Implicit Bias in Peer Review: Through the Lens of Language Abstraction

Xulang Zhang, Rui Mao, Erik Cambria
College of Computing and Data Science, Nanyang Technological University, Singapore
{xulang.zhang,rui.mao,cambria}@ntu.edu.sg

Abstract

Peer review is essential for the scholarly publishing process. However, its credibility is increasingly brought to questions. Bias is one of the aspects worthy of investigation. Existing research mostly focuses on predefined, explicit bias types, which are insufficient for analyzing the myriad of implicit biases in peer review. Thus, we proposed to study the bias in peer review through the lens of language abstraction, informed by the cognitive theories which suggest that frequency of abstraction in descriptions plays a latent yet important role in bias transmission. Hence, we trained a model to assess the abstraction level of text, and applied it to a review dataset to examine the connection between abstraction and the implicit biases in peer reviews. Results show that there are indeed observable quantitative differences in the abstraction use of reviews recommending to reject versus recommending to accept. Furthermore, reviews for the rejected papers tend to be more abstract than ones for the accepted papers, indicating possible transmission of implicit bias. To the best of our knowledge, our study is the first to study generalized Linguistic Intergroup Bias in the academic text domain.

Keywords: bias analysis, linguistic intergroup bias, linguistic category model

1. Introduction

Peer review remains a crucial and indispensable component of academic publishing, particularly for AI conferences. However, the rapid expansion and diversification of research communities, together with the widespread adoption of generative AI in paper writing and reviewing (Cambria et al., 2026), have intensified the burden on reviewers and raised growing concerns about the system's transparency (Wicherts, 2016), arbitrariness (Brezis and Birukou, 2020), quality (Rennie, 2016), and, most notably, bias (Roberts and Verhoef, 2016; Tomkins et al., 2017; Stelmakh et al., 2021). In this context, automating bias analysis in peer reviews becomes increasingly valuable (Nadeem et al., 2025a).

Existing NLP research on biased text mostly focuses on predefined, explicit bias, e.g., inappropriate language use related to gender, race, nationality, etc (Hartvigsen et al., 2022; Raza et al., 2024; Nadeem et al., 2025b). However, there are subtler biases of varying types and manifestations in peer reviews (Lee et al., 2013), which cannot be encapsulated by such an approach. Instead of tackling explicit or categorical biases, we analyze them with a more generalized approach, measured by abstraction to examine the usage of biased language. The core idea of the Linguistic Intergroup Bias (LIB) theory (Maass, 1999) is that, language abstraction is considered to be integrally linked with bias. This is rooted in the fact that the more abstract a statement is, the more it suggests temporal, cross-situational, and dispositional consistency, and the more difficult it is to be verified and to present refuting evidence (Semin and Fiedler, 1988).

The LIB is first proposed as the subconscious phenomenon that people tend to use abstract terms to describe in-group positive and out-group negative behavior, and concrete terms for the opposite (Maass et al., 1996), where the abstraction level is measured by the Linguistic Category Model (LCM) (Semin and Fiedler, 1991).

The theory is later extended to interpersonal scenarios, suggesting that differential usage of language abstraction not only indicates the bias of the speaker, but also contributes to bias formation of the audience (Maass, 1999; Wigboldus et al., 2000). Thus, in the context of peer review, it is reasonable to hypothesize that, because abstraction is generalizing and difficult to disprove, if a reviewer is intending to rejecting a paper, they would use comparatively more abstract terms in the review; and that abstract reviews might transmit the bias and have a sway on meta-reviewer's decision.

In this work, we aim to answer the following research questions:

(R1) Does the reviewer's preference for accepting/rejecting the paper manifest in the differential use of abstraction?

(R2) Is there a distinction between the abstraction levels of reviews for accepted and rejected papers?

(R3) Does the usage of abstractions in reviews affect the meta-reviewer's decision-making?

Figure 1 illustrates the analytical procedure of this work, including model training, data processing, and abstraction score computation. By analyzing a public dataset with ICLR peer reviews from 2017 to 2019, we find that, reviews recommending to reject a paper tend to be more abstract than ones recommending to accept (**R1**); reviews of rejected

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9569–9580
11–16 May 2026.

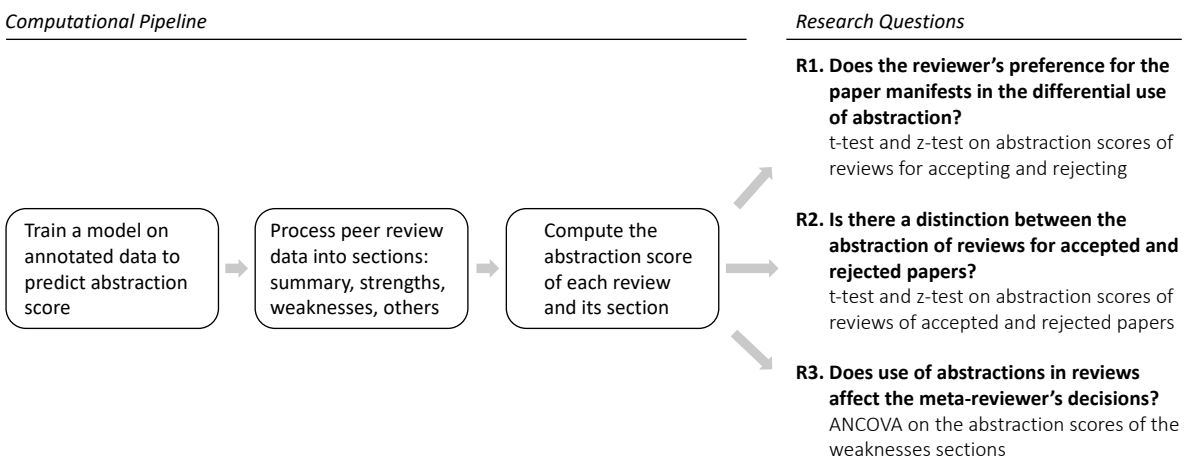

Figure 1: The analytical procedure for peer review bias analysis.

papers tend to be more abstract than accepted papers (**R2**); and abstraction use, particularly in the weaknesses aspect of a review, has an effect on whether a paper gets accepted (**R3**). To the best of our knowledge, this is the first work to apply LIB in the academic domain and the first to examine the peer review biases from the perspective of LIB.

2. Related Work

2.1. NLP-based Bias Analysis

Currently, NLP tools are increasingly employed for the peer review process. For instance, Ramachandran et al. (2017); Ghosal et al. (2022) assessed the quality of peer reviews by aspects such as relevance, expertise, tone, length, etc. Other research utilized Large Language Models (LLM) to assist or even generate peer review comments (Noothigattu et al., 2021; Dycke et al., 2022; Du et al., 2024; Kumar et al., 2024a), though not without criticism about their reliability and ethics (Yuan et al., 2022; Schintler et al., 2023; Ye et al., 2024).

Bias is one of the core challenges faced by the peer review process (Lee et al., 2013; Kuznetsov et al., 2024). Abundant studies have been conducted to investigate the various forms of bias in peer reviews, e.g., gender (Goldberg, 1968; Régner et al., 2019), prestige (Peters and Ceci, 1982; Tomkins et al., 2017), nationality (Ross et al., 2006), race and ethnicity (Ginther et al., 2011; Strauss et al., 2023), commensuration bias (Lee, 2015; Heesen, 2022), confirmation bias (Mahoney, 1977), resubmission bias (Stelmakh et al., 2021), etc. However, most studies do not focus on texts such as peer review comments. Limited text-based research exists; for example, Manzoor and Shah (2021) quantified bias in review text linked to visible identity indicators, restricted to predefined subgroups like gender and affiliation.

This is in accordance with the overall trends in the NLP research community, where studies on bias in text are primarily concerned with detecting language use that is harmful towards certain demographics (Hartvigsen et al., 2022; Pan et al., 2023; Raza et al., 2024; Kumar et al., 2024b; Yu et al., 2024; Ge et al., 2025) or biased predictions (Mao et al., 2023; Mei et al., 2024; Zhang et al., 2024; Yeo et al., 2025; Mao et al., 2025; Zhang et al., 2025). These works do not help in analyzing the varying and often implicit biases in domains that purport neutrality, such as peer review, leaving a research gap for a more generalized approach to bias analysis. As such, in this paper, we proposed an approach to examine whether the text in the review itself indicates patterns of bias.

2.2. LCM-based Bias Analysis

In psycholinguistics, language abstraction is considered to inherently express and transmit bias of the speaker, because of its perceived high level of stability, dispositional inference, and likelihood of repetition (Maass, 1999). Specifically, the abstraction levels are measured by the LCM, which defines four categories of terms. The most concrete terms are Descriptive Action Verbs (DAV), which describe specific, observable events such as *"A hit B"*. The second most concrete category is the Interpretive Action Verbs (IAV), which describe a specific event in a slightly more abstract way, e.g., *"A hurt B"*. The second most abstract category is the State Verbs (SV), which describe prolonged psychological states that go beyond a specific event, such as *"A hates B"*. Lastly, the most abstract terms are the adjectives (ADJ), which describe general, cross-situational dispositions such as *"A is aggressive"*. In bias analyses, DAVs and IAVs are considered as concrete terms, whereas SVs and ADJs are considered as abstract terms.

A variety of LCM-based studies have been conducted to analyze bias in different contexts, e.g., news articles regarding immigration (Geschke et al., 2010; Dragojevic et al., 2017), sports and political reports (Maass, 1999), political debates and interviews (Anolli et al., 2006), etc. Existing works have attempted to automate the LCM coding process, but they are mostly lexicon-based (Snefjella and Kuperman, 2015; Bhatia and Walasek, 2016; Reyt et al., 2016; Seih et al., 2017; Johnson-Grey et al., 2020). Such an approach lacks robustness and exhaustiveness, as the vocabularies are limited and cannot account for the varying word senses in different contexts. Thus, in this paper, we employed pretrained language models instead to predict the abstraction level of a given review.

3. Preliminary

In this section, we will introduce the LCM coding schema in detail.

The LCM (Semin and Fiedler, 1988; Coenen et al., 2006) is a framework for measuring levels of abstraction in descriptive text. It distinguishes four categories of abstraction terms, namely, Descriptive Action Verbs (DAV), Interpretive Action Verbs (IAV), State Verbs (SV), and Adjectives (ADJ). Note that the definition of abstractness varies across use cases (Coltheart, 1981; Crutch and Warrington, 2005; Brysbaert et al., 2014), and LCM defines a specific taxonomy of abstraction for the purpose of cognitive and interpersonal bias analysis.

The most concrete terms in LCM are DAVs, which provide a concrete and objective description of a specific behavioral event. Specifically, all actions that a DAV can describe invariantly share a physical feature. For instance, all the actions that *"talk"* can be applied to involve the physical feature of the mouth. Examples of DAV include: *walk, read, dance, ask, talk, discuss* and more.

IAVs are slightly more abstract as they describe a larger class of behaviors, although they maintain a clear reference to a specific behavior in a specific situation. They differ from DAV in that they do not share an invariant physical aspect, i.e., it is difficult to have a clear visualization of the behavior they refer to. Examples of IAV include: *do, have, make, study, exercise, follow, attend, review*, and more.

The most abstract verb category is SV, which describes lasting psychological states that generalize beyond specific situations and behaviors. These states can be either cognitive or affective, and cannot be objectively verified. Examples of SV include: *think, understand, know, love, hate*. Additionally, there is a special type of action verbs that are coded as SV. These verbs express an emotional consequence of a specific action, such as *angered, feared, surprised*, and more.

	Train	Dev	Test
No. of samples	1,939	718	719
No. of 0s	23	8	5

Table 1: The LCM-based dataset used for model training and testing. No. of 0s denotes the number of samples invalid for LCM coding.

Lastly, the most abstract category in LCM is ADJ, which includes adjectives and adverbs describing a general disposition that applies across situations, behaviors, and objects. Examples of ADJ include: *helpful, honest, reliable, aggressive, softly, lovingly*, and more.

Generally, in bias analyses, DAVs and IAVs are considered as concrete terms, whereas SVs and ADJs are considered as abstract terms. For the abstraction score calculation, they are assigned the weights 1, 2, 3, and 4, respectively. The abstraction score of a given sentence is computed as the weighted sum of their occurrences:

$$\text{abst} = \frac{\text{DAV} + 2 \cdot \text{IAV} + 3 \cdot \text{SV} + 4 \cdot \text{ADJ}}{\text{DAV} + \text{IAV} + \text{SV} + \text{ADJ}},$$

where DAV, IAV, SV, and ADJ denote the number of terms from each category in the sentence.

4. Materials

Data for Training. To train the model for abstraction scoring, we employed the data from Johnson-Grey et al. (2020), which consists of multiple sub-datasets containing text manually annotated with abstraction scores according to the LCM manual (Coenen et al., 2006). Specifically, we used the Study 1 dataset for training, and randomly divided the Feature dataset in half for validation and testing. Details are shown in Table 1.

Data for Analyses. To investigate implicit bias and lanugage abstraction in peer review, we employed the ICLR dataset (Chakraborty et al., 2020), which contains 8,151 reviews for 5,289 ICLR submissions from 2017, 2018, and 2019, including the ratings and decisions. For the purpose of this study, we utilized the deepseek-chat(v3) API[1] to sort the sentences in each review into four categories: summary of the paper, strengths, weaknesses, and other comments, as it is proven to demonstrate reliable text classification capabilities (Etaiwi and Alhijawi, 2025; Gao et al., 2026). Additionally, for the purpose of our analysis, we define a review as recommending acceptance when the rating is above 6, and recommending rejection when below 5. The rest are considered borderline cases. Details are shown in Table 2.

[1] platform.deepseek.com

Total	Recommend to accept	Recommend to reject	Borderline	Accepted	Rejected
8,151	2,291	2,341	3,519	3,147	5,004

Table 2: Details of the ICLR dataset, indicating the number of reviews of papers that are recommending to accept, recommending to reject, and borderline, as well as the number of reviews of papers that are accepted and rejected.

5. Methodology

In this section, we will introduce how we train an abstract scoring model using LCM-based annotated data. An illustration of the proposed model is shown in Figure 2.

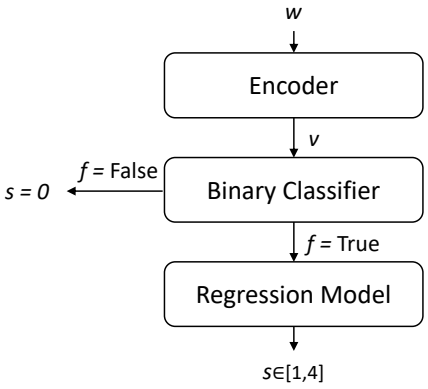

Figure 2: An illustration of the abstract scoring model. w denotes the input text, v denotes the embedding, f denotes the binary flag, and s denotes the predicted abstraction score.

As introduced in Section 3, according to the LCM manual (Coenen et al., 2006), the abstraction score is a continuous value between 1 and 4. In the case where the input is not qualified to be coded by LCM, the score is set to 0. Therefore, our model is a two-step pipeline where a binary classifier detects whether the input is valid for LCM coding, and a regression model outputs the abstraction score when valid.

Specifically, given an input w, a RoBERTa-base (Liu et al., 2019) encoder outputs its embedding:

$$v = Encoder(w).$$

The embedding is fed into two Feedforward Networks (FFN) to obtain the binary flag f:

$$f = \text{Sigmoid}(FFN_2^b(\text{ReLu}(FFN_1^b(v)))).$$

When f is true, the embedding v is fed through two FFNs and a sigmoid to obtain the continuous score $s \in [1, 4]$:

$$s = 1 + 3 \cdot \text{Sigmoid}(FFN_2^r(\text{ReLu}(FFN_1^r(v)))).$$

The binary classifier is optimized by Binary Cross-Entropy loss, and the regression model by Mean Squared Error loss.

The model is trained for 20 epochs using the AdamW optimizer (Loshchilov, 2017). Batch size is set to 16. Learning rate is set to 1e-5. The model achieved 80.12% R^2 score on the test set.

We then applied our model on the ICLR dataset to obtain the abstraction scores of the sections and the entirety of each review, shown in Section 6. We conducted human evaluation on 100 randomly selected samples, where two 2 annotators score the input sentences guided by the LCM manual (Coenen et al., 2006). Disagreements are resolved via discussions, with a 0.9 Cohen's Kappa. The model achieved 76.83% R^2.

6. Results

6.1. Analysis on Recommendations

To investigate how language abstraction correlates with reviewers' recommendations (**R1**), we conducted two-sample t-tests and z-tests on the abstraction scores of comments in different categories to examine whether there is a difference among the distributions of abstraction scores of reviews recommending acceptance and rejection. We adopted both statistical measures to ensure the robustness of the population variance assumptions. We discarded the Borderline reviews in this experiment due to their ambivalent preferences. The means and standard deviations are shown in Table 3.

Results show that the reviews recommending to accept ($M = 3.285, SD = 0.141$) are less abstract than the ones recommending to reject ($M = 3.298, SD = 0.142$), with $t = -3.126, z = -3.117$, both $p < 0.05$. Sentences in summary are more abstract when recommending to accept ($M = 3.313, SD = 0.181$) than to reject ($M = 3.304, SD = 0.194$), with $t = 1.632, p < 0.05$. Sentences in strengths are also more abstract when recommending to accept ($M = 3.405, SD = 0.289$) than to reject ($M = 3.371, SD = 0.287$), with $t = 4.017, z = 4.006$, both $p < 0.05$. In contrast, sentences in weaknesses are less abstract when recommending to accept ($M = 3.298, SD = 0.294$) than to reject ($M = 3.342, SD = 0.293$), with $t = -5.101, z = -5.086$, both $p < 0.05$. Other comments in the reviews recommending to accept a paper ($M = 3.263, SD = 0.178$) are also less abstract than recommending to reject ($M = 3.275, SD = 0.153$), with $t = -2.458, z = -2.452$, both $p < 0.05$.

	Recommend to accept		Recommend to reject		Average	
	M	SD	M	SD	M	SD
Summary	3.313	0.181	3.304	0.194	3.307	0.189
Strengths	3.405	0.289	3.371	0.287	3.384	0.288
Weaknesses	3.298	0.294	3.342	0.293	3.325	0.294
Others	3.263	0.178	3.275	0.153	3.270	0.163
Total	3.285	0.141	3.298	0.142	3.293	0.141

Table 3: Means (M) and Standard Deviations (SD) of the abstraction scores of different categories of sentences in the reviews that recommend to accept or reject.

	Accepted		Rejected	
	M	SD	M	SD
Summary	3.311	0.179	3.301	0.201
Strengths	3.395	0.268	3.376	0.289
Weaknesses	3.292	0.291	3.330	0.286
Others	3.268	0.117	3.278	0.126
Total	3.284	0.142	3.295	0.146

Table 4: Means (M) and Standard Deviations (SD) of the abstraction scores of different categories of sentences in the reviews of accepted and rejected papers.

6.2. Analysis on Decisions

To investigate whether the abstraction level of the reviews correlates with whether a paper is rejected or not (**R2**), we conducted t-tests and z-tests on the abstraction scores of review sentences of the accepted papers and the rejected papers (Table 4). Results indicate that reviews from the accepted papers ($M = 3.284, SD = 0.142$) are significantly less abstract than the rejected ones ($M = 3.295, SD = 0.146$), with $t = -3.368, z = -3.368$, both $p < 0.05$. The summaries in the reviews from the accepted papers ($M = 3.311, SD = 0.179$) are more abstract than the rejected ones ($M = 3.301, SD = 0.201$), with $t = 2.340, z = 2.341$, both $p < 0.05$. The strengths described in the reviews from the accepted papers ($M = 3.395, SD = 0.268$) are more abstract than the rejected ones ($M = 3.376, SD = 0.289$), with $t = 3.022, z = 3.023$, both $p < 0.05$.

On the other hand, the weaknesses described in the reviews from the accepted papers ($M = 3.292, SD = 0.291$) are less abstract than the rejected ones ($M = 3.330, SD = 0.286$), with $t = -5.778, z = -5.778$, both $p < 0.05$. The other comments from the accepted papers ($M = 3.268, SD = 0.117$) are significantly less abstract than the rejected ones ($M = 3.278, SD = 0.146$), with $t = -3.408, z = -3.408$, both $p < 0.05$. Overall, the trends of abstraction remain the same as the reviewers' recommendations. Interestingly, from Avg in Table 3, we found that the strengths in the reviews are described with significantly more abstraction than the weaknesses, regardless of the recommendations and the decisions.

Furthermore, from both Tables 3 and 4, we observe that while strength is the most abstract category, summary is more abstract than weaknesses in reviews recommending to accept ($t = 2.079, p < 0.05$) and in Accepted ($t = 3.119, p < 0.05$); whereas summary is less abstract than weaknesses in reviews recommending to reject ($t = -5.232, p < 0.05$) and in Rejected ($t = -5.868, p < 0.05$); other comments is always the most concrete category. Notably, the difference of abstraction scores of summary in recommendations ($\Delta = 0.009$) and decisions ($\Delta = 0.010$) are much smaller than the weaknesses ($\Delta = 0.044$) for recommendations and $\Delta = 0.038$ for decisions). These suggest that the distributions of abstraction scores of the weaknesses aspect for opposing recommendations/decisions are distinctly different than others. The different distributions for accepted and rejected papers are shown in Figure 3.

To further investigate whether the abstraction level of the reviews affects the meta-reviewers' decisions (**R3**), we perform ANCOVA (Analysis of Covariance) with binary logistic regression, where the abstraction score of weaknesses as the independent variable, the decision (Accepted, Rejected) as the dependent variable, and the recommendation (reviews recommending to accept, reject, and borderline) as a covariate. The overall model is statistically significant, $\chi^2(3) = 3195.6, p < 0.05$, with pseudo-$R^2 = 0.294$. To investigate **R3**, we found that the abstraction score of weaknesses had a small but significant negative effect on whether a paper is accepted or not ($\beta = -0.042$, $SE = 0.021$, $z = -2.026$, $p < 0.05$), after controlling for the recommendation. This suggests that, accounting for the impact of reviewers' scoring, there could be a higher chance of the paper being rejected when the weaknesses in the reviews are described abstractly, which aligns with the LIB's intuition that abstract statements are cognitively harder to refute with evidence.

6.3. Case Study

We examined a rejected paper and an accepted paper with polarizing ratings. The full reviews are shown in Table 6 and Table 7, respectively. The av-

Decision	Rating	Summary	Strengths	Weaknesses	Others	Total	Human
Reject	9	3.348	3.480	3.078	2.055	3.171	2.900
	4	3.394	3.519	3.422	3.047	3.326	3.238
	3	3.158	3.636	3.338	3.046	3.301	3.156
Accept	3	2.620	3.545	3.467	2.919	3.372	3.296
	6	2.999	3.543	3.348	2.907	3.075	2.973
	6	3.172	3.388	2.991	3.046	3.086	2.990

Table 5: The average abstraction scores of the reviews of a rejected paper and an accepted paper. Human indicates the abstraction scores of the entire review manually given by human annotators.

Review #1 (rating = 9):
this is a good paper. first of all, it presents a large-scale corpus for visual speech recognition. second, it demonstrates a visual speech recognition system based on open-vocabulary that gives the state-of-the-art recognition accuracy. the paper is very well written and all the technical details are clearly laid out. i, for one, would like to thank the authors for this meticulous work to the community. this is by far the largest dataset and the most impressive performance for vsr i have even seen in the asr/vsr community. i enjoyed reading this paper. i extend this review based on the replies. one of the arguments is that the work presented in this paper is a great success in engineering but it lacks technical novelty and therefore can not be accepted by the conference, which i think otherwise. first of all, the authors put together a very detailed and carefully designed technical pipeline for creating a very large visual speech recognition dataset, which is a valuable contribution to be community. (i assumed that the databset will become available to the community when reviewing the paper, which turned out not to be totally accurate. my apologies. i do hope the dataset will be made public. this is a major reason i gave a high score.) second, the authors have built systems that give the state-of-the-art performance on visual speech recognition. although the models and architectures are already out there, the impressive performance itself is an impact to the field. this is not simply achieved by piling in a large amount of data (although it does play a role). this is a system paper but its impact and its performance should at least get it in to the conference.
Review #2 (rating = 4):
the paper presents a non-trivial data processing pipeline, a large data set, and a system based on ctc and fsts for automatic lipreading from videos. the review of the previous work is comprehensive. the authors are also awared of the state of the art in speech recognition, a highly related task. the collection of the data set is definitely a contribution, but other than that, the technical novelty is scarce, since all of the techniques have been proposed either in lipreading from video or in speech recognition. the numbers in table 1 are impressive, but it is hard to tell where the improvement is coming from. it is worth running a few more experiments a) with the label set fixed while changing the network architecture b) with the network architecture fixed while changing the label set c) with the network and the label set fixed while changing dropout or group normalization. seq2seq is an odd child in this case, because you cannot really compare it to other settings. the result in table 2 is also impressive, but it would be nice to have the proposed system trained on lrs3-ted and compare against tm-seq2seq. it is generally a consensus that a large model paired with a large amount of data gives you improvement, and this type of improvement is not considered a contribution. it is then the authors' responsibility to have a comprehensive experiments showing that the improvement is not just due to having a larger model and more data. here are some minor details: p.6. note that there must be a blank between the 'e' characters to avoid collapsing ... –>this is actually not true, at least not in the original ctc formulation, where removing the duplicates and blanks have to be done in that order. to explain why modeling characters with ctc is problematic, ... –>this argument is not theoretically sound, so the question is does this happen in practice? the loss only measures at the independence level, but this doesn't prohibit the network to learn dependencies before the loss.
Review #3 (rating = 3):
the paper presents a large-scale lipreading system - no surprises there. this is good work and probably the strongest general purpose lip-reading system out there at this time, but i don't see both the work and the paper as a good fit for iclr. the authors take a large corpus of youtube videos (on which google has already trained direct acoustics-to-word speech recognizers, and which is manually transcribed), filter it, and extract regions that can be used for lipreading. they then describe a scalable preprocessing, and train a phone-based acoustic model using ctc. they seem to be using the (miao et al., 2015) and google wfst based decoding framework, and achieve a word error rate of ca 40%. that is impressive, but i don't see any novelty here, and the paper is full of contradictions, and leaves some important open questions: - the authors argue for "phonemes and ctc", and no speech person would disagree with them; in fact (miao et al., 2015) and many other papers show that the wers with a good phoneme based dictionary in english are lower than with a character based model. it's just easier if one does not need a dictionary. - why are the authors not using a viseme dictionary, or map their phoneme dictionary to a viseme dictionary. in visual space, their own "homonym" argument applies, too, and "mop" (or "mom") and "pop" should be mapped to the same "viseme" sequence - and the resulting uncertainty should be handled by the decoder, and not the classifier. - how did the authors generate the one million word phoneme vocabulary? even google used around 100,000k words in their whole-word experiments, if i remember correctly? what happens if the authros reduce the vocabulary? could you provide some error analysis or at least deletions/ insertions/ substitutiosn, and compare them against an audio system? - lipnet and the proposed architecture seem to be very similar - maybe you could provide some insight into which changes made the biggest difference? - is the data going to be available? - what is a "production-level speech decoder"? how come your model "is the first to combine a deep learning-based phoneme recognition model with production-grade word-level decoding techniques" if google does essentially the same ("in production")? - in section 1, you say that "by design, the trained model only performs well when videos are shot at specific angles when a subject is facing the camera, [...] it does not perform well in other contexts". in section 5, you demonstrate the "generalization power of our v2p approach"and find that it "is able to generalize well" - please clarify - "speech impaired patients" often have non-canonical articulation, the proposed system may not work well for them - it would be interesting to also know the absolute levels of insertions/ deletions/ substitutions for words and/ or phonemes, and for the audio only and visual systems, to be able to diagnose what the problems are. - finally, figure 10 is really hard to view - i'd be happy to be shown fewer faces, the main message is that the quality of the face detection is really good?

Table 6: Case study material of the rejected paper.

erage abstraction scores of each aspect of the two studied cases are shown in Table 5. The human-annotated abstraction scores are given by two annotators based on the LCM manual, and disagreements are resolved through discussions. From Table 5, we can see that for the rejected paper, the abstraction score of the review recommending to accept (Rating=9) is lower than the ones recommending to reject (Rating=4 and Rating=3).

The abstraction scores of weakness in the reviews recommending rejection are much higher than those recommending acceptance. These are consistent with the trends observed in Sections 6.1 and 6.2. The score of the strength section of the review recommending acceptance is the highest among all its aspects, but lower than the reviews recommending rejection, which could be attributed to individual differences in abstraction use.

Review #1 (rating = 3):
update: from the perspective of a \"broader ml\" audience, i cannot recommend acceptance of this paper. the paper does not provide even a clear and concrete problem statement due to which it is difficult for me to appreciate the results. this is the only paper out of all iclr2019 papers that i have reviewed / read which has such an issue. of course for the conference, the area chair / program chairs can choose how to weigh the acceptance decisions between interest to the broader ml audience and the audience in the area of the paper.

---

this paper addresses the problem that often features are obtained as a set, whereas certain orders of these features are known to allow for easier learning. with this motivation the goal of this paper is to learn a permutation of the features. this paper makes the following three main contributions: 1. the idea of using pairwise comparison costs instead of position-based costs 2. the methodological crux of how to go from the pairwise comparison costs to the permutation (that is, solving eqn. 2) using eqn. (1)) 3. an empirical evaluation i like the idea and the empirical evaluations are promising. however, i have a major concern about the second contribution on the method. there is a massive amount of literature on this very problem and a number of algorithms are proposed in the literature. this literature takes various forms including rank aggregation and most popularly the (weighted) minimum feedback arc set problem. the submitted paper is oblivious to this enormous literature both in the related work section as well as the empirical evaluations. i have listed below a few papers pertaining to various versions of the problem (this list is by no means exhaustive. with this issue, i cannot give a positive evaluation of this submitted paper since it is not clear whether the paper is just re-solving a solved problem. that said, i am happy to reconsider if the related work and the empirical evaluations are augmented with comparisons to the past literature on the methodological crux of the submitted paper (e.g., why off-the-shelf use of previously proposed algorithms may or may not suffice here.) unweighted feedback arc set: a fast and effective heuristic for the feedback arc set problem, eades et al. efficient computation of feedback arc set at web-scale, simpson et al. how to rank with few errors, kenyon-mathieu et al. aggregating inconsistent information: ranking and clustering, ailon et al. hardness results: the minimum feedback arc set problem is np-hard for tournaments, charbit et al. weighted feedback arc set: a branch-and-bound algorithm to solve the linear ordering problem for weighted tournaments, charon et al. exact and heuristic algorithms for the weighted feedback arc set problem: a special case of the skew\u2010symmetric quadratic assignment problem, flood approximating minimum feedback sets and multicuts in directed graphs, even et al. random inputs: noisy sorting without resampling, braverman et al. stochastically transitive models for pairwise comparisons: statistical and computational issues, shah et al. on estimation in tournaments and graphs under monotonicity constraints, chatterjee et al. survey (slightly dated): an updated survey on the linear ordering problem for weighted or unweighted tournaments, charon et al. convex relaxation of permutation matrices: on convex relaxation of graph isomorphism, afalo et al. facets of the linear ordering polytope, grotschel.

Review #2 (rating = 6):
the authors introduce a method to learn to permute sets end-to-end. they define the cost of a permutation as the sum of pairwise costs induced by the permutation, where the pairwise costs are learned. permutations are made differentiable by relaxing them to doubly stochastic matrices which are approximated with the sinkhorn operator. in the forward pass of the algorithm, a good permutation (ie one with low cost) is obtained with a few steps of gradient descent (the forward pass itself contains an optimization procedure). this permutation is then either used directly as the output of the algorithm or is used to permute the original inputs and feed the permuted sequence to another module (such as an rnn or a cnn). the method can easily be adapted to other structures such as lattices by considering row-wise and column-wise pairwise relations. the proposed method is benchmarked on 4 tasks: 1. sorting numbers, where they obtain very strong generalization results. 2. re-assembling image mosaics, on which they obtain encouraging results. 3. image classification through image mosaics. 4. visual question answering where the permuted inputs are fed to an lstm whose final latent state is fed back into the baseline model (a bilinear attention network). doing so improves over feeding the inputs to an lstm without learning the order.for which the output is the permutation itself and classification from image mosaics and visual question answering which require to learn an implicit permutation. the method is most similar to learning latent permutations with gumbel-sinkhorn networks (mena et al) but considers pairwise relations when producing the permutation. this can have important advantages (such as taking local relations into account, as shown by the strong sorting results) but also drawbacks (inability to differentiate inputs with similar content), but in any case this represents a good step towards exploring with different cost functions. the method can be quite unpractical (cubic time complexity in set cardinality, optimization in forward pass, having to preprocess the set into a sequence for another module can be resource expensive). experimental results on toy tasks (tasks 1, 2 and 3) are encouraging. the approach improves over a relatively strong baseline (task 4) although it isn't clear that it would still hold true when controlling for number of parameters and compute. i have a few comments about the presentation (for which i would be willing to change my score to a 6): - when possible, please use the numbers reported by mena et al and consider reporting error (instead of accuracy) as they do to ease comparison. the results that you report using their method are quite worse than what they report, so i think it would be fair to include both your reimplementation and the initial results in the table. - it would be interested to have some insights on what function f is learned (for the sorting task and re-assembling image mosaics for example).- clarity would be improved with figures representing which neural networks are used at what part of the process.
##
updated review: the authors have greatly improved presentation and have addressed concerns about the increase in parameters and computation time. i have changed my score to a 6.

Review #3 (rating = 6):
this paper proposed an interesting idea of learning representations of sets by permutation optimizations. through learning a permutation of the elements of a set, the proposed algorithm can learn a permutation-invariant representation of that set. to deal with the underlying difficult combinatorial optimization problem, the authors proposed to relax the optimization constraints and instead optimize over the set of doubly-stochastic matrices with reparameterization using the sinkhorn operator. the cost function of this optimization is related to a pairwise ordering cost, which compares the order for each pair of the elements. the idea of using pairwise comparison information to learn permutations is interesting. the total cost function utilizes the comparison information and optimization over this cost function can lead to a permutation-invariant representation of the set. the idea of using the sinkhorn operator to reparameterize the doubly-stochastic matrices makes the optimization objective differentiable. also, the experiment results compared with some baseline algorithms showed the success of the proposed methods in many different tasks. my major concern of the proposed method is on whether this method can be applied to large sets. since the algorithm compares all pairs of elements in the set, we need o(n^2) comparisons for a set of size n and hence the proposed method might be slow if n is large. is it possible to improve the efficiency for large sets? questions and suggestions: 1. since the authors wants to approximately solve the objective function in equation (2), it is better if we can see a proof showing why this optimization problem is difficult. 2. for the experiment in section 4.2, it seems that all methods (including the proposed methods and the baseline methods) are not performing well if the images are split to at least 4 * 4 equal-size tiles. i understand that currently the authors applied their method to the case of grid permutation by simply adding all cost functions of all rows and columns. is it possible to extend the proposed method to the grid case in another way so that the results under this setting is better? 3. it will be better if the authors can propose some more insights (probably with some theoretical analysis) when can the po-u method performs better and when can the po-la method performs better. 4. the authors mentioned that, the proposed method can get good permutations even for only t=4 steps. what if we continue running the algorithm? will the permutation converges stably? 5. the authors proposed to update the permutation matrix parameters in an alternative way (equation (7)) and mentioned that this update works significantly better in the experiments. it will be great if the authors can have a theoretical analysis on why this is true since p and \\tilde p can be quite different from each other for an arbitrary \\tilde p matrix. minor comment: i think there is a typo in equation (5). the entry \\tilde p_{pq} is related to not only the entry p_{pq}, but also the other entries of the matrix p. hence, i think equation (5) should be modified as a matrix multiplication.

Table 7: Case study material of the accepted paper.

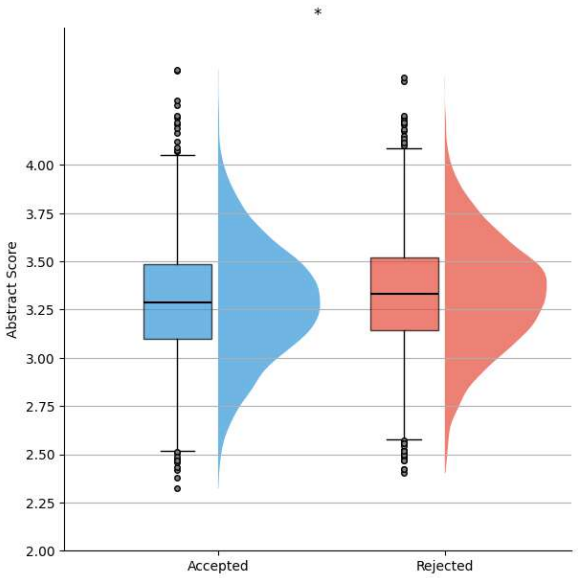

Figure 3: abstraction scores of weaknesses in the reviews of accepted and rejected papers (* denotes $p < 0.05$).

It can also be seen that for the accepted paper, the abstraction scores of the entirety and of the weaknesses of the review recommending to reject (Rating=3) are both higher than the ones of the other two borderline reviews (Rating=6), which aligns with the observation that reviews recommending to reject tend to be more abstract, especially for the weaknesses section (Section 6.1).

Comparing the two papers, it can be observed that the average abstraction score of the reviews for the rejected paper ($M = 3.266$) is higher than the average of the accepted paper ($M = 3.178$). The average abstraction score of the weaknesses aspects of the rejected paper ($M = 3.279$) is also higher than the average of the accepted paper ($M = 3.269$), reflecting the trends observed in Section 6.2. Hence, we can see that quantitatively, the reviews recommending rejection made a more abstract and convincing case that may lead to the paper being rejected.

Comparing the human-annotated scores, we can see that the trend of abstraction level observed from the model-predicted scores aligns with human evaluation. However, on the whole, the human-annotated scores tend to be lower than the model-predicted ones. This is likely because in the reviews, there are a lot of AI and machine learning jargons that contain adjectives, which by the definition of the LCM should not be counted as ADJ, e.g., *"neural"* in "neural network". As the training data is not from the AI research domain, the trained model might not be able to distinguish such adjectives, thus leading to overall higher abstraction level estimates.

We hypothesize that by adding a small amount of domain-specific data for fine-tuning, the model would yield more accurate predictions. Since all reviews contain various such jargons, we believe statistically it averages out and does not diminish the trends observed. Furthermore, as seen in Review #2 and #3 in Table 6, some reviews contain quotes from the paper, which are not statements made about the reviewed paper and thus would affect the abstraction score. Employing a filter to exclude such quotes could further improve the scoring accuracy.

7. Discussion

From the t-test and z-test results of Section 6.1, we find that with statistical significance, the reviews recommending to reject a paper are more abstract than the ones recommending to accept, which support our hypothesis in **R1**. A closer examination of the different aspects of reviews indicates that the strengths part of reviews recommending to accept is more abstract than the ones recommending to reject, whereas the opposite occurs for the weaknesses part in the reviews. These findings are consistent with the intuition that reviewers tend to use more abstract terms to describe the aspects that affirm their decisions, due to the generalizing and elusive nature of abstraction, which is also in line with the interpersonal theory of LIB.

Furthermore, the t-test and z-test also suggest that the summary parts are more abstract in the reviews recommending acceptance, while the other comments are more abstract in the reviews recommending rejection. The former is likely because when preferring to accept, reviewers tend to use more adjectives to qualify the methods proposed by the paper in their summary, e.g., *"the paper proposed a novel and effective model that ..."*, whereas concrete outlining of the methodology would be used when preferring to reject.

For the latter, we hypothesize that this phenomenon occurs because the other comment sections generally contain questions, suggestions, and justifications. Thus, reviewers tend to give more concrete, actionable advice for papers they recommend to accept, and raise questions with a higher abstraction level, i.e., harder to counter, for the ones they recommend to reject.

Additionally, from Table 3, it can be observed that the strengths obtain higher abstraction scores than all other aspects regardless of the recommendation preferences. It is likely because in peer review, reviewers are not inclined to detail what the authors did well, which aligns with previous findings in LIB studies for certain journalistic domains (Maass, 1999).

The t-test and z-test results of Section 6.2 show that the same trends of abstraction usage occur for different aspects of reviews in accepted and rejected papers, which affirms the research question **R2**. Interestingly, as mentioned in the section, while the strengths and the other comments remain the most and the least abstract aspect across all review types, for reviews recommending rejection and of rejected papers, the weaknesses are the second most abstract aspect; whereas for reviews recommending acceptance and of accepted papers, the weaknesses are the second least abstract aspect. This suggests that the abstraction score of weaknesses is the most prominent signifier of biased language use, as shown in Figure 3.

Moreover, with the ANCOVA results, we can further conclude that, although the recommendations (rating scores) of reviews affect the meta-reviewer's decision, statistical significance suggests that usage of abstraction in the weakness aspect of reviews may also play a potential role in the final decision (**R3**), indicating that implicit biases are transmitted through language abstraction, in accordance with LIB.

Since the use of language abstraction is generally not consciously controlled (Maass, 1999), it is worth exploring how this form of bias transmission may be mitigated for future works. For instance, the abstraction score of a review could be used as an indicator for reviewers and meta-reviewers to assess whether the review contains enough concrete arguments for its recommendation. Furthermore, automated review rewriting tools could be developed to decrease the abstraction level of reviews so as not to transmit implicit bias to the meta-reviewers.

8. Limitations

In this paper, we only focus on implicit linguistic interpersonal bias as defined by the LIB theories. There are many types of bias present in the peer review process, e.g., stereotypes and affiliation bias, as introduced in Section 2. These are all important aspects of bias research, but falls outside of the scope of this study.

Additionally, since the data we trained and tested our model on are not produced in the academic writing setting, the quality of the abstract scores outputted by the model is not quite leveled with the standard of human scoring, as indicated by results reported in Section 5 and Section 6.3.

Lastly, the findings of the paper are based on the experiments on the ICLR dataset. In future works, we will investigate their generalizability on peer review datasets in different scientific domains and distributions.

9. Ethics Considerations

The data and models involved in this paper are based on public corpora that do not contain private data nor offensive content. Our method is proposed for researching linguistic intergroup and interpersonal bias, and should not be used outside of research contexts. Our method should not be used for bias analysis on any individual's or group's private texts without their consent.

10. Bibliographical References

Luigi Anolli, Valentino Zurloni, and Giuseppe Riva. 2006. Linguistic intergroup bias in political communication. *The Journal of general psychology*, 133(3):237–255.

Sudeep Bhatia and Lukasz Walasek. 2016. Event construal and temporal distance in natural language. *Cognition*, 152:1–8.

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

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Erik Cambria, Rui Mao, Amir Hussain, Keith Oatley, and Geoffrey Hinton. 2026. Artificial intelligence as the fourth decentering revolution: From cosmic, biological, and psychological displacement to cognitive decentering. *Cognitive Computation*, 18(20):1–13.

Souvic Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2020. Aspect-based sentiment analysis of scientific reviews. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 207–216.

Linda HM Coenen, Liselotte Hedebouw, and Gün R Semin. 2006. Measuring language abstraction: The linguistic category model (lcm).

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33:497.

Sebastian J Crutch and Elizabeth K Warrington. 2005. Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3):615–627.

Marko Dragojevic, Alexander Sink, and Dana Mastro. 2017. Evidence of linguistic intergroup bias

in us print news coverage of immigration. *Journal of Language and Social Psychology*, 36(4):462–472.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. 2024. LLMs assist NLP researchers: Critique paper (meta-) reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Nlpeer: A unified resource for the computational study of peer review. *arXiv preprint arXiv:2211.06651*.

Wael Etaiwi and Bushra Alhijawi. 2025. Comparative evaluation of chatgpt and deepseek across key nlp tasks: Strengths, weaknesses, and domain-specific performance. *Array*, page 100478.

Tianchen Gao, Jiashun Jin, Zheng Tracy Ke, and Gabriel Moryoussef. 2026. A comparison of deepseek and other llms. *The American Statistician*, 80(1):164–176.

Mengshi Ge, Rui Mao, and Erik Cambria. 2025. Discovering the cognitive bias of toxic language through metaphorical concept mappings. *Cognitive Computation*, 17:65.

Daniel Geschke, Kai Sassenberg, Georg Ruhrmann, and Denise Sommer. 2010. Effects of linguistic abstractness in the mass media. *Journal of Media Psychology*.

Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *Plos one*, 17(1):e0259238.

Donna K Ginther, Walter T Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L Haak, and Raynard Kington. 2011. Race, ethnicity, and nih research awards. *Science*, 333(6045):1015–1019.

Philip Goldberg. 1968. Are women prejudiced against women? *Trans-action*, 5(5):28–30.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Remco Heesen. 2022. The necessity of commensuration bias in grant peer review. *Ergo*, 8:423–443.

Kate M Johnson-Grey, Reihane Boghrati, Cheryl J Wakslak, and Morteza Dehghani. 2020. Measuring abstract mind-sets through syntax: Automating the linguistic category model. *Social Psychological and Personality Science*, 11(2):217–225.

Asheesh Kumar, Tirthankar Ghosal, Saprativa Bhattacharjee, and Asif Ekbal. 2024a. Towards automated meta-review generation via an nlp/ml pipeline in different stages of the scholarly peer review process. *International Journal on Digital Libraries*, 25(3):493–504.

Shanu Kumar, Gauri Kholkar, Saish Mendke, Anubhav Sadana, Parag Agrawal, and Sandipan Dandapat. 2024b. Socio-culturally aware evaluation framework for llm-based content moderation. *arXiv preprint arXiv:2412.13578*.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. 2024. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*.

Carole J Lee. 2015. Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283.

Carole J Lee, Cassidy R Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for information Science and Technology*, 64(1):2–17.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Anne Maass. 1999. Linguistic intergroup bias: Stereotype perpetuation through language. In *Advances in experimental social psychology*, volume 31, pages 79–121. Elsevier.

Anne Maass, Roberta Ceccarelli, and Samantha Rudin. 1996. Linguistic intergroup bias: Evidence for in-group-protective motivation. *Journal of Personality and Social Psychology*, 71(3):512.

Michael J Mahoney. 1977. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1:161–175.

Emaad Manzoor and Nihar B Shah. 2021. Uncovering latent biases in text: Method and application to peer review. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4767–4775.

Rui Mao, Guanyi Chen, Xiao Li, Mengshi Ge, and Erik Cambria. 2025. A comparative analysis of metaphorical cognition in ChatGPT and human minds. *Cognitive Computation*, 17(35):1–12.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14:1743–1753.

Xin Mei, Rui Mao, Xiaoyan Cai, Libin Yang, and Erik Cambria. 2024. Medical report generation via multimodal spatio-temporal fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 4699–4708.

Mohammad Nadeem, Shahab Saquib Sohail, Erik Cambria, and Shagufta Afreen. 2025a. South Asian biases in language and vision models. *Nature Machine Intelligence*, 7:1775–1777.

Mohammad Nadeem, Shahab Saquib Sohail, Erik Cambria, Björn W Schuller, and Amir Hussain. 2025b. Gender bias in text-to-video generation models: A case study of Sora. *IEEE Intelligent Systems*, 40(3):10–15.

Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. 2021. Loss functions, axioms, and peer review. *Journal of Artificial Intelligence Research*, 70:1481–1515.

Zihao Pan, Kai Peng, Shuai Ling, and Haipeng Zhang. 2023. For the underrepresented in gender bias research: Chinese name gender prediction with heterogeneous graph attention network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14436–14443.

DP Peters and SJ Ceci. 1982. Peer-review practices of psychological journals: The fate of published articles, submitted again. reprinted from "the behavioral and brain sciences. In *Peer Commentary on Pcer Review: A Case Study in Scientific Quality Control*. Cambridge University Press Cambridge, UK.

Lakshmi Ramachandran, Edward F Gehringer, and Ravi K Yadav. 2017. Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education*, 27:534–581.

Shaina Raza, Muskan Garg, Deepak John Reji, Syed Raza Bashir, and Chen Ding. 2024. Nbias: A natural language processing framework for bias identification in text. *Expert Systems with Applications*, 237:121542.

Isabelle Régner, Catherine Thinus-Blanc, Agnès Netter, Toni Schmader, and Pascal Huguet. 2019. Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature human behaviour*, 3(11):1171–1179.

Drummond Rennie. 2016. Let's make peer review scientific. *Nature*, 535(7610):31–33.

Jean-Nicolas Reyt, Batia M Wiesenfeld, and Yaacov Trope. 2016. Big picture is better: The social implications of construal level for advice taking. *Organizational Behavior and Human Decision Processes*, 135:22–31.

Seán G Roberts and Tessa Verhoef. 2016. Double-blind reviewing at evolang 11 reveals gender bias. *Journal of Language Evolution*, 1(2):163–167.

Joseph S Ross, Cary P Gross, Mayur M Desai, Yuling Hong, Augustus O Grant, Stephen R Daniels, Vladimir C Hachinski, Raymond J Gibbons, Timothy J Gardner, and Harlan M Krumholz. 2006. Effect of blinded peer review on abstract acceptance. *Jama*, 295(14):1675–1680.

Laurie A Schintler, Connie L McNeely, and James Witte. 2023. A critical examination of the ethics of ai-mediated peer review. *arXiv preprint arXiv:2309.12356*.

Yi-Tai Seih, Susanne Beier, and James W Pennebaker. 2017. Development and examination of the linguistic category model in a computerized text analysis method. *Journal of Language and Social Psychology*, 36(3):343–355.

Gün R Semin and Klaus Fiedler. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*, 54(4):558.

Gün R Semin and Klaus Fiedler. 1991. The linguistic category model, its bases, applications and range. *European review of social psychology*, 2(1):1–30.

Bryor Snefjella and Victor Kuperman. 2015. Concreteness and psychological distance in natural language use. *Psychological science*, 26:1449.

Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. 2021. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proceedings of the ACM*

on Human-Computer Interaction, 5(CSCW1):1–17.

Dana Strauss, Sophia Gran-Ruaz, Muna Osman, Monnica T Williams, and Sonya C Faber. 2023. Racism and censorship in the editorial and peer review process. *Frontiers in Psychology*, 14:1120938.

Andrew Tomkins, Min Zhang, and William D Heavlin. 2017. Single versus double blind reviewing at wsdm 2017. *arXiv preprint arXiv:1702.00502*.

Jelte M Wicherts. 2016. Peer review quality and transparency of the peer-review process in open access and subscription journals. *PloS one*, 11(1):e0147913.

Daniel HJ Wigboldus, Gün R Semin, and Russell Spears. 2000. How do we communicate stereotypes? linguistic bases and inferential consequences. *Journal of Personality and Social Psychology*, 78(1):5.

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. 2024. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint arXiv:2412.01708*.

Wei Jie Yeo, Rui Mao, Moloud Abdar, Erik Cambria, and Ranjan Satapathy. 2025. Debiasing clip: Interpreting and correcting bias in attention heads. *arXiv preprint arXiv:2505.17425*.

Zehui Yu, Indira Sen, Dennis Assenmacher, Mattia Samory, Leon Fröhling, Christina Dahn, Debora Nozza, and Claudia Wagner. 2024. The unseen targets of hate: A systematic review of hateful communication datasets. *Social Science Computer Review*, page 08944393241258771.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Xulang Zhang, Rui Mao, and Erik Cambria. 2024. Multilingual emotion recognition: Discovering the variations of lexical semantics between languages. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.

Xulang Zhang, Rui Mao, and Erik Cambria. 2025. A systematic analysis of biases in large language models. *arXiv preprint arXiv:2512.15792*.