Causal Matching with Text Embeddings: A Case Study in Estimating the Causal Effects of Peer Review Policies

Raymond Z. Zhang[1] **Neha Nayak Kennard**[2] **Daniel Scott Smith**[1]
Daniel A. McFarland[1] **Andrew McCallum**[2] **Katherine A. Keith**[3]
[1]Stanford Graduate School of Education
[2]University of Massachusetts Amherst
[3]Williams College
ray.zhang@alumni.stanford.edu {kennard, mccallum}@cs.umass.edu
{danielscottsmith,mcfarland}@stanford.edu kak5@williams.edu

Abstract

A promising approach to estimate the causal effects of peer review policies is to analyze data from publication venues that shift policies from single-blind to double-blind from one year to the next. However, in these settings the content of the manuscript is a confounding variable—each year has a different distribution of scientific content which may naturally affect the distribution of reviewer scores. To address this textual confounding, we extend variable ratio nearest neighbor matching to incorporate text embeddings. We compare this matching method to a widely-used causal method of stratified propensity score matching and a baseline of randomly selected matches. For our case study of the ICLR conference shifting from single- to double-blind review from 2017 to 2018, we find human judges prefer manuscript matches from our method in 70% of cases. While the unadjusted estimate of the average causal effect of reviewers' scores is -0.25, our method shifts the estimate to -0.17, a slightly smaller difference between the outcomes of single- and double-blind policies. We hope this case study enables exploration of additional text-based causal estimation methods and domains in the future.

1 Introduction

For over two hundred years, peer review has been the key means of evaluating scholarly work and establishing scientific legitimacy (Birukou et al., 2011). Although many claim double-blind peer review reduces evaluation biases due to known author identities (Tvina et al., 2019; Sun et al., 2022; Kern-Goldberger et al., 2022), others claim there is little statistical evidence for a preference over single-blind (Haffar et al., 2019).

In this work, we argue that studying the impact of peer review anonymization policies is inherently a *causal* question. If we intervene and assign a manuscript to double-blind review, what is resulting effect on the manuscript's review score compared to what the score would have been under single-blind review? The gold standard for these types of causal estimation questions is randomized controlled trials (RCTs) which produce unbiased effect estimates (Holland, 1986; Pearl, 2009). However, in the case of peer review, an RCT is unethical because applying different review policies to different manuscripts could potentially harm the dissemination of scientific findings and researchers' careers.

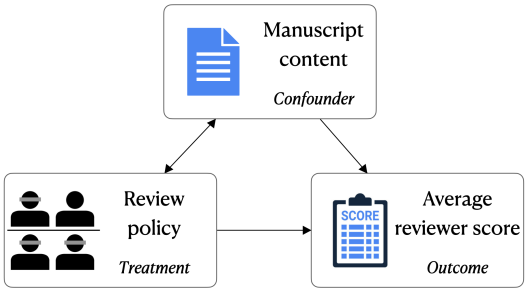

Figure 1: Causal diagram for our case study that estimates the causal effects of single- versus double-blind peer review policies.

In the absence of an RCT, one can use observational (non-randomized) data from publication venues before and after a policy change to estimate causal effects. However, a major obstacle to unbiased effect estimation for observational data is accounting for confounding variables that affect both treatment and outcome. In the case of peer review, we represent our domain assumptions via the causal diagram in Figure 1. The content of a manuscript affects peer review scores—popular scientific content might naturally have higher reviewer scores—and is correlated with review policy—the distribution of content in submitted manuscripts might be different in years with single- versus double-blind review.[1] Many methods have been proposed

[1]The diagram in Figure 1 is an Acyclic Directed Mixed Graph (ADMG) (Richardson, 2003) which contains both directed edges—denoting direct causal dependence—and bidirected edges. Manuscript content and review policy have a

Findings of the Association for Computational Linguistics: ACL 2023, pages 1284–1297
July 9-14, 2023

to statistically adjust for confounding variables in general (Rosenbaum and Rubin, 1983; Pearl, 2009; Morgan and Winship, 2015). We follow the framework presented by Keith et al. (2020) who review settings for which text data is a proxy for confounding variables.

From this prior work, we distill **three important criteria** for choosing text-based confounding adjustment method for case studies like ours: the method should (1) allow for empirical checks of causal overlap, (2) incorporate modern text representations, and (3) enable human validation of intermediate steps. Causal *overlap*—a necessary condition for an estimate to be causal—requires any unit to have a non-zero probability of assignment to each treatment condition for all possible values of confounders (Morgan and Winship, 2015). If causal overlap is not satisfied, one has to either abandon the project or shift the target causal estimand[2]. Second, text data contains many layers of linguistic granularity and there are challenges to operationalizing a variable like "manuscript content". General-purpose language representations have greatly improved performance of predictive natural language processing (NLP) tasks, e.g. Peters et al. (2018); Devlin et al. (2019); Cohan et al. (2020), and we hypothesize we could use them to help find semantically similar treated and untreated documents. Finally, human validation is important because unlike prediction settings, causal settings have no ground-truth (Holland, 1986). In our case study, we do not have access to counterfactual outcomes for the same manuscript under both treatment settings. Thus, we are essentially combining the "black box" of causal estimation with another "black box" of NLP techniques, so it is crucial that we are able to evaluate intermediate steps of the causal estimation pipeline in order to lend validity to our results.

In summary, we contribute the following:

- We combine causal matching approaches with the NLP embedding literature and implement variable ratio nearest neighbor matching with replacement and a caliper on cosine distance of document embeddings, *Variable Ratio Matching with Embeddings (VRM-E)*[3]. We demonstrate that this method satisfies our three criteria above.
- We apply VRM-E to a case study of peer review data, consisting of ratings from 1400 manuscripts from the International Conference on Learning Representations (ICLR) in 2017 (single-blind peer review) and 2018 (double-blind).
- For our case study, we find human domain-experts prefer matches between treated and untreated manuscripts from VRM-E over 70% of the time compared to a baseline of stratified propensity score matching (Rosenbaum and Rubin, 1983) and randomly selected matches.
- While the baseline unadjusted estimate of the average treatment effect on the control (ATC) of aggregated reviewers' scores (on a 10-point scale) is -0.25 with 95% confidence interval of [-0.39, -0.11], VRM-E shifts the ATC to -0.17 [-0.29, -0.05], a slightly smaller difference between the outcomes of single- and double-blind policies.

2 Related Work

Methods for text-based confounding adjustment. We describe gaps in existing work based on our three criteria of overlap (O), incorporating modern NLP representations (R), and human validation (V). The text adjustment method proposed by Roberts et al. (2020) uses human judgements experiments for validation (V+) but relies on topic modeling (R-). While Veitch et al. (2020) make use of state-of-the-art NLP in the form of BERT (Devlin et al., 2019) to jointly estimate treatment and counterfactual outcomes (R+), one cannot validate intermediate representations (I-) or empirically check for overlap (O-). Wood-Doughty et al. (2018) use classifiers to adjust for textual confounding variables; however, many settings, including ours, do not have gold-standard labels of low-dimensional confounders necessary for these classifiers. Mozer et al. (2020) propose a framework for human judgement of text matches (V+), but their empirical results are domain-dependent and do not generalize to our case study. Many other applications of text in causal inference use stratified propensity score matching (SPSM) (Rosenbaum and Rubin, 1983), e.g. De Choudhury et al. (2016); De Choudhury

bi-directed edge because manuscript content does not *directly cause* the review policy but unmeasured common causes affect both variables. Conditioning on manuscript content still blocks the "backdoor path" between treatment and outcome, so we call it a confounding variable in the remainder of this work.

[2]Typically, one has to shift to only the subset of the population with "common support" in the sample; see Morgan and Winship (2015) Section 4.6.1.

[3]Code for VRM-E and other experiments from this paper can be found at `https://github.com/raymondEDS/VRM-E/`

and Kiciman (2017); Olteanu et al. (2017); Kiciman et al. (2018); Saha et al. (2019). Because SPSM is widely-used and satisfies our three criteria, we empirically compare to this method in Section 3.4.

Peer review studies. Some argue double-blind review reduces bias associated with reputation, race, gender, and institution (Tvina et al., 2019). Experiments show that single-blind reviewers bid (Tomkins et al., 2017) and recommend acceptance (Okike et al., 2016) at higher rates for famous authors from top institutions. In a non-causal study using the same ICLR dataset, Sun et al. (2022) show changing from single- to double-blind review results in decreased scores for prestigious authors. Manzoor and Shah (2021) also use the ICLR dataset but focus on text as a causal outcome rather than as a confounding variable.

3 Methods and empirical pipeline

3.1 Case study data

We use titles, abstracts, and review ratings for ICLR 2017 and 2018 submissions, scraped from OpenReview by Zhang et al. (2022). In both years, each submission was rated by multiple reviewers on a 10-point scale; we use the mean rating as the causal outcome. ICLR 2017 used single-blind reviewing, and had 490 submissions; ICLR 2018 used double-blind reviewing and received 910 submissions.

3.2 Set-up for causal estimation

To estimate causal effects, ideally we would have counterfactual outcomes for each unit i—in our case study, a unit is a single manuscript—for both treatment settings, $T = 0$ and $T = 1$. Using the potential outcomes framework (Rubin, 1974, 2005), we denote these counterfactual outcomes as $Y_i(T_i = 0)$ and $Y_i(T_i = 1)$. The average treatment effect (ATE) for population of n units is

$$\tau = \frac{1}{n} \sum_i \Big(Y_i(T_i = 1) - Y_i(T_i = 0) \Big) \quad (1)$$

However, the *fundamental problem of causal inference* is that we do not have access to both counterfactual outcomes for a single unit (Holland, 1986). Instead, a naive approach estimates the ATE as a difference in means between the treated and untreated groups

$$\hat{\tau}_{\text{naive}} = \frac{1}{n_1} \sum_{i:T_i=1} Y_i - \frac{1}{n_0} \sum_{i:T_i=0} Y_i \quad (2)$$

with n_1 and n_0 being the number of units in the treated and untreated groups respectively. This naive estimate can be biased in the presence of confounding variables, C. To address this confounding, one can use the backdoor adjustment formula (Pearl, 2009) to statistically adjust for C

$$\hat{\tau}_{\text{BDA}} = \sum_c \Big(E[Y|T = 1, C = c] - E[Y|T = 0, C = c] \Big) P(C = c) \quad (3)$$

Eq. 3 is an unbiased estimate of the ATE under certain necessary causal identification assumptions such as no unmeasured confounding and *causal overlap*: $\forall c, 0 < P(T = 1|C = c) < 1$. Intuitively, if causal overlap is satisfied, then the terms in Eq. 3 can be estimated from data because then there are at least one treated and untreated unit for each c. However, D'Amour et al. (2021) show that overlap becomes increasingly difficult to satisfy as the dimensionality of C grows. Thus, for text-based confounding settings, practitioners face tradeoffs between the linguistic granularity for which they operationalize C and satisfying causal overlap.[4]

3.3 VRM-E

To satisfy the three criteria described in Section 1, we combine previous work from the NLP representation learning literature (Le and Mikolov, 2014; Wu et al., 2018; Zamani et al., 2018) with the causal literature on variable ratio matching (Ming and Rosenbaum, 2001; Stuart, 2010). As Stuart (2010) notes, causal matching has the advantage that one can empirically check regions for overlap, whereas alternatives, like regression, would rely on extrapolation for those same regions. Our method, *Variable Ratio Matching with Embeddings (VRM-E)*, operationalizes C as many clusters of semantically similar documents where each cluster has at least one treated and non-treated manuscript, thus explicitly satisfying overlap. Note, like all causal estimation approaches, the validity of our

[4]To illustrate this tradeoff, consider the following hypothetical scenario: the size of the vocabulary is 10 and C is operationalized as vector of word indicators for each document. This gives $2^{10} = 1024$ possibilities for c and there must be least one $T = 0$ and $T = 1$ document for each c to satisfy overlap (2048 total documents). This minimum number of documents needed is more than the total documents we have in this case study and grows exponentially with the size of the vocabulary. Thus, practitioners must choose a different operationalization of text or abandon the project.

method is contingent on the assumptions we stated previously; see Limitations section for more discussion. We subsequently describe the five steps of VRM-E and provide suggestions to navigate bias-variance trade-offs and trade-offs between causal overlap and the granularity of textual semantic similarity between treated and untreated groups.

Step 1: Obtain embeddings for each unit of text.

Step 2: Set the "anchor" group as the treated or untreated group with the fewest number of units. This will ensure each unit in the smaller group is matched with at least one unit in larger group, satisfying causal overlap.[5]

Step 3: Run agglomerative clustering on the cosine distance[6] of the embeddings for all anchor units, with a maximum distance threshold a. This step ensures extremely similar units in the anchor group are matched with the same non-anchor units and reduces the runtime of Step 4.

Step 4: For each cluster centroid from Step 3, find the k-nearest neighbors in the non-anchor group by embedding cosine distance, limiting to a maximum distance of b. Our b variable is analogous to a *caliper* in other causal matching literature; see Rosenbaum and Rubin (1985a,b); Stuart (2010).

Step 5: Use a matching estimator (see Appendix A) to estimate the causal effect.

Setting hyperparameters. An advantage of our method is that it is agnostic to the choice of text embedding and domain experts should choose the embedding that is best-suited for their domain. Additionally, our method only has three free hyperparameters: k, a, and b. As Gelman and Loken (2013) discuss, limiting the number of free parameters can help mitigate "garden of forking paths" issues in data analysis. We make several recommendations on navigating the tradeoffs for these hyperparameters. Choices of k correspond to bias-variance trade-offs; a higher k may result in increased bias while decreasing variance (Stuart, 2010). The a threshold is meant to be arbitrarily small to create tight clusters in the anchor group. Setting b to a small value increases the chance of textual semantic similarity between units, but could result in a violation of causal overlap for anchor units that have no matches; thus, we recommend selecting the minimum value of b that still satisfies overlap.

Case study hyperparameters. For our case study, we use SPECTER (Cohan et al., 2020), a pre-trained language model which generates embeddings for scientific manuscripts using their titles and abstracts, and outperforms alternative models on benchmark scientific tasks. After qualitative inspection of initial results, we set $k = 10$ and $a = 0.1$; see Appendix D.3 for robustness to these choices. We set $b = 0.23$ since this is the lowest value such that all manuscripts in our anchor group ($T = 0$; 2017 manuscripts) have at least one match in the non-anchor group; see Appendix B.2. In Step 4, we allow for matching with replacement and further investigate this decision in Appendix C.1.1.

3.4 SPSM

We compare VRM-E to stratified propensity score matching (SPSM). To do so, we train a logistic regression model for the propensity scores, $P(T = 1|X)$. For our case study, we operationalize X as the same SPECTER embeddings used in VRM-E to compare the two methods fairly. We use cross-fitting (Hansen, 2000; Newey and Robins, 2018) with cross validation within the training folds to ensure the models are not overfitting. Using the trained models, we infer propensity scores for each unit in the corresponding inference folds and then we stratify the scores into the standard five buckets (Neuhäuser et al., 2018). See Appendix D.1 for more details. Empirically, we find SPSM is limited when incorporating text embeddings. First, although SPSM satisfies overlap, approximately 95% of the data is distributed in stratum 3 (between scores 0.4 and 0.6); see Figure 4 and Table 7. Additionally, the model only has 62% accuracy on the training folds. This shows propensity score modeling's limitation—it collapses rich text data into a single score whereas VRM-E maintains more fine-grained matches.

4 Results for Peer Review Case Study

Because there is no ground-truth in causal inference, we first manually evaluate matches to assess validity, and then estimate the causal effects. In our Limitations section, we discuss potential threats to validity.

[5]While this satisfies overlap, it can change the estimand to the average treatment effect on the control (ATC) or average treatment effect on the treated (ATT). This is still preferred to shifting the estimate to only the subsample of the population with "common support" (Morgan and Winship, 2015).

[6]This is equivalent to one minus the cosine similarity. This metric has become standard for embedding similarity in NLP (Chandrasekaran and Mago, 2021; Mohammad and Hirst, 2012)

Method	Raw prefs.	Pref. %
VRM-E	212	70%
Random matches	38	13%
SPSM	53	17%

Table 1: Domain-expert human preferences (pref.) for the treated-untreated manuscripts matches from the three methods. VRM-E is statistically different from both random matches and SPSM with $p < 0.01$ (two-sided T-test; see Appendix D.2).

Method	ATC	95% CI
VRM-E	-0.17	[-0.29, -0.05]
Naive (unadjusted)	-0.25	[-0.39, -0.11]
SPSM	-0.26	[-0.38, -0.14]

Table 2: For ICLR 2017 and 2018 data, estimates of the average treatment effect on the control (ATC) and 95% confidence interval (CI). ATC for VRM-E is statistically different from other methods with $p < 0.01$ (paired bootstrap; see Appendix D.2).

4.1 Human judgements on matches

Causal matching aims to find units that "look similar" but receive different treatments, allowing researchers to approximate counterfactuals. Both VRM-E and SPSM allow for human evaluation of matches, satisfying one of our three criteria in Section 1.[7] To empirically analyze the differences between these matching methods, three authors compared 100 randomly sampled manuscript titles from ICLR 2017 to their matched titles from ICLR 2018.[8] For the same ICLR 2017 manuscript, three 2018 manuscript matches were judged: one randomly sampled match, one match from VRM-E, and one match from SPSM. These matches were permuted (and method names masked) and the judges were instructed to select the most similar of the three matches; ties were allowed. See Appendix B.1 for additional details. To obtain each method's preference percentage, we add together the preferences across all three judges and then divided by the total number of preferences. Table 1 shows judges prefer our method 70% of the time. The agreement rate between judges is 0.56 for Fleiss' Kappa (Davies and Fleiss, 1982), a low value but not unreasonable given the difficultly of the task. The high preference for VRM-E lends validity to our final causal estimates from this approach.

4.2 Causal effect estimates

In Table 2, we compare the average treatment on the control (ATC) from VRM-E to the ATC from the unadjusted estimate and estimate from SPSM. We calculate the 95% confidence interval (CI) given the percentile bootstrap method (Hahn, 1995) resampling the 2017 papers; see Appendix D.2. All three methods estimate an ATC with a negative sign. This result suggests that for the same manuscript, shifting from single-blind to double-blind would decrease the average reviewer score. However, while the unadjusted estimate of the ATC is -0.25 with 95% CI [-0.39, -0.11], our method shifts the ATC to -0.17 [-0.29, -0.05], a slightly smaller difference between the outcomes of single- and double-blind policies.

5 Conclusion and future work

In this work, we implement VRM-E, a method for text-based causal confounding adjustment that satisfies our three criteria of empirically checking causal overlap, incorporating modern NLP embeddings, and human validation of intermediate steps. For our case study, we find domain-experts prefer VRM-E matches 70% of the time compared to random matches and stratified propensity score matches. While the sign of the causal effect—negative—of switching from single blind to double blind reviewing on average reviewer scores is consistent across all methods, VRM-E estimates a slightly less negative effect.

Future work could investigate the causal mechanisms behind the negative causal effect and explore heterogeneous treatment effects due to author identity. Additional directions could examine different causal outcomes on the text of peer reviews such as discourse-level sentence labels (Kennard et al., 2022) or politeness (Danescu-Niculescu-Mizil et al., 2013). We hope this case study enables exploration of additional text-based causal estimation methods and domains in the future.

[7] Some prior work on text-based causal adjustment uses structured metadata to evaluate adjustment, e.g., Roberts et al. (2020); Sridhar et al. (2018). However, many studies do not have access to this type of structured data, or, as in our case, are not able to use it as ground truth (see Appendix C.2).

[8] Although SPECTER embeddings encode both the title and abstract, our human judges only evaluate the title for consistency and speed. Future work could investigate human judgements with the abstract.

Limitations

Our work is limited in several ways. We use human judgements on our case study data to demonstrate a preference of VRM-E versus SPSM. However, additional case studies in other domains such as education, healthcare, legal studies etc. are necessary in order to gather empirical evidence that preference for VRM-E generalizes.

Threats to validity. There are several threats to interpreting our case study estimates as causal. Like any causal study with observational data, our case study relies on untestable causal identification assumptions such as no unmeasured confounding. Other unmeasured confounding likely does exist. For example, our document embeddings do not necessarily measure the "quality" of the manuscripts or the "novelty" of the ideas, both of which could affect reviewers' scores. Regarding estimation, by allowing for matching with replacement, Appendix C.1.1 shows that several manuscripts are reused with high frequency. This will introduce bias within our model as noted in Stuart (2010). Additionally, our choice of b satisfies overlap but at the expense of very similar semantic matches between manuscripts. This could explain why there was only a moderate amount of agreement between the human judges as many matches are less semantically similar than we would prefer.

Ethics Statement

Data. Our case study data comes from Zhang et al. (2022) who aggregated data from OpenReview and other venues. Our work falls in line with the intention of Zhang et al. (2022): to investigate peer review. While individuals and research groups may not have intended for their work to be studied in the manner of our case study, we believe that the risk is minimal because researchers have agreed to publish their work via the OpenReview platform. The original intent of OpenReview was to create more transparency within the peer review process and allow for the analysis of various policies (Soergel et al., 2013). Risks are further minimized since we do not analyze individual manuscripts but rather focus on aggregate policy implications.

Peer review. A second ethical implication of our work concerns acting on our substantive findings about peer review. Our work primarily focuses on comparing text-based causal matching methods, so we do not focus on a sophisticated quasi-experimental design for the case study, and we do not analyze additional confounders other than the title and abstract of manuscripts. As mentioned in Sections 1 and 2, peer review is a complicated topic with conflicting analyses based on context. There are many different perspectives on what makes a peer review process "good". We hope our work is a step towards an improved peer review process, but we caution against using the results of this study in isolation as a basis for setting or changing any peer review policies.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under Grant Number 2022435, and in part by the Chan Zuckerberg Initiative under the project Scientific Knowledge Base Construction. KK is grateful for support from a Young Investigator Grant from the Allen Institute for Artificial Intelligence. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Aliaksandr Birukou, Joseph Wakeling, Claudio Bartolini, Fabio Casati, Maurizio Marchese, Katsiaryna Mirylenka, Nardine Osman, Azzurra Ragone, Carles Sierra, and Aalam Wassef. 2011. Alternatives to peer review: Novel approaches for research evaluation. *Frontiers in Computational Neuroscience*, 5.

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Comput. Surv.*, 54(2).

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors.

Mark Davies and Joseph L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, 38(4):1047–1051.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *International AAAI Conference on Web and Social Media (ICWSM)*.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. 2021. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348:1–17.

Samir Haffar, Fateh Bazerbachi, and M. Hassan Murad. 2019. Peer review bias: A critical review. *Mayo Clinic Proceedings*, 94(4):670–676.

Jinyong Hahn. 1995. Bootstrapping quantile regression estimators. *Econometric Theory*, 11(1):105–121.

Bruce E Hansen. 2000. Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603.

Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.

Neha Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1234–1249, Seattle, United States. Association for Computational Linguistics.

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

Emre Kiciman, Scott Counts, and Melissa Gasser. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *Twelfth International AAAI Conference on Web and Social Media*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

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.

Kewei Ming and Paul R Rosenbaum. 2001. A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics*, 10(3):455–463.

Saif M. Mohammad and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey. *CoRR*, abs/1203.1858.

Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Reagan Mozer, Luke Miratrix, Aaron Russell Kaufman, and L Jason Anastasopoulos. 2020. Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*.

Markus Neuhäuser, Matthias Thielmann, and Graeme D. Ruxton. 2018. The number of strata in propensity score stratification for a binary outcome. *Archives of Medical Science*, 14(3):695–700.

Whitney K Newey and James R Robins. 2018. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Kanu Okike, Kevin T. Hug, Mininder S. Kocher, and Seth S. Leopold. 2016. Single-blind vs Double-blind Peer Review in the Setting of Author Prestige. *JAMA*, 316(12):1315–1316.

Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proceedings of the 2017 ACM Conference on*

Computer Supported Cooperative Work and Social Computing, pages 370–386. ACM.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, Second edition. Cambridge University Press.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

Thomas Richardson. 2003. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157.

Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Paul R Rosenbaum and Donald B Rubin. 1985a. The bias due to incomplete matching. *Biometrics*, pages 103–116.

Paul R Rosenbaum and Donald B Rubin. 1985b. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. volume 66, page 688. American Psychological Association.

Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 440–451.

David Soergel, Adam Saunders, and Andrew McCallum. 2013. Open scholarship and peer review: a time for experimentation. In *ICML 2013 Workshop on Peer Reviewing and Publishing Models*. ICML.

Dhanya Sridhar, Aaron Springer, Victoria Hollis, Steve Whittaker, and Lise Getoor. 2018. Estimating causal effects of exercise from mood logging data. In *IJCAI/ICML Workshop on CausalML*.

Elizabeth A. Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1).

Mengyi Sun, Jainabou Barry Danfa, and Misha Teplitskiy. 2022. Does double-blind peer review reduce bias? evidence from a top computer science conference. *Journal of the Association for Information Science and Technology*, 73(6):811–819.

Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1).

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

Alina Tvina, Ryan Spellecy, and Anna Palatnik. 2019. Bias in the peer review process: Can we do better? *Obstetrics & Gynecology*, 133(6).

Victor Veitch, Dhanya Sridhar, and David M Blei. 2020. Adapting text embeddings for causal inference. In *UAI*.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4598.

Lingfei Wu, Ian En-Hsu Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J Witbrock. 2018. Word mover's embedding: From word2vec to document embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4524–4534.

Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 497–506, New York, NY, USA. Association for Computing Machinery.

Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. 2022. Investigating fairness disparities in peer review: A language model enhanced approach. *arXiv preprint arXiv:2211.06398*.

A Causal estimators for ATC

In this section, we describe the estimator for average treatment effect on the control (ATC). Let T_0 be the set of units for which we observe $T = 0$ and T_1 be the set of units that we observe $T = 1$; let N_{T_0} and N_{T_1} be the number of units in those two sets respectively. Then the theoretical ATC (with counterfactual terms) is

$$\tau^{ATC} = \frac{1}{N_{T_0}} \sum_{i \in T_0} \Big(Y_i(T_i = 1) - Y_i(T_i = 0) \Big) \tag{4}$$

The naive estimator assumes that the mean outcome over all units in T_1 suffices as the approximate counterfactual for every unit in T_0

$$\hat{\tau}^{ATC}_{\text{naive}} = \frac{1}{N_{T_0}} \sum_{i \in T_0} \Big((\frac{1}{N_{T_1}} \sum_{j \in T_1} Y_j) - Y_i \Big) \tag{5}$$

Following Stuart (2010); Morgan and Winship (2015), we use the following ATC matching estimator for VRM-E and SPSM. For each control unit, $i \in T_0$ and its corresponding matches M_i—matches are treated units in the same cluster for VRM-E and treated units in the same strata for SPSM—the estimator creates a "counterfactual" outcome from the mean of the matches

$$\hat{Y}_i(1) = \frac{1}{|M_i|} \sum_{j \in M_i} Y_j \tag{6}$$

which is substituted into

$$\hat{\tau}^{ATC}_{\text{match}} = \frac{1}{N_{T_0}} \sum_{i \in T_0} \Big(\hat{Y}_i(1) - Y_i \Big) \tag{7}$$

Intuitively, this estimator weights the $P(C = c)$ term Eq. 3 as the number of T_0 manuscripts in each cluster in VRM-E and each strata in SPSM.

B Hyperparameter choice and robustness

B.1 Human judgements on matches

title_2017	title_2018	Rating
#Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning	A Framework for the Quantitative Evaluation of Disentangled Representations	0
#Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning	Learning Dynamic State Abstractions for Model-Based Reinforcement Learning	0
#Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning	Learning to Mix n-Step Returns: Generalizing Lambda-Returns for Deep Reinforcement Learning	1
A Context-aware Attention Network for Interactive Question Answering	Unsupervised Representation Learning by Predicting Image Rotations	0
A Context-aware Attention Network for Interactive Question Answering	Incremental Learning through Deep Adaptation	0
A Context-aware Attention Network for Interactive Question Answering	Topic-Based Question Generation	1
A Learned Representation For Artistic Style	XGAN: Unsupervised Image-to-Image Translation for many-to-many Mappings	1
A Learned Representation For Artistic Style	Reinforcement Learning on Web Interfaces using Workflow-Guided Exploration	0
A Learned Representation For Artistic Style	Wasserstein Auto-Encoders	0
A Neural Stochastic Volatility Model	Learning Sparse Latent Representations with the Deep Copula Information Bottleneck	0
A Neural Stochastic Volatility Model	Representing Entropy : A short proof of the equivalence between soft Q-learning and policy gradients	1
A Neural Stochastic Volatility Model	Initialization matters: Orthogonal Predictive State Recurrent Neural Networks	0

Figure 2: Spreadsheet used for human judging of similarity of matches. Each judge was given a sheet with 300 rows of titles to rate based on our procedure described in Section 4.1. The same 2017 title is compared to three 2018 titles from three different methods and method names are masked in this spreadsheet. On average the three judges took 48 minutes (50, 44, and 50 minutes dis-aggregated) for the task.

B.2 Choosing b hyperparameter for VRM-E

In Figure 3, we show a visualization of our choice of b in VRM-E for our case study data. We select b such that it is the minimal value (minimizing bias in the estimates) while satisfying overlap (a necessary condition for causal estimation).

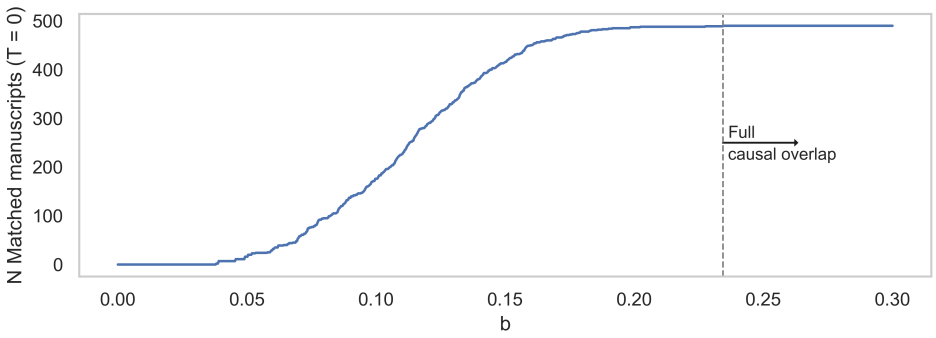

Figure 3: Plot for the choice of b for VRM-E. X-axis is the choice of b, the maximum cosine distance in Step 4 of VRM-E. Y-axis is the number of manuscripts in the anchor group (2017 manuscripts; $T = 0$) that have at least one match in the non-anchor group. Above the threshold of $b = 0.2343$, every manuscript in the anchor group is matched with at least one manuscript in the non-anchor group and thus overlap is satisfied.

C Qualitative evaluation and examples

C.1 VRM-E example

2017 title: Machine Comprehension Using Match-LSTM and Answer Pointer	Cosine Distance
2018 title: QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension	0.104
2018 title: Multi-Mention Learning for Reading Comprehension with Neural Cascades	0.11
2018 title: LEARNING TO ORGANIZE KNOWLEDGE WITH N-GRAM MACHINES	0.129
2018 title: FAST READING COMPREHENSION WITH CONVNETS	0.131
2018 title: ElimiNet: A Model for Eliminating Options for Reading Comprehension with Multiple Choice Questions	0.133
2018 title: FusionNet: Fusing via Fully-aware Attention with Application to Machine Comprehension	0.133
2018 title: Dynamic Integration of Background Knowledge in Neural NLU Systems	0.139
2018 title: Neural Compositional Denotational Semantics for Question Answering	0.142
2018 title: Phase Conductor on Multi-layered Attentions for Machine Comprehension	0.144
2018 title: Adaptive Memory Networks	0.149

Table 3: Example of 2017 manuscript matched with ten 2018 manuscripts. The right-most column is the cosine distances between embeddings of the respective manuscripts.

C.1.1 Repeated matches example

Title of 2018 Manuscript	Repeat Count
Neumann Optimizer: A Practical Optimization Algorithm for Deep Neural Networks	55
LSH Softmax: Sub-Linear Learning and Inference of the Softmax Layer in Deep Architectures	34
Revisiting Bayes by Backprop	32
Latent Space Oddity: on the Curvature of Deep Generative Models	32
A Bayesian Perspective on Generalization and Stochastic Gradient Descent	31

Table 4: Examples of the top 5 repeated manuscripts in VRM-E for our case study data. The right-most column is the number of times the 2018 manuscript has been matched to a different 2017 manuscript.

C.2 Structured keywords

In both 2017 and 2018, submissions to ICLR were accompanied by a list of keywords selected by the authors. Initially, we attempted to use these keywords as ground-truth by which we could evaluate VRM-E by comparing the Standard Difference in Means (SDM) (Stuart, 2010) for each keyword.

However, we found that this was not a valid approach for this ICLR dataset as the semantic function of keywords changed between 2017 and 2018. In 2017, all submissions selected from a set of 15 general keywords, e.g., `deep_learning` or `natural_language_processing`. In contrast, submissions in 2018

used a more varied set of keywords with finer granularity. The 1748 keywords used in 2018 included specific topics such as `attention` or `word_embeddings`.

D Additional empirical settings and results

D.1 Training logistic regression model for SPSM

In Section 3.4, we train a logistic regression model using the `scikit-learn` Python package (Pedregosa et al., 2011). We conduct a grid search, resulting in the best parameters listed in Table 5. This model was used to estimate the propensity scores of the manuscripts. The performance metrics from training are listed in Table 6. Figure 4 and Table 7 give the distribution of the propensity scores at inference time.

Parameter	Input
Model	LogisticRegression
l1_ratio	0.1
solver	saga
max_iter	20000
tol	0.001
penalty	elasticnet
dual	False
class_weight	balanced
random_state	42
model__C	0.01
run time	19.3 seconds

Table 5: Best Parameters for logistic regression

Metric	Score
Accuracy	0.62
Average Precision Score	0.54
Calibration RSME	0.19
F1	0.54
Mean Prediction Binary	0.47
Mean Prediction Decimal	0.49
Mean Prediction	0.35
ROC AUC	0.68

Table 6: Logistic regression performance metrics for the training folds.

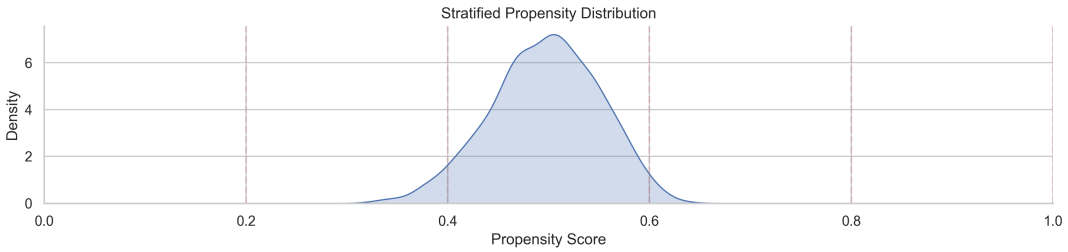

Figure 4: Propensity score distribution on the inference folds. Visualization is created using a kernel density estimate.

Strata	$P(T=1)$ range	N_{T_0} (2017 manuscripts)	N_{T_1} (2018 manuscripts)
1	[0.0, 0.2]	0	0
2	[0.2, 0.4]	9	45
3	[0.4, 0.6]	472	851
4	[0.6, 0.8]	9	14
5	[0.8, 1.0]	0	0

Table 7: Distribution of manuscripts within each strata for SPSM.

D.2 Statistical significance

Table 1. For the human judgement results presented in Table 1, we conduct a two-sided T-test on the distributions of preferences for pairs of methods. Comparing VRM-E to random matches and SPSM,

we obtain T-statistic of 10.8 and 8.7 respectively. Both have a p-value of less than 0.01 (far below the threshold of rejection).

Table 2. To obtain confidence intervals and test for statistical significance in Table 2, we use bootstrapping. Since we evaluate the average treatment effect on the control, we sample with replacement the 2017 manuscripts 5000 times. For each bootstrap sample, we then calculate the ATC estimates for VRM-E, SPSM, and the naive (unadjusted) approaches. We calculate the 95% confidence interval (CI) given the percentile bootstrap method (Hahn, 1995). We use the 97.5 percentile and 2.5 percentile of the bootstrap samples to determine the confidence interval.

To determine if there is a statistically significant difference between pairs of method, we use a paired bootstrap approach (Tibshirani and Efron, 1993). Specifically, we follow the algorithm from Berg-Kirkpatrick et al. (2012)'s Figure 1. When comparing the difference between the ATC for VRM-E and Naive and VRM-E and SPSM, both obtain a p-value of less than 0.01 (far below the threshold of reject).

D.3 Robustness to hyperparameter selection

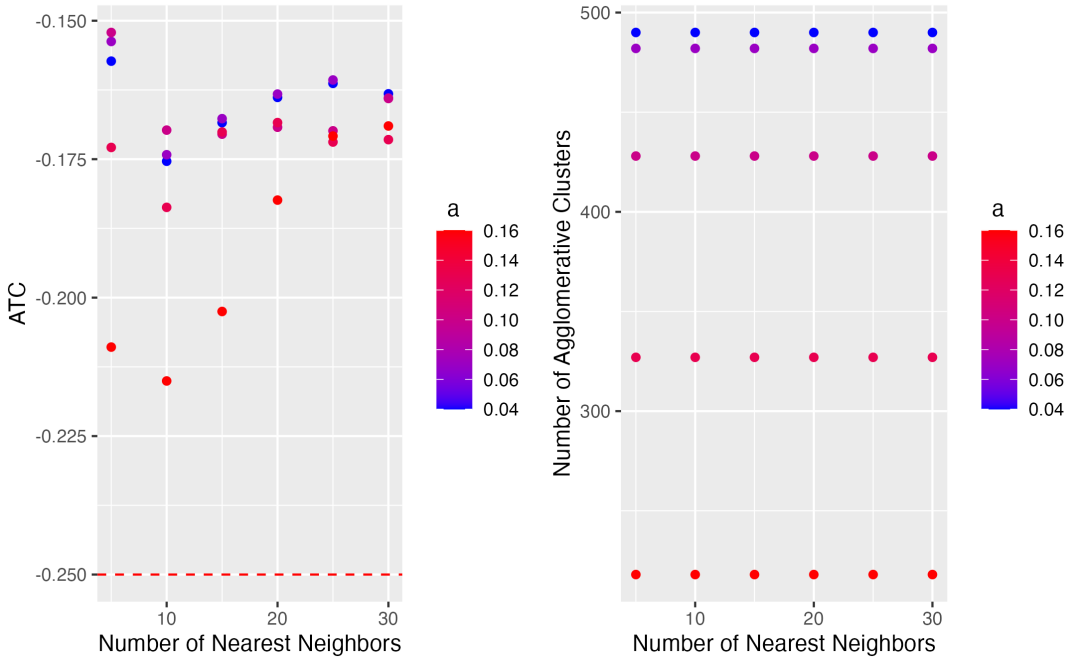

Figure 5: We re-run VRM-E for different choices of k, maximum number of nearest neighbors (x-axis) for different values of a (colors). **Left.** The y-axis is the ATC. The dashed red horizonal line is the unadjusted (baseline) ATC. **Right.** The y-axis is the number of agglomerative clusters (Step 3 in VRM-E) given the choice of a.

In Figure 5, we compare the ATC and number of agglomerative clusters given different hyperparameter choices. We do not use these plots to select hyperparameters (since there is no ground-truth in causal estimation) but rather to inspect our results' robustness to these choices post-hoc. As expected, increasing a—the hyperparameter which specifies the cosine distance threshold for which we count anchor units to be similar—decreases the number of agglomerative clusters (Step 3 in VRM-E) but also changes the ATC to be slightly more negative. As shown in the left plot of Figure 5, the estimates of the ATC for all our choices of hyperparameters are still less negative than the baseline unadjusted ATC.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation section

☑ A2. Did you discuss any potential risks of your work?
The ethics statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1 introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

OpenReview data see section 3.1

☑ B1. Did you cite the creators of artifacts you used?
section 3.1

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics Statement

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethics statement

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
section 3.1

C ☑ Did you run computational experiments?

section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix F.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix F

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.