig. 1 Overview of the observational research design and sample structure. Full-decision sample (ICLR) is used for acceptance models; rating sample (ICLR/NeurIPS) for score models; accepted-only sample for descriptive comparisons only.

3.3 Adjusted regression models

To examine associations between predictors and outcomes, we employ multiple linear regression for ratings and logistic regression for acceptance decisions ([Tranmer](#)

and Elliot 2008; Kleinbaum et al. 2002; Aldrich and Nelson 1984). Figure 1 summarizes the observational research design. OpenReview records, manuscript PDFs, and titles/abstracts are used to construct observable metadata, communicative-structural features, and an abstract-level semantic-content proxy. The analysis then distinguishes full-decision samples, rating samples, and accepted-only cross-venue data.

3.4 Covariate-balancing robustness adjustment

We use propensity-score matching and generalized propensity-score weighting as covariate-balancing sensitivity checks to assess whether main associations persist after balancing observed covariates. Because unobserved confounders cannot be fully controlled, estimates are interpreted as covariate-balanced associations rather than causal effects.

For binary focal variables (e.g., US affiliation), propensity-score matching pairs focal and comparison papers with similar probabilities of focal variable status given observed covariates, creating more comparable matched subsets. For continuous focal variables (e.g., number of formulas), generalized propensity-score weighting balances observed covariates across focal variable levels via inverse-probability weights, allowing us to estimate associations adjusted for observable confounders.

To partially account for abstract-level semantic and framing signals, we construct a text proxy by fine-tuning a SciBERT (Beltagy et al. 2019) model to predict the average reviewer rating from the paper’s title and abstract. This proxy is not intended to capture intrinsic scientific quality—novelty, correctness, rigor, and theoretical contribution cannot be inferred from titles and abstracts alone—but rather to adjust for observable topical and framing variation. Because titles and abstracts also contain communicative choices, this adjustment may absorb some presentation-related variation; estimates should therefore be interpreted as conservative adjusted associations.

3.5 Independent variables

To construct interpretable models of peer review outcomes, we define a set of independent variables organized into three groups. The first group comprises *metadata variables* (author- and affiliation-level information), which capture observable social and organizational characteristics. The second group comprises *presentation-related* or *communicative-structural variables*, including title length, sentence length, text length, and the placement of formulas, figures, and tables. We avoid treating all extracted manuscript features as purely “non-content” variables. Features such as formulas, figures, tables, and references are content-bearing, but they also reflect how scientific arguments are organized and communicated. We therefore refer to them as presentation-related or communicative-structural features. The third group is an *abstract-level text proxy* (SciBERT-predicted rating). These variables are chosen for their analyzability and potential to relate to reviewer decisions.

3.5.1 Metadata

Meta-data, collected via the OpenReview API, describe the submission context and include the venue and primary research area, which may reflect field-specific conventions (e.g., theoretical papers are more common at ICML, while ICLR papers often contain extensive appendices). In addition, we capture author-level information such as the number of authors, the first author’s inferred gender, academic position (e.g., student, postdoc, professor), and the type (university vs. industry) and geographic region (e.g., US, China, EU) of their affiliation. These variables are included as exploratory contextual controls rather than as primary focal variables. Gender is inferred from first names and is subject to measurement limitations discussed in Section 4.1.

3.5.2 Presentation features

Presentation-related features are extracted from the PDF files of manuscripts using Grobid and quantify formatting choices that may be associated with readability and perceived rigor. We include metrics of textual extent, such as the length of the abstract, title, main body text, and full text, as well as the average sentence length. Mathematical content is captured by the automatically detected number and total length of formula expressions, with a distinction between those appearing in the main body versus the full paper. Visual and tabular elements are represented by the automatically detected counts of figures and tables, and by the total number of table elements (cells), again separated by location (main body vs. full paper). Finally, we record the number of literature references and total references to gauge the depth of engagement with prior work. All variables listed in Table 2 are used in subsequent regression and robustness analyses; where appropriate, skewed variables are log-transformed to improve model fit.

3.6 Dependent variables

We define two primary dependent variables: the final acceptance decision and reviewer ratings. Acceptance includes all papers accepted as poster, spotlight, or oral presentations. Rejection encompasses all submissions that were not accepted, including reject, withdraw, and desk reject decisions. For submissions to ICLR 2024/2025 and NeurIPS 2023/2024, reviewers provide ratings across multiple dimensions: soundness, presentation, and contribution (each on a 1–5 scale), along with an overall rating (on a 1–10 scale). In our regression analyses, we use the average overall rating as the continuous dependent variable.

4 Data description and exploratory analysis

4.1 Data description

We collect data from three top machine learning conferences (ICLR 2024/2025, NeurIPS 2023/2024, and ICML 2023/2024) comprising $N = 29,279$ manuscripts (16,634 accepted, 12,645 rejected, see Table 3). Complete accept/reject outcomes are available primarily for ICLR, which serves as the main sample for acceptance models.

Table 2 Notations and explanations of independent variables.

Notation	Explanation
N.Author	Number of authors.
US	Indicator for whether the first author’s affiliation is in the United States.
Student	Indicator for whether the first author is a student.
University	Indicator for whether the first author is affiliated with a university.
Female	Indicator for whether the first author is female.
N.Formu	Number of formulas in the full paper.
N.FormuMain	Number of formulas in the main body.
N.Ref	Number of literature references.
N.RefAll	Total number of references.
N.Fig	Number of figures in the full paper.
N.FigMain	Number of figures in the main body.
N.Tab	Number of tables in the full paper.
N.TabMain	Number of tables in the main body.
ElemTab	Number of table elements in the full paper.
ElemTabMain	Number of table elements in the main body.
LenFormu	Total length of formulas in the full paper.
LenFormuMain	Total length of formulas in the main body.
LenAbs	Abstract length.
LenTitle	Title length.
LenText	Text length in the full paper.
LenTextMain	Text length in the main body.
LenSentsAve	Average sentence length.

NeurIPS and ICML provide more limited information on rejected papers; accepted papers from these venues are used only to characterize presentation conventions among successful papers and are not treated as representative of the full submission pool.

To avoid conflating different data-generating mechanisms, we distinguish three analytic samples (Table 1). The *full-decision sample* (ICLR 2024/2025) contains submissions with both accepted and rejected outcomes observable and serves as the primary basis for acceptance-related analyses. The *rating sample* includes ICLR and NeurIPS submissions with publicly available reviewer ratings and is used for score analyses. The *accepted-paper cross-venue sample* includes accepted papers from all three venues and is used only for descriptive comparisons of presentation conventions among successful papers.

We collect metadata such as author information and primary area from the OpenReview platform API (The OpenReview Team 2023). To quantify presentation features, we parsed the PDF files of all manuscripts using Grobid (Lopez 2009), a tool for extracting structured information from scholarly documents. From each paper we derived fine-grained presentation features, including the number of formulas, figures,

Table 3 Statistics of Collected Data

	Accepted	Rejected	Total
ICML2023	1828	0	1828
ICML2024	2610	0	2610
NeurIPS 2023	2357	0	2357
NeurIPS 2024	4025	200	4225
ICLR 2024	2246	4943	7189
ICLR 2025	3568	7502	11070
Total	16634	12645	29279

tables, and textual length metrics. We computed these statistics separately for the full paper and for the main body only (i.e., excluding appendices and references). The constructed features are listed in Table 2. Continuous variables exhibiting right-skewed distributions (e.g., formula counts: $M = 33.4$, $SD = 48.8$) are log-transformed to approximate normality. Overall, the dataset contains 29,279 examples.

For each submission from ICLR and NeurIPS, the dataset includes the following attributes: paper identifier, primary research area, number of authors, author-level information (e.g., academic position, affiliation region, gender), title, abstract, keywords, the presentation features listed in Table 2, reviewer ratings and scores (overall rating, soundness, presentation, contribution, and reviewer confidence), and the final decision (oral, spotlight, poster, or reject). For ICML submissions, all of the above information is available except for the reviewer ratings and scores. Thus, for analyses involving reviewer ratings, we restrict the sample to ICLR and NeurIPS submissions ($N = 24,834$). An illustrative example entry is provided in D.

Gender variable. The gender of the first author is inferred from first names using a gender inference tool. We acknowledge that inferred gender is subject to measurement error: some names are ambiguous, some authors may not identify with binary gender categories, and the inference tool may misclassify names from certain cultural backgrounds. The female indicator in our dataset therefore represents an approximate binary classification rather than a definitive gender identity. Given these measurement limitations, we treat gender-related results as exploratory and interpret them cautiously.

Because the disclosure of rejected submissions varies across conferences and years, our analyses use different samples depending on the outcome of interest. We do not assume that accepted-only observations from NeurIPS or ICML represent the full submission pool of those venues. Instead, these observations are used to enrich the characterization of presentation patterns among successful papers and to examine directional consistency across top-tier ML venues. The strongest evidence for acceptance-related associations comes from samples where both accepted and rejected submissions are available. This distinction is important for interpreting the pooled models: they are informative about broad regularities in the observed data but should not be read as unbiased estimates of the complete cross-venue acceptance process.

A key limitation is that not all venues disclose rejected submissions. Consequently, the pooled dataset does not constitute a random sample of all submitted papers across the three conferences. We therefore avoid interpreting cross-venue pooled acceptance models as estimating a universal acceptance rule. Instead, we use them to examine whether the direction of associations observed in more complete decision samples remains consistent when successful papers from other comparable venues are included. Some venues disclose only accepted papers or provide incomplete rejected-paper records. Therefore, the total dataset is not a complete census of submissions for all venues. Analyses are conducted on venue-specific analytic samples depending on outcome availability.

Table 4 Descriptive statistics ($N = 24,834$) of ICLR 2024/2025 and NeurIPS 2023/2024.

Variable	mean	std.dev.	2.5%	97.5%
<i>N.Author</i>	5.20	2.76	2	11
<i>US</i>	0.24	0.42	0	1
<i>Student</i>	0.37	0.48	0	1
<i>University</i>	0.77	0.42	0	1
<i>Female</i>	0.10	0.30	0	1
<i>N.Formu</i>	33.40	48.80	0	178
<i>N.FormuMain</i>	13.30	10.59	0	39
<i>LenFormu</i>	3505	7138	0	22496
<i>LenFormuMain</i>	1225	3752	0	4091
<i>N.Tab</i>	6.75	5.26	0	19
<i>N.TabMain</i>	2.99	2.19	0	8
<i>ElemTab</i>	273.6	450.3	0	1121
<i>ElemTabMain</i>	109.8	128.5	0	440
<i>N.RefAll</i>	119	54.1	39	245
<i>N.Ref</i>	42.94	18.54	14	87
<i>LenTitle</i>	73.54	20.25	34	116
<i>LenAbs</i>	1346	305	786	1984
<i>N.Fig</i>	8.64	6.18	1	24
<i>N.FigMain</i>	5.07	3.60	0	14
<i>LenText</i>	51073	19884	25688	97393
<i>LenTextMain</i>	29397	4516	21731	36710
<i>LenSentsAve</i>	133.9	16.3	51.8	203.6

4.2 Exploratory analysis of independent variables

We compare the overall statistics of the collected data from ICLR2024/2025 and NeurIPS2023/2024. The summary table of the selected variables from the metadata and presentation features is given by Table 4. We provide the average, standard deviation, 2.5% and 97.5% quartiles of each variable in the collected data. By comparing the quartiles with the mean, we can learn that most of the full-paper quantitative

statistics are right skewed with long tails. This indicates that logarithm or other transformations (e.g., Box-Cox) can be applied to convert the original distribution into a near-normal distribution for these variables.

All manuscript-level features are measured prior to review and are therefore temporally prior to reviewer ratings and final decisions.

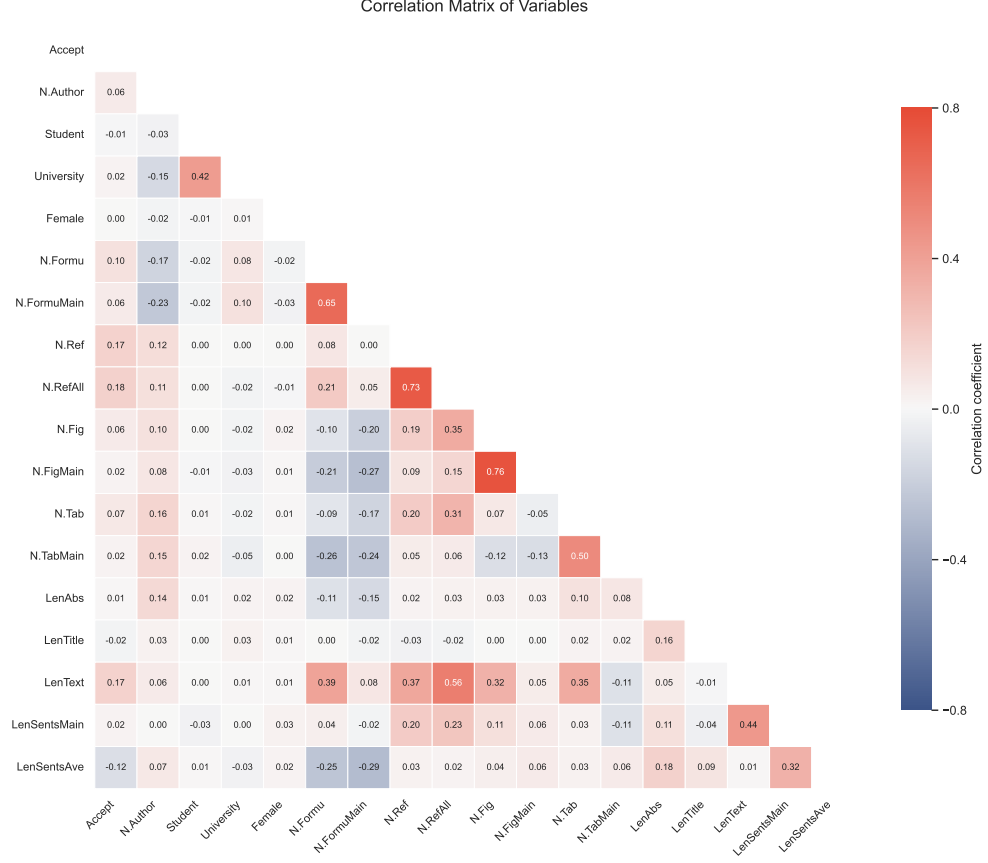


Fig. 2 Pairwise correlations among selected author-level and manuscript-level variables. Correlations are descriptive and do not adjust for venue, year, topic or semantic content.

To motivate our subsequent analyses and provide an initial characterization of the relationships between the attributes of the manuscript and the results of the peer review, we calculated pairwise Pearson correlations for the full set of independent variables and the final acceptance decision, as depicted in Figure 2. This exploratory analysis reveals several patterns that align with and extend previous descriptive work. Figure 2 presents pairwise correlations among the main manuscript-level and author-level variables used in the analysis. The correlations show that acceptance is positively

associated with several indicators of manuscript elaboration and collaboration, including the number of authors and the presence of mathematical or visual elements, while it is negatively associated with several length-related indicators such as title length and average sentence length. Critically, we observe non-trivial correlations between meta-data and presentation-related features and review outcomes. The number of authors (**N.Author**) shows a consistent, although modest, positive correlation with review outcomes ($0.04 \leq r \leq 0.12$), consistent with a collaborative capacity interpretation.

Regarding presentation features, both the volume of mathematical content and visual elements exhibit differential correlations with acceptance. The number of formulas in the main body (**N.FormuMain**) is positively correlated with acceptance ($r \approx 0.10$), while the number of figures (**N.Fig**) also shows positive correlations. In contrast, **LenTitle** exhibits a consistent negative correlation with acceptance, suggesting that conciseness in titling is associated with more favorable evaluations. Although these correlations provide preliminary evidence that formatting and metadata are intertwined with perceived quality, they cannot distinguish whether this is due to a direct association with reviewer assessments or merely a reflection of higher-quality work possessing different stylistic traits. Furthermore, the generally low correlations between the presentation features themselves (e.g., $r = 0.28$ between **N.FormuMain** and **N.FigMain**) indicate that multicollinearity is not a primary concern for subsequent regression analyses. Their distinct correlational profiles underscore the need for a more refined analytical investigation into their specific patterns.

4.3 Exploratory and visual analysis of acceptance decisions

4.3.1 Associations between metadata and acceptance decisions

We compared acceptance rates across different research areas, first-author positions, and institutional regions, as illustrated in Figure 3. Among research areas, theory (Learning Theory) exhibited the highest acceptance rate (approximately 49.0%), while reinforcement learning (33.0%) and X-supervised learning (26.8%) showed lower rates. Generative models, meanwhile, accounted for the largest share of oral presentations.

These patterns may reflect variations in the intensity and efficiency of research in different stages of the career. Regarding first-author position, papers led by assistant professors achieved the highest observed acceptance rate (34.5%), while undergraduate and master’s students had the lowest rate (26.4%). Raw group-level differences attenuate after adding semantic-content proxy and presentation covariates. Geographically, first authors affiliated with institutions in the United States and the European Union had the highest observed acceptance rates (34.8% and 34.9%, respectively), while those from China (including Hong Kong, Macau and Taiwan) had the lowest rate (30.3%). The British Commonwealth countries and other East Asian regions were in between. These raw group differences are treated as descriptive patterns rather than evidence of reviewer bias, author quality differences, or causal effects; they may partly reflect unobserved topic composition, institutional resources, language conventions, and sample-selection mechanisms.

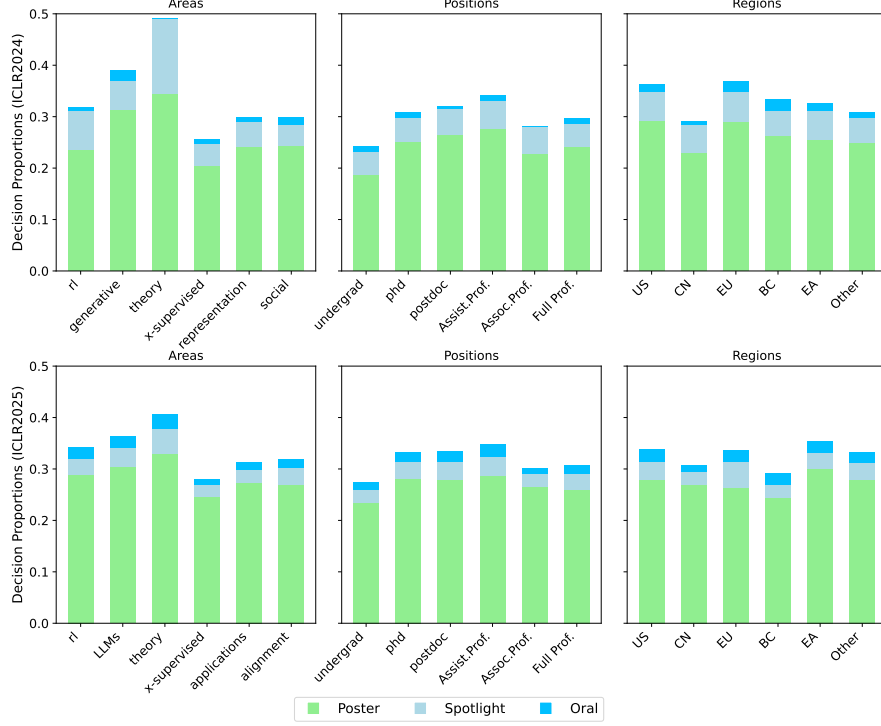


Fig. 3 The comparison of the final decision proportions of ICLR 2024&2025 based on primary areas, positions of the first authors and regions of the first author’s institution. The segments from bottom to top of each bar are the proportions of Accept (poster), Accept (spotlight) and Accept (oral), respectively.

4.3.2 Associations between presentation features and acceptance decisions

Figure 4 compares the distribution of several presentation features between accepted and rejected papers in ICLR 2024/2025. In general, accepted papers tended to contain more extensive content (longer full text, more references, and a larger total number of formulae), suggesting a larger volume of work. In particular, accepted papers had a higher number of automatically detected formulae in the main body and a greater total formula length, although the presence of numerous outliers indicates substantial heterogeneity. Regarding figures and tables, accepted papers contained more figures and tables in the full paper, yet the number of tables in the main body was slightly lower than in rejected papers. The variance of most presentation features was larger among accepted papers, reflecting greater stylistic flexibility.

Figure 5 presents the linear associations between the four presentation features and manuscript acceptance probability, based on binned empirical effect points. Linguistic variables exhibit the strongest negative linear correlations with acceptance probability across all evaluation outcomes: title length and average sentence length. The binned

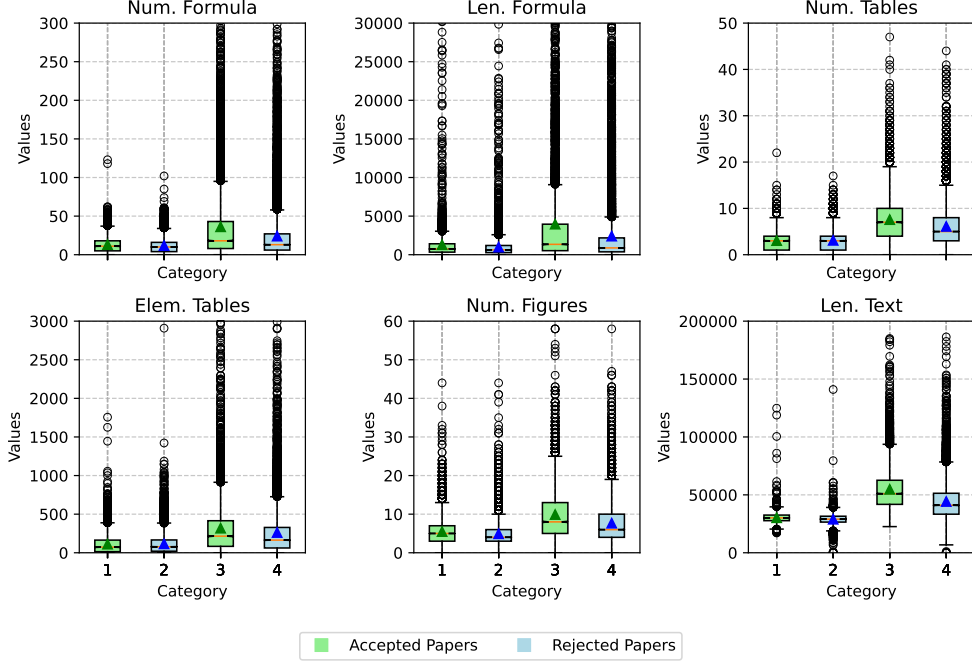


Fig. 4 Comparison of presentation features between accepted and rejected papers in ICLR 2024&2025. The light green boxes and light blue boxes are for accepted and rejected papers. The left two boxes are for main-body statistics, and the right two boxes are for statistics of the full paper.

empirical effect points are visually consistent with downward-sloping trend lines, suggesting a monotonic negative association across the observed range of variable values. For formal and visual exposition variables, the number of figures and the number of formulas in the main body show robust positive linear correlations with acceptance probability. Binned effect points track the upward-sloping trend lines for both variables. Correlations are computed on binned conditional means and are intended only as exploratory visual summaries. Extra results about the association between presentation features and ratings are provided in the online supplement.

5 Empirical Results and Sensitivity Analyses

This section reports adjusted associations between manuscript-level features and peer-review outcomes. All estimates are interpreted as observational associations rather than causal effects. Rating analyses use the rating sample (ICLR and NeurIPS submissions with available reviewer ratings); acceptance analyses are restricted to the ICLR full-decision sample. Cross-venue accepted-only data are used only for descriptive comparisons in the appendix.

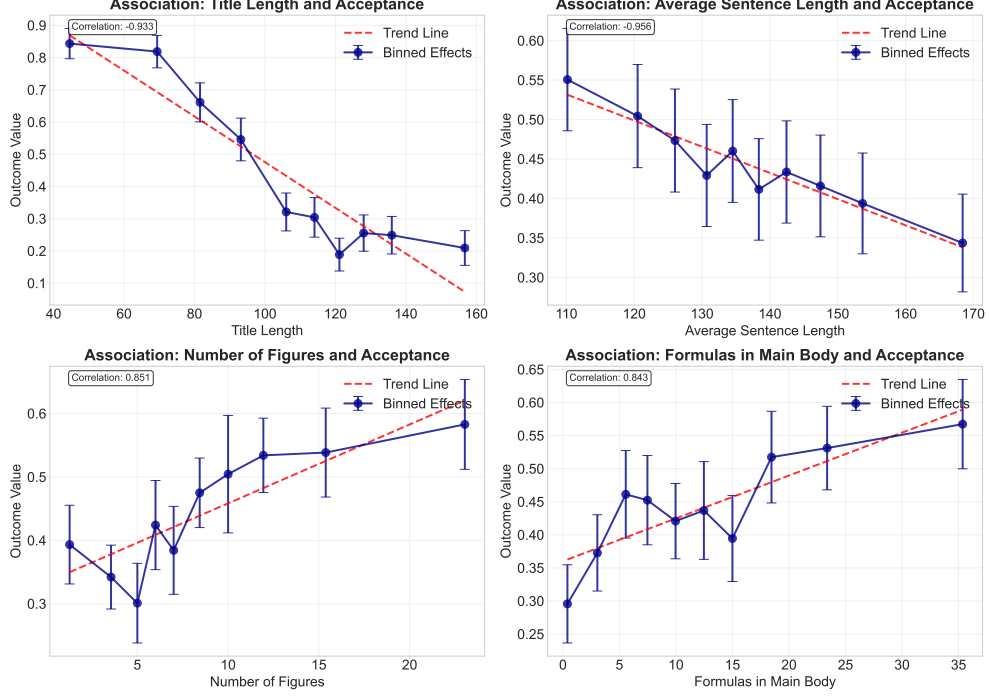


Fig. 5 Binned mean associations between core presentation features and acceptance in the ICLR full-decision sample. Pearson correlation coefficients are calculated across binned means and are intended for visualization only; they should not be interpreted as paper-level correlations or causal effects.

5.1 Regression analysis for ratings

Table 5 reports associations with review scores among submissions from ICLR 2024/2025 and NeurIPS 2023/2024 for which review ratings are available ($N = 24,834$). To quantify the association between metadata and presentation-related features and review outcomes, we estimate a multiple linear regression model where the dependent variable is the average rating received by a paper:

$$\text{AveRating}_i = \beta_0 + \sum_{k \in \mathcal{K}_{\text{meta}}} \beta_k X_{ik}^{\text{meta}} + \sum_{j \in \mathcal{J}_{\text{pres}}} \gamma_j X_{ij}^{\text{pres}} + \epsilon_i, \quad (1)$$

where $\mathcal{K}_{\text{meta}}$ denotes the set of metadata variables (e.g., author count, affiliation region) and $\mathcal{J}_{\text{pres}}$ represents presentation features (e.g., formula counts, text length metrics). Continuous variables with skewed distributions are log-transformed prior to estimation.

In this and all subsequent tables, the prefix “L.” indicates a logarithmic transformation, and underlined variables are rescaled by a factor of 10^{-5} . Several robust patterns emerge across all rating dimensions. The number of authors exhibits a strong positive association with all scores ($\beta = 0.040$, $p < 0.001$ for average rating), suggesting that collaborative work is associated with more favorable evaluations. Regarding

Table 5 Regression models of review scores among submissions with available ratings (ICLR 2024/2025 and NeurIPS 2023/2024).

Indep. Variables	Rating	Soundness	Contribution	Presentation
Const	-4.03 (0.298)***	-0.428 (0.113)***	-1.21 (0.117)***	0.133 (0.124)
N.Author	0.0402 (0.003)***	0.0118 (0.001)***	0.0155 (0.001)***	0.0134 (0.001)***
US	0.0690 (0.016)***	0.0185 (0.006)**	0.0170 (0.006)**	0.0343 (0.007)***
Student	-0.0442 (0.015)**	-0.0170 (0.006)**	-0.0142 (0.006)*	-0.0127 (0.006)*
University	-0.0544 (0.017)**	-0.0187 (0.007)**	-0.0228 (0.007)***	-0.0133 (0.007)
Female	-0.0174 (0.022)	-0.0038 (0.008)	-0.0133 (0.009)	-0.0063 (0.009)
N.Formu	0.0398 (0.020)*	0.0204 (0.007)**	0.0145 (0.008)	-0.0131 (0.008)
N.FormuMain	0.0124 (0.001)***	0.0042 (0.000)***	0.0043 (0.000)***	0.0020 (0.001)***
N.Ref	0.0049 (0.001)***	0.0018 (0.000)***	0.0029 (0.000)***	0.0024 (0.000)***
N.RefAll	0.0017 (0.000)***	0.0006 (7.8e-5)***	0.0001 (8.09e-5)	0.0006 (8.6e-5)***
N.Fig	0.0726 (0.016)***	-0.0164 (0.006)**	-0.0009 (0.006)	0.0165 (0.006)*
N.FigMain	0.0165 (0.003)***	0.0066 (0.001)***	0.0056 (0.001)***	0.0070 (0.001)***
L.N.Tab	0.119 (0.020)***	0.0139 (0.008)	0.0208 (0.008)**	0.0070 (0.008)
N.TabMain	-0.004 (0.005)	0.0015 (0.002)	9.86e-5 (0.002)	0.0029 (0.002)
L.ElemTab	-0.0256 (0.008)***	-0.0165 (0.003)***	-0.0069 (0.003)*	-0.0025 (0.003)
ElemTabMain	0.0002 (7.01e-5)*	6.8e-5 (2.65e-5)**	6.65e-5 (2.75e-5)*	7.82e-5 (2.9e-5)**
L.LenFormu	0.0109 (0.010)	0.0007 (0.004)	0.0009 (0.004)	0.0004 (0.004)
LenFormuMain	-0.945 (0.228)***	-0.154 (0.086)	-0.097 (0.089)	-0.2619 (0.095)**
LenAbs	-2.48 (2.26)	-1.68 (0.856)*	1.08 (0.886)	-2.04 (0.942)*
LenTitle	-0.0019 (0.000)***	-0.0007 (0.000)***	-0.0006 (0.000)***	-0.0009 (0.000)***
L.LenText	0.793 (0.030)***	0.295 (0.012)***	0.337 (0.012)***	0.224 (0.013)***
LenTextMain	1.33 (0.199)***	0.161 (0.075)*	0.0539 (0.078)	0.535 (0.083)***
LenSentsAve	-0.0047 (0.000)***	-0.0022 (0.000)***	-0.0021 (0.000)***	-0.0016 (0.000)***
Adj- R^2	0.224	0.180	0.184	0.125
F-stats	326.3	248.5	254.7	162.0
AIC	7.22e+04	2.39e+04	2.57e+04	2.87e+04
BIC	7.23e+04	2.41e+04	2.58e+04	2.89e+04

Notes: Coefficients are adjusted associations. ***, **, and * indicate 0.001, 0.01, and 0.05 levels of significance, respectively.

textual features, title length is consistently associated with lower ratings ($\beta = -0.002$, $p < 0.001$), while full-text length is strongly positively associated ($\beta = 0.793$, $p < 0.001$). Average sentence length is negatively associated with all evaluation metrics ($p < 0.001$), consistent with the interpretation that shorter sentences are associated with more favorable evaluations.

For mathematical content, a larger number of automatically detected main-body formulas (**N.FormuMain**) is consistently associated with more favorable scores ($\beta = 0.012$ for rating, $p < 0.001$), whereas total formula count shows limited significance. Visual elements exhibit differential patterns: main-body figures (**N.FigMain**) are positively associated with all outcomes, while table-related variables show mixed effects.

The model explains approximately 22% of the variance in overall ratings (adjusted $R^2 = 0.224$), indicating that observable metadata and presentation-related features contain substantial information about reviewer evaluations.

Table 6 Logistic regression models of acceptance in the ICLR full-decision sample.

	ICLR2024	ICLR2025
Const	-20.3 (1.48)***	-13.3 (1.09)***
N.Author	0.0791 (0.012)***	0.0493 (0.008)***
US	0.2129 (0.065)***	0.0965 (0.051)
Student	-0.0703 (0.065)	0.0021 (0.050)
University	-0.1725 (0.071)*	-0.108 (0.058)
Female	0.0458 (0.091)	0.0584 (0.072)
N.Formu	-0.0340 (0.086)	0.120 (0.064)
N.FormuMain	0.0188 (0.006)***	0.0152 (0.004)***
N.Ref	0.0062 (0.002)*	0.0008 (0.002)
N.RefAll	0.0040 (0.001)***	0.0054 (0.001)***
L.N.Fig	0.0652 (0.069)	0.219 (0.053)***
N.FigMain	0.0283 (0.012)*	0.0166 (0.009)
L.N.Tab	0.229 (0.085)**	0.106 (0.066)
N.TabMain	-0.0029 (0.019)	0.0077 (0.015)
L.ElemTab	-0.0496 (0.032)	0.0181 (0.026)
ElemTabMain	8.44e-6 (0.000)	-0.0002 (0.000)
L.LenFormu	0.0435 (0.043)	-0.0120 (0.031)
<u>LenFormuMain</u>	0.469 (1.080)	-0.446 (1.074)
<u>LenAbs</u>	5.87 (9.52)	5.44 (7.46)
LenTitle	-0.0041 (0.001)**	-0.0039 (0.001)***
L.LenText	1.75 (0.150)***	0.9937 (0.111)***
<u>LenTextMain</u>	-0.919 (0.848)	1.09 (0.687)
LenSentsAve	-0.0046 (0.002)*	-0.0029 (0.001)*
Pseudo R-squ.:	0.1221	0.08990
LLR p-value:	1.279e-216	2.179e-249

Notes: Coefficients are adjusted associations. ***, **, and * indicate 0.001, 0.01, and 0.05 levels of significance, respectively. These results are based on the ICLR full-decision sample. Models include year and primary-area fixed effects. Fixed-effect coefficients are omitted for readability.

5.2 Acceptance associations in the ICLR full-decision sample

The primary inferential sample for acceptance outcomes is the ICLR full-decision sample (ICLR 2024 and ICLR 2025); accepted-only cross-venue data are used only for descriptive comparison of presentation conventions among successful papers.

The models estimate adjusted associations between observable pre-review features and peer-review outcomes. Reviewer ratings are analyzed as evaluation outcomes rather than as predictors of final decisions in the main analysis. We complement the rating analysis by modeling the final acceptance decision using logistic regression. Let p_i denote the probability of acceptance for paper i . The log-odds are specified as a linear function of the same metadata and presentation features:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_{k \in \mathcal{K}_{\text{meta}}} \beta_k X_{ik}^{\text{meta}} + \sum_{j \in \mathcal{J}_{\text{pres}}} \gamma_j X_{ij}^{\text{pres}}. \quad (2)$$

Table 6 reports the results for the ICLR full-decision sample, where both accepted and rejected submissions are observable. The primary inferential scope of these models is limited to ICLR 2024/2025; we do not generalize these findings to NeurIPS or ICML, where rejected submissions are largely unobserved. Within the ICLR full-decision sample, the findings largely mirror those from the rating analysis. Author count remains a robust positive covariate ($\beta = 0.049$ to 0.079 , $p < 0.001$), a larger number of automatically detected main-body formulas (**N.FormuMain**) is associated with higher predicted probabilities of acceptance ($\beta = 0.015$ to 0.019 , $p < 0.001$), title length shows a negative association ($\beta = -0.004$, $p < 0.01$), and full-text length (**L.LenText**) is strongly associated with acceptance ($\beta = 0.99$ to 1.75 , $p < 0.001$).

Notably, the association of US affiliation is significant in ICLR2024 ($\beta = 0.213$, $p < 0.001$) but attenuates in ICLR2025 ($\beta = 0.097$, $p = 0.058$), suggesting that institutional-level associations may be context-dependent and correlated with observed semantic-content proxy and presentation differences. These cross-year differences underscore that even within the same venue, associations may vary with changing submission pools and review conditions.

To aid interpretability, we translate selected logistic coefficients into odds ratios. In the ICLR full-decision sample, each additional author is associated with an odds ratio of approximately $\exp(0.049)$ to $\exp(0.079) \approx 1.05$ – 1.08 , corresponding to a 5%–8% higher predicted odds of acceptance per additional author in the fitted models. A 10-character increase in title length corresponds to approximately $\exp(-0.04) \approx 0.96$ times the odds of acceptance, corresponding to a 4% lower predicted odds of acceptance per 10 additional characters. Supplementary directional consistency checks using partially observed cross-venue data are provided in Appendix B; these should be interpreted as exploratory only, because accepted-only observations from other venues do not permit acceptance-related inference.

Because of the large sample size, statistical significance should not be equated with substantive importance. Most associations are modest in magnitude and should be interpreted as weak to moderate communicative signals rather than deterministic predictors of acceptance.

5.3 Regression analysis with a text-based semantic-content proxy

Table 7 adds the SciBERT-based text proxy to assess whether manuscript-structure associations persist after adjusting for topical-semantic variation observable from titles

and abstracts. The proxy is an abstract-level text measure, not a proxy for intrinsic scientific quality; it may capture topic, framing, and writing style, while leaving substantial unobserved quality variation.

The model is trained on the corpus of submissions from ICLR 2024/2025 and NeurIPS 2023/2024, divided into training, validation, and test sets (80:10:10). We use the Adam optimizer with differential learning rates: 2×10^{-5} for the pre-trained SciBERT parameters and 2×10^{-4} for the linear prediction head attached to the “[CLS]” token embedding. Training runs for five epochs with a weight decay of 0.01 and an effective batch size of 64.

On the held-out test set ($N = 2,481$), the SciBERT model achieves the following performance: Pearson $r = 0.49$, $R^2 = 0.24$, RMSE = 1.28, MAE = 1.02. These metrics indicate a moderate correlation between the proxy and actual reviewer ratings, leaving substantial residual variation that may reflect content quality, novelty, or other unobserved factors.

We augment our baseline regression models by including the SciBERT-predicted rating as an additional covariate. Because the proxy is trained to predict reviewer ratings, its coefficient is not interpreted as an independent substantive finding. Crucially, these regressions are estimated exclusively on the held-out test set, which was never used during fine-tuning. This reduces the risk of training-test leakage in constructing the proxy.

Table 7 reports the results. Model 1 in each pair replicates the baseline specification (without content control) on this smaller test sample; Model 2 includes Pred. The predicted rating enters both models with strong statistical significance. For the rating equation, the coefficient $\beta = 0.438$ ($p < 0.001$) indicates that a one-point increase in the predicted score is associated with a 0.438-point increase in the actual average rating, holding other factors constant. For the acceptance equation, the coefficient $\beta = 0.708$ ($p < 0.001$) represents the change in the log-odds of acceptance. The inclusion of Pred improves model fit: adjusted R^2 for ratings increases from 0.274 to 0.309, and pseudo R^2 for acceptance increases from 0.151 to 0.171.

Notably, several presentation-related features retain their significance and magnitude even after the semantic-content proxy is included. The number of authors remains robustly positive for both ratings ($\beta = 0.051$, $p < 0.001$) and acceptance ($\beta = 0.098$, $p < 0.001$).¹ Similarly, automatically detected main-body formulas (N.FormuMain), figures (N.Fig), full-text length (L.LenText), and title length (LenTitle) all maintain their directions and significance levels. **The stability of these key coefficients before and after inclusion of the SciBERT proxy (e.g., N.Author: 0.054 to 0.051 for ratings; N.FormuMain: 0.020 to 0.019 for ratings) suggests that the observed associations are at least partially robust to the inclusion of the available semantic-content proxy, although unobserved dimensions of quality may still contribute.**

These results suggest that the associations between presentation-related features and review outcomes are not fully explained by abstract-level text variation. However,

¹For the logistic regression, the coefficient 0.098 corresponds to an odds ratio of $\exp(0.098) \approx 1.10$, implying a 10% higher predicted odds of acceptance per additional author in the fitted model. The positive association may reflect a capacity channel: larger teams have more diverse expertise and greater opportunity for internal revision before submission.

Table 7 Regression models with and without the SciBERT-based semantic-content proxy (held-out test set).

Variable	Rating		Acceptance	
	Model 1	Model 2	Model 1	Model 2
Const	-4.056(1.098)***	-5.288(1.078)***	-13.845(2.806)***	-16.169(2.881)***
N.Author	0.054(0.010)***	0.051(0.010)***	0.100(0.022)***	0.098(0.022)***
US	0.016(0.060)	-0.021(0.059)	0.025(0.132)	-0.027(0.134)
Student	0.002(0.062)	-0.007(0.061)	0.098(0.140)	0.100(0.143)
University	-0.173(0.064)**	-0.121(0.063)	-0.138(0.143)	-0.065(0.145)
Female	0.007(0.085)	0.024(0.083)	0.237(0.186)	0.269(0.189)
Assist.Prof.	0.153(0.092)	0.140(0.090)	0.363(0.204)	0.376(0.207)
Assoc.Prof.	-0.038(0.096)	-0.014(0.094)	0.050(0.215)	0.094(0.218)
N.Formu	0.002(0.001)	0.001(0.001)	-0.001(0.002)	-0.002(0.002)
N.FormuMain	0.020(0.005)***	0.019(0.004)***	0.032(0.010)**	0.032(0.010)**
N.Ref	0.001(0.002)	0.001(0.002)	0.000(0.004)	-0.001(0.004)
N.RefAll	0.001(0.002)	0.001(0.002)	0.002(0.003)	0.003(0.003)
N.Fig	0.024(0.007)**	0.024(0.007)**	0.045(0.016)**	0.045(0.016)**
N.FigMain	0.004(0.012)	-0.001(0.011)	-0.010(0.025)	-0.016(0.025)
L.N.Tab	0.096(0.073)	0.122(0.071)	-0.028(0.163)	0.034(0.166)
N.TabMain	0.015(0.018)	0.007(0.017)	0.034(0.040)	0.021(0.041)
L.ElemTab	0.015(0.028)	0.028(0.027)	0.080(0.064)	0.098(0.065)
ElemTabMain	12.633(28.708)	12.923(27.997)	-68.240(65.255)	-69.248(66.067)
L.LenFormu	0.037(0.019)	0.028(0.018)	0.100(0.047)*	0.084(0.047)
<u>LenFormuMain</u>	-2.060(1.480)	-2.162(1.443)	-7.440(3.942)	-7.641(3.989)
<u>LenAbs</u>	9.248(8.635)	2.275(8.452)	27.702(19.574)	14.360(20.144)
LenTitle	-0.007(0.001)***	-0.006(0.001)***	-0.022(0.002)***	-0.020(0.002)***
L.LenText	0.782(0.116)***	0.674(0.114)***	1.217(0.285)***	1.066(0.290)***
<u>LenTextMain</u>	1.353(0.788)	1.406(0.768)	-2.020(1.737)	-1.906(1.762)
LenSentsAve	-0.003(0.002)	-0.002(0.002)	0.000(0.004)	0.002(0.004)
Pred		0.438(0.045)***		0.708(0.106)***
R-squared:	0.283	0.319	-	-
Adj. R-squared:	0.274	0.309	-	-
Pseudo R-squared:	-	-	0.151	0.171
AIC	5425.2	5334.7	1981.8	1937.7
BIC	5562.9	5477.9	2119.5	2080.9

Notes: All models include year and primary-area fixed effects; rating models additionally include venue fixed effects. Fixed-effect coefficients are omitted for readability. Coefficients are adjusted associations. ***, **, and * indicate 0.001, 0.01, and 0.05 levels of significance, respectively.

we cannot conclude that these structural variables are entirely independent of paper content or research quality. The proxy’s moderate predictive performance ($R^2 = 0.24$) indicates substantial residual variation that may reflect unobserved dimensions of quality.

In contrast, some variables that were marginally significant in earlier full-sample regressions lose significance once the semantic-content proxy is included. For example, the US affiliation indicator becomes statistically indistinguishable from zero in both equations, and the university indicator loses significance in the rating model. This suggests that certain institutional associations observed in Section 5.1 and 5.2 may partially reflect correlated content and presentation quality rather than persisting as an independent group-level association after observed adjustments. The average sentence length (`Len.Sents.Ave`) also becomes insignificant in this subsample, possibly due to the reduced sample size or because sentence-level readability is partially captured by the SciBERT embedding of the abstract.

We note that the coefficient magnitudes in Table 7 differ from those reported in Tables 6 and Tables 5 because the regressions here are estimated on the smaller SciBERT test set rather than the entire corpus. However, the direction and relative importance of the key predictors remain consistent, strengthening the robustness of our findings.

5.4 Covariate-balancing sensitivity analysis for binary variables

To assess whether the observed associations are sensitive to imbalance in observed covariates, we employ propensity-score matching as a covariate-balancing sensitivity check. We match papers that share similar presentation features and SciBERT-predicted ratings, reducing imbalance in observed confounders. The matching analyses are conducted on the held-out test subset ($N \approx 2,481$) for which the SciBERT proxy is available.

Our matching procedure achieved effective covariate balance. The Standardized Mean Differences (SMD) for all key features fell below the 0.1 threshold. Within the limits of observed covariate adjustment, we find that most examined author characteristics do not yield a statistically significant adjusted association with acceptance decisions in the ICLR full-decision sample. For example, the significant US-affiliation association observed in regression vanishes after matching. Similarly, we found no significant adjusted associations for student status or inferred gender. The association of being an assistant professor remains marginally significant ($p \approx 0.05$), which suggests a potential though weak link for this specific group that warrants further investigation.

Overall, these findings suggest that, within the limits of observed covariate adjustment, author-identity variables exhibit limited independent explanatory power for review outcomes in the ICLR full-decision sample. The attenuation of raw group differences after matching suggests that these differences are at least partly explained by observable differences in semantic-content proxies and presentation-related characteristics, rather than reflecting reviewer preferences independent of observable characteristics. This does not rule out potential fairness concerns or structural inequality, but it suggests that raw disparities should be interpreted with caution.

Table 8 Covariate-balancing results for binary metadata variables using propensity-score matching

Variable	N		Rating				Acceptance			
	Focal	Comp.	Diff.	S.E.	SMD ₀	SMD ₁	Diff.	S.E.	SMD ₀	SMD ₁
US	591	1998	-0.07	0.066	0.048	0.024	-0.026	0.029	0.048	0.024
Student	930	1659	-0.082	0.055	0.036	0.015	-0.006	0.023	0.036	0.015
Female	252	2337	0.09	0.101	0.104	0.041	0.064	0.045	0.104	0.041
University	1580	1009	-0.077	0.052	0.066	0.018	0.01	0.023	0.066	0.018
Assist.Prof	282	2307	0.181	0.093	0.13	0.075	0.082	0.042	0.13	0.075
Associate.Prof	214	2375	0.07	0.109	0.043	0.048	0.023	0.048	0.043	0.048

Notes: Estimates are covariate-balanced associations after balancing observed covariates, including year and primary-area fixed effects (coefficients omitted for readability).

5.5 Covariate-balancing sensitivity analysis for continuous variables

For continuous focal variables, we employ the Generalized Propensity Score (GPS) framework combined with inverse-probability weighting as a covariate-balancing sensitivity check. The GPS analyses are conducted on the held-out test subset ($N \approx 2,481$) for which the SciBERT proxy is available. We consider a comprehensive set of base covariates, including venue, *N.Author*, *Student*, *Female*, *US*, *University*, *LenAbs*, *LenText*, and the SciBERT-predicted rating.

The analysis proceeds in three steps. First, the GPS is estimated by modeling the conditional distribution of each focal variable given the covariates via Ridge regression. Second, we calculate stabilized weights based on the ratio of marginal to conditional densities, applying trimming at the 5th and 95th percentiles. Finally, we estimate GPS-weighted marginal associations using Weighted Least Squares (WLS) for ratings and weighted Generalized Linear Models (GLM, Binomial family) for acceptance decisions in the ICLR full-decision sample. Table 9 reports the estimated associations, standard errors, odds ratios (OR), and the post-weighting SMDs.

The covariate-balancing sensitivity checks reveal several adjusted associations that are consistent with presentation-related patterns not fully absorbed by observed controls. Among presentation-related features, a larger number of automatically detected main-body formulae (*N.FormuMain*) shows a significant adjusted association with both ratings (association = 0.018) and acceptance (OR = 1.023) in the ICLR full-decision sample, while the total number of formulae (*N.Formu*) shows a negligible association. This asymmetry is consistent with the interpretation that main-body mathematical exposition is more strongly associated with favorable review outcomes than formulae appearing elsewhere, although this pattern may also reflect paper-type differences. Similarly, the number of figures (*N.Fig*) and tables (*N.Tab*) produce positive and significant adjusted associations (OR = 1.021 and 1.035, respectively), indicating that richer visual and empirical evidence is associated with more favorable evaluation outcomes after balancing observed covariates.

Table 9 Covariate-balanced associations for continuous communicative-structural variables

Variable	Effective N		Rating			Acceptance			
	Lower range	Higher range	Assoc.	S.E.	SMD after weighting	Assoc.	OR	S.E.	SMD after weighting
N.Formu	1120	1142	0.003	0.0004	0.087	0.002	1.002	0.0004	0.087
N.FormuMain	1042	1220	0.018	0.002	0.039	0.022	1.023	0.004	0.039
N.Ref	1108	1154	0.003	0.0003	0.052	0.005	1.005	0.001	0.052
N.Fig	1094	1168	0.019	0.003	0.044	0.021	1.021	0.005	0.044
N.Tab	1026	1236	0.021	0.004	0.053	0.034	1.035	0.006	0.053
N.ElemTab	1131	1131	0.000	<0.0001	0.022	0.000	1.000	<0.0001	0.022
N.FigMain	860	1402	0.013	0.006	0.025	0.013	1.014	0.009	0.025
N.TabMain	836	1426	0.004	0.011	0.023	0.007	1.007	0.019	0.023
ElemTabMain	1131	1131	0.000	<0.0001	0.014	0.000	1.000	<0.0001	0.014
LenTextMain	1131	1131	0.000	<0.0001	0.060	0.000	1.000	<0.0001	0.060
LenSentsAve	1131	1131	-0.003	0.001	0.016	-0.008	0.992	0.003	0.016
LenFormu	1131	1131	0.000	<0.0001	0.098	0.000	1.000	<0.0001	0.098
LenFormuMain	1131	1131	0.000	<0.0001	0.039	0.000	1.000	<0.0001	0.039

Notes: Estimates are covariate-balanced marginal associations from the ICLR full-decision sample. Models include year and primary-area fixed effects (coefficients omitted for readability). Lower/higher range indicates the effective sample sizes in the lower and upper ranges of each focal variable distribution after weighting.

Notably, the average sentence length (*LenSentsAve*) exhibits a negative adjusted association with ratings (association = -0.003 , S.E. = 0.001) and acceptance (OR = 0.992 , S.E. = 0.003). This is consistent with the possibility that conciseness may be associated with review outcomes, although sentence length is a coarse proxy for readability and does not capture argument quality.

In contrast, several variables, including *N.ElemTab*, *LenTextMain*, and *LenFormu*, yield adjusted associations of effectively zero. The null association for main-body text length is notable given its significance in earlier regression specifications. This discrepancy suggests that main-body length likely captures content-related variation in simple regressions but shows no independent association once content is controlled. The post-weighting SMDs are uniformly below the 0.10 threshold, validating the balance achieved by the weighting procedure.

6 Discussion

6.1 Main findings

This study shows that several manuscript-level communicative and structural features are systematically associated with peer-review ratings and decisions in observable OpenReview data from major AI conferences. In particular, larger author teams,

shorter titles, shorter average sentences, and more visible main-body mathematical or visual elements tend to co-occur with more favorable review outcomes in the ICLR full-decision sample and the rating sample. These associations are statistically robust across multiple regression specifications and persist after adjusting for a SciBERT-based semantic-content proxy derived from titles and abstracts.

Although several coefficients are statistically significant, their magnitudes should be interpreted as incremental associations rather than deterministic predictors of acceptance. For example, the odds ratio associated with one additional main-body formula is modest ($OR \approx 1.02$), and the association between title length and review score, while statistically robust, is small in substantive terms. We therefore interpret these patterns as weak to moderate communicative signals rather than as practically decisive factors in peer review.

The associations for visual and tabular elements are less uniform across specifications than those for author count, title length and sentence length. We therefore interpret them as suggestive evidence that visible evidential organization may be related to review outcomes, rather than as robust evidence of a universal presentation association.

6.2 Possible interpretations

These associations are consistent with several possible interpretations, but none can be confirmed causally with the present observational data. Larger teams may have greater research capacity, division of labor, access to complementary expertise or institutional resources; however, author count may also be associated with unobserved project scale, topic choice or network position. Concise titles and shorter sentences may reduce processing costs for reviewers operating under time constraints; at the same time, these measures are crude proxies for readability and do not capture argument quality or scientific novelty. Main-body formulas, figures, and tables may make technical evidence more visible and inspectable; nevertheless, these elements may also be associated with paper type, methodological tradition or empirical maturity.

The Elaboration Likelihood Model ([Petty and Cacioppo 1986](#)) provides one possible interpretive lens: when evaluators lack the capacity to process all arguments centrally, peripheral cues such as presentation fluency may become more salient. Similarly, Signaling Theory ([Spence 1978](#); [Connelly et al. 2011](#)) suggests that observable features prominently displayed in the main body may function as signals under information asymmetry. We do not claim to test these mechanisms directly; rather, our findings are consistent with their predictions in the context of high-stakes expert evaluation.

6.3 Boundary conditions and limitations

The results should not be read as prescriptive advice to manipulate manuscript style, nor as evidence of reviewer bias. They instead document observable regularities in how manuscript presentation, author-team characteristics, and review outcomes co-vary in a specific high-stakes conference-review environment.

Several boundary conditions limit the interpretation. Our SciBERT-based text proxy captures only abstract-level semantic and framing signals; it cannot measure novelty, correctness, rigor, or theoretical contribution. Residual unobserved confounding likely remains, and the generalizability of our findings may depend on field-specific conventions (Nair and Gibbert 2016; Guo et al. 2018).

The findings have fairness implications, but they do not by themselves identify unfairness. If communicative-structural cues are correlated with institutional resources, language background or access to mentoring, then evaluation outcomes may partly reflect inequalities in the capacity to produce legible and conventionally persuasive manuscripts. Future research with stronger identification designs would be needed to distinguish legitimate evaluation of clarity from inequitable sensitivity to presentation conventions.

Our data cover 2023–2025, a period of rapid growth in AI conference submissions. Review norms during this period may differ from earlier or later years, and our findings may not generalize to journals, lower-volume fields or disciplines with different writing conventions. Not all venues disclose rejected submissions comprehensively, which introduces sample-selection considerations for cross-venue comparisons.

The structural variables extracted from PDFs should be interpreted as automated proxies rather than error-free measurements; GROBID-based parsing may misidentify formulas, figures, tables, or main-body boundaries in conference PDFs with heterogeneous LaTeX templates. In a manual evaluation of randomly selected 60 papers, we found that some GROBID-derived quantities (e.g., LenFormuMain, ElemTabMain) tend to be lower than the true values in many papers, while the overall correlations between the estimated and true quantities range from 0.84 to 0.99.

Despite these limitations, the consistency of the observed associations across multiple analytic approaches suggests that communicative-structural features are a stable correlate of evaluative outcomes in this empirical setting.

6.4 Comparison with prior work

Unlike previous predictive studies that focused on forecasting acceptance (Skorikov and Momen 2020; Kang et al. 2018; Rimkus et al. 2023), we emphasize interpretable adjusted associations and control for abstract-level semantic-content variation using a text-based proxy. Our approach documents whether presentation-related associations persist after balancing observable covariates. In contrast to descriptive analyses of demographic disparities (Tran et al. 2020; Ganapathy et al. 2022), we show that many institutional associations are attenuated after controlling for semantic-content proxies, while several presentation-related patterns remain. This methodological orientation aligns with recent calls for more transparent and assumption-aware analysis in computational social science (Imai and Van Dyk 2004; Caliendo and Kopeinig 2008). These findings indicate that evaluative outcomes are systematically associated with how manuscripts organize and make visible their scientific claims, without implying that reviewers consciously rely on presentation cues or that such cues are illegitimate.

Our findings also connect with the broader scientometric literature on manuscript features and scholarly impact. Jamali and Nikzad (2011) found that title characteristics predict downloads and citations, Hasan and Breunig (2021) documented positive

associations between article length and citation outcomes, and [Nair and Gibbert \(2016\)](#) showed that title structure significantly predicts citation counts in management science. Our study extends this line of inquiry by focusing on the pre-publication peer-review stage and by examining how a broader set of communicative-structural features—including sentence length, formula placement, and figure counts—are associated with expert evaluations. This shift from post-publication citation outcomes to pre-publication review outcomes provides new insights into the role of manuscript presentation in scientific evaluation.

7 Conclusion

This paper provides a large-scale observational mapping of how manuscript-level structural and communicative features are associated with peer-review ratings and decisions in OpenReview-based AI conferences. In the ICLR full-decision sample and the rating sample, larger author teams, shorter titles, shorter average sentence length, and more main-body mathematical exposition are consistently associated with more favorable review outcomes. These adjusted associations are statistically robust but modest in magnitude; they are descriptive regularities rather than causal claims, evidence of reviewer bias, or prescriptive advice.

The findings suggest that manuscript presentation and document structure are non-negligible components of the peer-review environment. For scientometric research, the results highlight both the value and the boundaries of using OpenReview data for document-level peer-review analysis. Future research should examine whether these associations reflect communication efficiency, underlying research quality, reviewer cognitive constraints, or field-specific writing conventions, using designs with stronger causal identification.

Statements and Declarations

Funding

This paper is supported by the Fundamental Research Funds for Central Universities (Grant No. D5000240054).

Competing interests

The authors declare no competing interests.

Ethics statement

This study uses publicly accessible records from OpenReview and automatically extracted manuscript-level features from publicly available paper PDFs. The study did not involve any intervention, interaction with individuals, or collection of private communications. All analyses were conducted at the aggregate level to examine associations between manuscript-level characteristics and peer-review outcomes. Author-level contextual variables, such as affiliation region, institutional type, career status, and inferred gender, were used only for statistical adjustment and exploratory

aggregate analysis. We do not make individual-level judgments about authors, reviewers, or institutions, and we do not interpret group-level differences as evidence of reviewer bias, author ability, or causal effects. Data sharing is limited to paper identifiers, derived variables, analysis code, and documentation consistent with platform terms and privacy considerations.

Data, materials and code availability

The data used in this study were collected from the OpenReview platform (<https://openreview.net>). A subset of the OpenReview paper data used in this study is provided in the supplementary materials, including paper IDs, metadata, and extracted features. Raw PDFs and review records are not redistributed where the platform’s terms restrict redistribution; users can retrieve the original records from OpenReview using the provided paper IDs. Upon acceptance, the repository will be archived on Zenodo to ensure long-term accessibility.

Author contributions

A/Prof. Renlong Jie collected the data, performed the data analysis, and wrote the first draft. Prof. Chen Chu contributed to the methodology and paper revision. A/Prof. Danyang Jia contributed to the paper revision and conceptual discussion. A/Prof. Ruibin Geng contributed to the conceptual discussion. A/Prof. Tingyu Lu contributed to the paper revision. A/Prof. Guiwen Guan contributed to the visualization. Prof. Zhen Wang contributed to the validation and resource. All authors read and approved the final manuscript.

Appendix A Comparison between venues

We compare the presentation features of accepted papers in different venues including ICLR, NeurIPS and ICML from 2023 to 2025. We calculate a subset of presentation features for the accepted papers in each venue, including the average number of authors, the average number of references, number of automatically detected formulas/tables/figures and the average length of text in the main context and appendix. The summary is given in Table A1. For each venue, we report the average values and standard deviations (in parentheses below the average values) of each quantity explained above. We learn that ICML generally includes more formulas in its main body and full paper, NeurIPS has more references, and ICLR uses more figures. Although ICML has the largest space in the main body, it uses relatively fewer figures and tables in the full paper.

Appendix B Supplementary directional checks using partially observed cross-venue data

We do not estimate pooled logistic regression models combining ICLR full-decision data with accepted-only NeurIPS and ICML records, because such models would mix

Table A1 Descriptive comparison of presentation features among accepted papers across venues. These data include accepted-only observations and should not be used to infer acceptance mechanisms.

Primary Area	N. Author	N. Formu	N. FormuMain	N. Ref	N. Tab	N. TabMain	N. Fig	N. FigMain	Len. Text	Len. TextMain
ICML2023	4.5 (2.4)	61.9 (65.4)	23.1 (14.7)	43.4 (16.4)	5.6 (4.5)	3.1 (2.5)	7.7 (6.0)	5.1 (3.8)	53024 (16548)	33519 (4218)
ICML2024	5.1 (3.2)	54.6 (64.6)	20.8 (14.8)	46.5 (18.2)	6.4 (5.0)	3.5 (2.7)	8.3 (6.4)	5.3 (3.7)	55137 (20596)	33646 (5555)
NeurIPS2023	4.8 (2.3)	43.6 (53.6)	17.2 (11.5)	50.5 (20.8)	5.4 (5.5)	2.6 (2.2)	7.9 (6.4)	4.9 (3.8)	47797 (18728)	29804 (4286)
NeurIPS2024	5.1 (2.5)	52.1 (67.9)	16.1 (11.5)	53.8 (21.5)	8.4 (5.5)	2.9 (2.2)	9.0 (6.6)	5.2 (3.9)	68204 (19028)	29332 (4654)
ICLR2024	5.3 (2.8)	36.8 (50.6)	13.2 (10.5)	44.1 (16.0)	7.3 (5.4)	3.2 (2.2)	9.5 (6.8)	5.4 (3.7)	52781 (19044)	30609 (5688)
ICLR2025	5.7 (3.7)	35.7 (50.4)	12.9 (10.4)	44.1 (16.7)	7.8 (5.4)	3.0 (2.2)	10.2 (6.9)	5.5 (3.7)	55995 (19826)	29985 (4592)

distinct sampling mechanisms. Researchers interested in cross-venue patterns should consult Appendix A, which reports descriptive statistics of accepted papers by venue.

Appendix C Robustness adjustment details

This appendix provides the detailed technical specification of the propensity-score-based methods used for robustness adjustment in Section 3.4.

C.1 Propensity score matching for binary variables

For binary focal variables (e.g., US affiliation indicator, student status), we employ propensity score matching. For a binary focal indicator $T_i \in \{0, 1\}$ and a vector of observed covariates X_i (including metadata and presentation features), the propensity score is the conditional probability of focal status:

$$e(X_i) = P(T_i = 1 \mid X_i),$$

estimated via logistic regression:

$$\log \left(\frac{e(X_i)}{1 - e(X_i)} \right) = \hat{\beta}_0 + \hat{\beta}^\top X_i,$$

where $\hat{\beta}$ represents the estimated coefficients. The estimated propensity score $\hat{e}(X_i)$ is used to match focal and comparison units with similar probabilities of focal status.

Matching relies on the assumption that, conditional on the propensity score, the distribution of observed covariates is similar between focal and comparison groups. We implement nearest-neighbor matching without replacement, using a caliper of 0.05

standard deviations of the logit propensity score. After matching, we assess covariate balance through standardized mean differences (SMD). The matched association is approximated by the average outcome difference between focal and matched comparison units. Standard errors are calculated using the Abadie-Imbens robust variance estimator.

C.2 Generalized Propensity Score (GPS) weighting for continuous variables

For continuous focal variables (e.g., number of formulas, text length), the GPS is the conditional density of the focal variable given covariates, estimated via Ridge regression:

$$r(t, X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t - \mu(X_i))^2}{2\sigma^2}\right).$$

The GPS method extends standard propensity-score matching to continuous focal variables. The key assumption is that, conditional on the GPS, the distribution of observed covariates is balanced across different levels of the focal variable. Positivity requires $r(t, X_i) > 0$ for all t and X_i , ensuring that all focal variable levels are represented in the data.

We estimate stabilized inverse-probability weights based on the ratio of marginal to conditional densities, applying trimming at the 5th and 95th percentiles to mitigate the influence of extreme weights. We then estimate GPS-weighted marginal associations using Weighted Least Squares (WLS) for ratings and weighted Generalized Linear Models (GLM, Binomial family) for acceptance decisions in the ICLR full-decision sample. This procedure is used to assess whether associations persist after balancing observed covariates.

C.3 Robustness and sensitivity checks

We conduct balance tests across GPS strata, ensuring that all standardized mean differences (SMD) fall below the 0.1 threshold after weighting. We also verify that the weighting procedure does not produce extreme inverse-probability weights that could destabilize the estimates, applying trimming at the 5th and 95th percentiles when necessary. We do not report Rosenbaum bounds because such bounds require effect-size benchmarks that are not well-defined for the present observational associations.

Appendix D Data example

```
{
  "paper_id": "<paper_id>",
  "primary_area": "<research area string>",
  "n_author": 4,
  "author_info_list": [
    {
      "name": "<Author A>",
      "institution": "<Institution A>",
      "country": "<Country code>",
      "position": "PhD student"
    }
  ]
}
```

```

    },
    {
      "name": "<Author B>",
      "institution": "<Institution B>",
      "country": "<Country code>",
      "position": "Assistant Professor"
    },
    {
      "name": "<Author C>",
      "institution": "<Institution C>",
      "country": "<Country code>",
      "position": "Associate Professor"
    },
    {
      "name": "<Author D>",
      "institution": "<Institution C>",
      "country": "<Country code>",
      "position": "PhD student"
    }
  ],
  "n_formula": 7,
  "n_formula_1": 7,
  "n_ref_uni": 29,
  "n_ref": 68,
  "n_ref_all": 79,
  "n_fig": 5,
  "n_tab": 1,
  "L_tab": 308,
  "n_element_tab": 21,
  "n_fig_1": 2,
  "n_tab_1": 1,
  "L_tab_1": 308,
  "n_element_tab_1": 21,
  "formula_len_all": 228,
  "formula_len_all_1": 228,
  "len_all": 96972,
  "len_all_1": 58049,
  "len_abs": 1506,
  "len_title": 124,
  "len_sents": 30553,
  "len_sents_1": 28827,
  "n_sents": 174,
  "n_sents_1": 164,
  "venue": "ICLR2024",
  "decision": "<decision label>",
  "L_title": 76,
  "L_abs": 1319,
  "title": "<Paper title text>",
  "abs": "<Abstract text describing the research problem,
    proposed method, key technical contributions, and main
    results.>",
  "keywords": [
    "<Keyword 1>",
    "<Keyword 2>",
    "<Keyword 3>",
    "<Keyword 4>"
  ],
  "rating_list": [6, 5, 3, 5, 5],
  "soundness_list": [3, 3, 2, 3, 3],
  "presentation_list": [2, 3, 2, 3, 3],

```

```

"contribution_list": [3, 2, 2, 3, 2],
"confidence_list": [2, 4, 4, 4, 4]
}

```

References

- Aczel, B., Barwich, A.-S., Diekman, A.B., Fishbach, A., Goldstone, R.L., Gomez, P., Gundersen, O.E., Hippel, P.T., Holcombe, A.O., Lewandowsky, S., *et al.*: The present and future of peer review: Ideas, interventions, and evidence. *Proceedings of the National Academy of Sciences* **122**(5), 2401232121 (2025)
- Ayal, S., Hochman, G., Zakay, D.: Two sides of the same coin: Information processing style and reverse biases. *Judgment and Decision Making* **6**(4), 295–305 (2011)
- Aldrich, J.H., Nelson, F.D.: *Linear Probability, Logit, and Probit Models. Quantitative Applications in the Social Sciences*, vol. 45. Sage, Beverly Hills, CA (1984)
- Ausloos, M., Nedi
vc, O., Dekanski, A.: Correlations between submission and acceptance of papers in peer review journals. *Scientometrics* **119**(1), 279–302 (2019) <https://doi.org/10.1007/s11192-019-03026-x>
- Bao, P., Hong, W., Li, X.: Predicting paper acceptance via interpretable decision sets. In: *Companion Proceedings of the Web Conference 2021*, pp. 461–467. ACM, Ljubljana, Slovenia (2021). <https://doi.org/10.1145/3442442.3451370>
- Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 3613 (2019). Association for Computational Linguistics
- Bornmann, L., Mutz, R., Daniel, H.-D.: A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLOS ONE* **5**(12), 14331 (2010)
- Bornmann, L.: Scientific peer review. *Annual Review of Information Science and Technology* **45**(1), 197–245 (2011)
- Connelly, B.L., Certo, S.T., Ireland, R.D., Reutzel, C.R.: Signaling theory: A review and assessment. *Journal of management* **37**(1), 39–67 (2011)
- Caliendo, M., Kopeinig, S.: Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* **22**(1), 31–72 (2008)
- Chaiken, S., Ledgerwood, A.: A theory of heuristic and systematic information processing. In: Van Lange, P.A.M., Kruglanski, A.W., Higgins, E.T. (eds.) *Handbook of Theories of Social Psychology* vol. 1, pp. 246–266. Sage Publications, London (2012)

- Drori, I., Te'eni, D.: Human-in-the-loop ai reviewing: Feasibility, opportunities, and risks. *Journal of the Association for Information Systems* **25**(1), 98–109 (2024)
- Ganapathy, K., Liu, E., Zarger, Z., Somepalli, G., Goldblum, M., Goldstein, T.: An Investigation into the Role of Author Demographics in ICLR Participation and Review. OpenReview. Retrieved from <https://openreview.net/forum?id=id=1DUwCRNAbA> (2022)
- Guo, F., Ma, C., Shi, Q., Zong, Q.: Succinct effect or informative effect: the relationship between title length and the number of citations. *Scientometrics* **116**(3), 1531–1539 (2018) <https://doi.org/10.1007/s11192-018-2805-8>
- Ghosal, T., Verma, R., Ekbal, A., Bhattacharyya, P.: Deepsentipeer: Harnessing sentiment in review texts to recommend peer review decisions. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1120–1130. Association for Computational Linguistics, Florence, Italy (2019). <https://doi.org/10.18653/v1/P19-1110>
- Hasan, S., Breunig, R.: Article length and citation outcomes. *Scientometrics* **126**(9), 7583–7608 (2021) <https://doi.org/10.1007/s11192-021-04083-x>
- Horbach, S.P., Halfman, W.: The ability of different peer review procedures to flag problematic publications. *Scientometrics* **118**(1), 339–373 (2019) <https://doi.org/10.1007/s11192-018-2969-2>
- Huang, S., Lu, W., Bu, Y., Huang, Y.: Revisiting the exploration-exploitation behavior of scholars' research topic selection: Evidence from a large-scale bibliographic database. *Information Processing & Management* **59**(6), 103110 (2022)
- Huang, Y., Lu, W., Liu, J., Cheng, Q., Bu, Y.: Towards transdisciplinary impact of scientific publications: A longitudinal, comprehensive, and large-scale analysis on microsoft academic graph. *Information Processing & Management* **59**(2), 102859 (2022)
- Huisman, J., Smits, J.: Duration and quality of the peer review process: The author's perspective. *Scientometrics* **113**(1), 633–650 (2017) <https://doi.org/10.1007/s11192-017-2310-5>
- Hudson, J.: An analysis of the titles of papers submitted to the UK REF in 2014: authors, disciplines, and stylistic details. *Scientometrics* **109**(2), 871–889 (2016) <https://doi.org/10.1007/s11192-016-2081-4>
- Imai, K., Van Dyk, D.A.: Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99**(467), 854–866 (2004)
- Joshi, D.J., Kulkarni, A., Pande, R., Kulkarni, I., Patil, S., Saini, N.: Conference

- paper acceptance prediction: Using machine learning. In: Machine Learning and Information Processing: Proceedings of ICMLIP 2020, pp. 143–152 (2021). Springer
- Jamali, H.R., Nikzad, M.: Article title type and its relation with the number of downloads and citations. *Scientometrics* **88**(2), 653–661 (2011) <https://doi.org/10.1007/s11192-011-0412-z>
- Jin, Y., Zhao, Q., Wang, Y., Chen, H., Zhu, K., Xiao, Y., Wang, J.: Agentreview: Exploring peer review dynamics with LLM agents. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 1208–1226. Association for Computational Linguistics, Miami, FL, USA (2024). <https://doi.org/10.18653/v1/2024.emnlp-main.70>
- Kang, D., Ammar, W., Dalvi, B., Van Zuylen, M., Kohlmeier, S., Hovy, E., Schwartz, R.: A dataset of peer reviews (peerread): Collection, insights and nlp applications. arXiv preprint arXiv:1804.09635 (2018)
- Kankanhalli, A.: Peer review in the age of generative ai. *Journal of the Association for Information Systems* **25**(1), 76–84 (2024)
- Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic Regression. Statistics for Biology and Health. Springer, New York, NY (2002). <https://doi.org/10.1007/b97379>
- Kang, K., Zhang, H., Li, Y., Luo, X., Wushour, S.: Automatic academic paper rating based on modularized hierarchical attention network. In: CCF International Conference on Natural Language Processing and Chinese Computing, pp. 669–681 (2022). Springer
- Lewison, G., Hartley, J.: What’s in a title? numbers of words and the presence of colons. *Scientometrics* **63**(2), 341–356 (2005) <https://doi.org/10.1007/s11192-005-0216-0>
- Lee, C.J., Moher, D., Sugimoto, C.R.: The state of the science of peer review: A survey. *Research Integrity and Peer Review* **8**(1), 5 (2023) <https://doi.org/10.1186/s41073-023-00129-4>
- Lopez, P.: Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27–October 2, 2009. Proceedings 13, pp. 473–474 (2009). Springer
- Manzoor, E.A., Shah, N.B.: Uncovering latent biases in text: Method and application to peer review. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 4767–4775. AAAI Press, Virtual Event (2021). <https://ojs.aaai.org/index.php/AAAI/article/view/16604>

- Nair, L.B., Gibbert, M.: What makes a ‘good’ title and (how) does it matter for citations? a review and general model of article title attributes in management science. *Scientometrics* **107**(3), 1331–1359 (2016) <https://doi.org/10.1007/s11192-016-1937-y>
- Petty, R.E., Cacioppo, J.T.: The elaboration likelihood model of persuasion. In: Berkowitz, L. (ed.) *Advances in Experimental Social Psychology* vol. 19, pp. 123–205. Academic Press, San Diego, CA (1986). [https://doi.org/10.1016/S0065-2601\(08\)60214-2](https://doi.org/10.1016/S0065-2601(08)60214-2)
- Ross-Hellauer, T.: What is open peer review? A systematic review. *F1000Research* **6**, 588 (2017) <https://doi.org/10.12688/f1000research.11369.2>
- Rimkus, L., Verbickas, J., Batista-Navarro, R.: Reviewer 2 must be stopped: Transformer-based approaches for predicting paper acceptance. In: *Annual International Conference on Information Management and Big Data*, pp. 249–260 (2023). Springer
- Squazzoni, F., Gandelli, C.: Saint Matthew strikes again: An agent-based model of peer review and the scientific community structure. *Journal of Informetrics* **6**(2), 265–275 (2012) <https://doi.org/10.1016/j.joi.2011.12.005>
- Shah, N.B.: Challenges, experiments, and computational solutions in peer review. *Communications of the ACM* **65**(6), 76–87 (2022)
- Staudinger, M., Kusa, W., Piroi, F., Hanbury, A.: An analysis of tasks and datasets in peer reviewing. In: *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pp. 257–268. Association for Computational Linguistics, Bangkok, Thailand (2024). <https://aclanthology.org/2024.sdp-1.24/>
- Skorikov, M., Momen, S.: Machine learning approach to predicting the acceptance of academic papers. In: *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pp. 113–117 (2020). IEEE
- Spence, M.: Job market signaling. In: *Uncertainty in Economics*, pp. 281–306. Elsevier, New York (1978). <https://doi.org/10.1016/B978-0-12-214850-7.50025-5>
- Stelmakh, I., Shah, N., Singh, A.: On testing for biases in peer review. *Advances in Neural Information Processing Systems* **32** (2019)
- Stelmakh, I., Shah, N.B., Singh, A., Daum
’e III, H.: A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 4785–4793. AAAI Press, Virtual Event (2021). <https://ojs.aaai.org/index.php/AAAI/article/view/16606>

- Silva, J.A., Bornemann-Cimenti, H., Tsigaris, P.: Optimizing peer review to minimize the risk of retracting COVID-19-related literature. *Medicine, Health Care and Philosophy* **24**, 21–26 (2021) <https://doi.org/10.1007/s11019-020-09990-z>
- Tranmer, M., Elliot, M.: Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)* **5**(5), 1–5 (2008)
- The OpenReview Team: OpenReview Python Client. GitHub repository. Version 1.27.3. Available at <https://github.com/iesl/openreview-py> (2023)
- Tran, D., Valtchanov, A., Ganapathy, K., Feng, R., Slud, E., Goldblum, M., Goldstein, T.: An open review of OpenReview: A critical analysis of the machine learning conference review process. arXiv preprint. arXiv:2010.05137 (2020). <https://arxiv.org/abs/2010.05137>
- Wang, G., Peng, Q., Zhang, Y., Zhang, M.: What have we learned from openreview? *World Wide Web* **26**(2), 683–708 (2023)
- Wang, K., Wan, X.: Sentiment analysis of peer review texts for scholarly papers. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 175–184 (2018)
- Wolfram, D., Wang, P., Hembree, A., Park, H.: Open peer review: promoting transparency in open science. *Scientometrics* **125**(2), 1033–1051 (2020) <https://doi.org/10.1007/s11192-020-03488-4>
- Xue, Z., He, G., Liu, J., Jiang, Z., Zhao, S., Lu, W.: Re-examining lexical and semantic attention: Dual-view graph convolutions enhanced bert for academic paper rating. *Information Processing & Management* **60**(2), 103216 (2023)
- Yan, E., Ding, Y.: Discovering author impact: A pagerank perspective. *Information processing & management* **47**(1), 125–134 (2011)
- Yao, R., Ye, Y., Zhang, J., Li, S., Wu, O.: Exploring developments of the ai field from the perspective of methods, datasets, and metrics. *Information Processing & Management* **60**(2), 103157 (2023)
- Zhang, Y., Yu, F.-Y., Schoenebeck, G., Kempe, D.: A system-level analysis of conference peer review. In: Proceedings of the 23rd ACM Conference on Economics and Computation, pp. 1041–1080. ACM, New York, NY, USA (2022). <https://doi.org/10.1145/3490486.3538301>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ICLR2025data.json](#)