How Effective is Your Rebuttal? Identifying Causal Models from the OpenReview System

Anonymous Author(s)
Affiliation
Address
email

Abstract

The peer review process is central to scientific publishing, with the rebuttal phase offering authors a critical opportunity to address reviewers' concerns. Yet the causal mechanisms underlying rebuttal effectiveness, particularly how author responses influence final review decisions, remain unclear. To uncover the mechanisms driving rebuttal effectiveness, we model it as a causal representation learning (CRL) problem. Using data from the OpenReview system for ICLR submissions, we examine how rebuttal characteristics of both reviewers and authors causally affect post-rebuttal rating changes. We introduce a weakly supervised disentangled CRL framework that leverages review subscores (e.g., openness, clarity, directness) as concept-level supervision. Theoretically, we establish identifiability conditions for latent variables across multiple distributions, showing that human-interpretable concepts can be recovered under mild assumptions. Empirically, our results uncover distinct causal patterns governing successful rebuttals, revealing how specific strategies differentially influence review criteria. These findings provide actionable guidance for authors in crafting effective rebuttals, while offering broader implications for transparency, fairness, and efficiency in the peer review process.

1 Introduction

Scientific progress is fundamentally dependent on peer review, a system designed to ensure the quality, validity, and integrity of published research [1, 2, 3, 4]. Recently, the machine learning community has embraced greater transparency through platforms like OpenReview, where reviews, rebuttals, and rating changes are publicly accessible [5, 6, 7, 8]. This transparency creates an unprecedented opportunity to analyze the dynamics of peer review, particularly the effectiveness of author rebuttals in influencing reviewers' final assessments. Despite its importance, little is empirically known about what makes a rebuttal effective and how specific strategies influence reviewers' final decisions.

Prior work on peer review has examined systemic properties such as bias [9], arbitrariness [10], and predictive validity [11, 12]. More recently, studies leveraging OpenReview data have explored reviewer behavior and rating consistency [13, 14], and randomized trials have tested reviewer anchoring effects [15]. The effectiveness of rebuttals has been touched upon. For instance, [16] report that rebuttals trigger rating changes in roughly 25% of cases, but causal mechanisms remain unexplored. Existing analyses focus largely on correlations, leaving an open question of why certain rebuttal strategies succeed while others fail, and under what contextual conditions they are effective.

In parallel, advances in causal representation learning (CRL) provide powerful tools for uncovering latent causal factors from high-dimensional data [17]. While many CRL methods operate in unsupervised settings [18, 19, 20, 21], recent work demonstrates that weak supervision from concept labels [22, 23, 24, 25] can better align latent factors with human-interpretable dimensions. Complementary strategies that exploit multiple distributions [21, 26], multiple modalities [20, 27],

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Table 1: **Performance comparison** of LLMs in generating subscores for rebuttal. Besides the average Time Cost (TC), L_2 norm between human annonation and LLM estimation is reported for other 10 subscores, including Clarity (CL), Directness (DI), Attitude (AT), Authors Openness (AO), Evidence (EV), Rigor (RI), De-Escalation (DE), Review Quality (RQ), Reviewer Openness (RO), and Concern Severity (CS). The final column reports the average L_2 error (AE) across all 10 subscores.

Models (LLMs)	Metrics											
	TC↓	**CL↓**	**DI↓**	**AT↓**	**AO↓**	**EV↓**	**RI↓**	**RI↓**	**RQ↓**	**RO↓**	**CS↓**	**AE↓**
DeepSeek-R1	18.33s	0.30	0.55	0.47	0.60	0.83	0.63	0.67	0.80	0.55	0.80	0.62
Grok-3-Latest	11.74s	0.45	0.50	0.74	0.60	0.39	0.53	0.67	1.00	0.90	1.30	0.71
Gemini-2.0-Flash-Lite	3.70s	0.40	0.75	0.53	0.95	0.89	0.58	1.00	0.30	1.10	0.65	0.71
ChatGPT-4.1-Mini	9.73s	0.35	0.65	0.95	0.95	0.94	0.63	1.00	1.25	0.60	0.65	0.80
Gemini-2.0-Flash	4.12s	0.45	1.10	0.79	0.95	1.11	0.74	0.83	0.75	1.05	0.45	0.82
ChatGPT-4.1	9.73s	0.55	0.70	0.79	1.05	0.89	0.84	0.67	1.40	0.90	0.65	0.84
ChatGPT-4.1-Nano	5.34s	0.35	0.55	1.05	1.30	0.72	0.58	2.50	0.50	0.50	0.75	0.88
Llama-4-Maverick	5.86s	0.50	0.75	1.26	1.35	1.28	0.89	2.00	0.75	0.70	0.75	1.02
Gemini-2.5-Flash-Preview-04-17	13.26s	0.65	0.75	0.74	1.35	0.94	0.84	1.17	1.45	1.20	1.25	1.03
Deepseek-V3-0324	5.94s	0.55	1.00	1.84	0.80	1.28	1.16	1.50	1.40	1.30	0.85	1.17
ChatGPT-4o-Latest	8.43s	0.60	1.30	2.00	1.40	1.17	0.95	2.17	2.00	1.05	0.80	1.34

nonstationarities [28, 29], or interventions [30, 31] provide variation that strengthens identifiability. At the same time, large language models (LLMs) now enable sophisticated analysis of scientific text [32, 33, 34], producing flexible representations that capture nuanced argumentation and discourse. Prior research in argument mining and persuasion [35, 36] highlights that textual properties such as clarity, directness, and evidence strength are central to rebuttal persuasiveness, yet these insights have not been systematically incorporated into causal analyses of peer review. To address this gap, we leverage pretrained LLMs to extract embeddings from rebuttal and review text, and analyze these representations within a causal framework, aiming to explain the underlying rebuttal process in OpenReview system and provide practical guidance to the broader machine learning community.

Using ICLR review and rebuttal data from OpenReview, we define 10 subscores (e.g., openness, clarity) that capture features of both reviewers and authors during the rebuttal process, extracted with state-of-the-art LLMs. We hypothesize that these subscores are linked to reviewers' rating changes and aim to verify it. We proposed a causal representation learning framework which leverages the subscores as weak supervision. To summarize, our contributions are threefold: (i) we introduce rebuttal effectiveness as a causal modeling problem and formalize it within a multi-distribution CRL framework, (ii) we establish theoretical identifiability results showing that human-interpretable concepts can be recovered under mild assumptions, and (iii) we provide an empirical analysis uncovering population-level causal patterns in rebuttal success. Ultimately, our goal is to reveal which factors are most causally related to rating change and to offer the machine learning community concrete guidance on how to craft more effective and efficient rebuttals in OpenReview.

2 Dataset and Observations

Data Collection and Processing. We analyze peer-review data from the Paper Copilot platform [37], which aggregates review data from major artificial intelligence (AI) conferences. We focus on ICLR 2024 (7304 submissions) and 2025 (11672 submissions) hosted on the OpenReview system, as these cycles provide full access to author–reviewer discussions and both pre- and post-rebuttal ratings, which are essential for studying rating change. Each submission includes metadata (e.g., title, author list, abstract), reviewer initial ratings and comments, author rebuttals, and follow-up discussions. Desk-rejected papers are removed. Since our goal is to analyze reviewer–author interactions rather than paper-level outcomes, each review–rebuttal thread is treated as an independent sample. To ensure meaningful interactions, we exclude cases without rebuttals or without reviewer responses. After filtering, we obtain 23922 valid reviewer–author discussions (each stands for one sample).

Self-defined Subscores and LLM Inference. For fine-grained analysis, we randomly select 10% of samples from each primary area (2393 in total) for annotation. We define ten interpretable subscores: seven author-related (*clarity, directness, positive attitude, acknowledgment of limitations, strength of evidence, technical rigor, handling of misunderstandings and de-escalation*) and three reviewer-related (*specificity, open-mindedness, severity of concerns*), each rated on a five-point ordinal scale. To scale beyond manual labeling, we benchmarked multiple large language models against a 20-

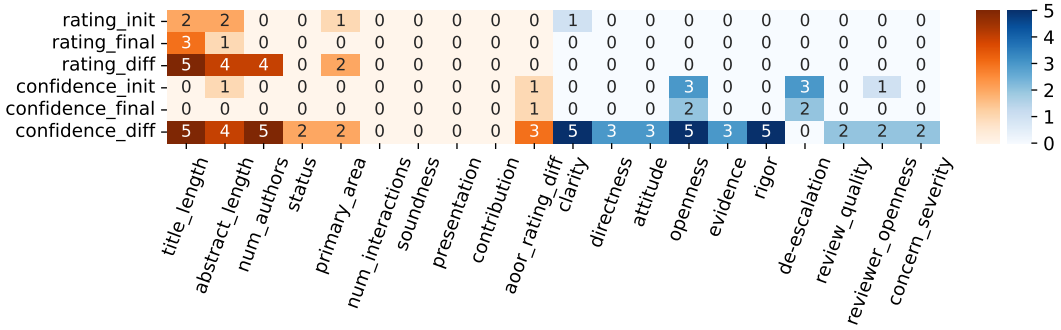


Figure 1: **Independence test results** between rating/confidence and metadata (red) or LLM-inferred subscores (blue). Each cell shows the number of tests (out of five different methods) failed to reject the null hypothesis of independence at significance level α=0.05. A value of 0 indicates strong evidence of dependence, while 5 indicates strong evidence of independence across all applied tests.

example seed set (see Tab. 1) and selected **DeepSeek-R1**, which showed the highest agreement with human annotations, to automatically label the 10% subset. Refer to App. A2 for more details.

Independence Tests. To analyze how rebuttal and reviewer attributes relate to rating change, we apply five independence tests: KCI [38], RCSI [39], HSIC [40], Chi-square [41], and G-square [42]. Fig. 1 reports the aggregated outcomes, and Fig. A1 shows the distributions of all metadata and subscores. Independence tests on metadata are conducted using the full set of 23922 samples, while tests on subscores are based on the annotated 2393-sample subset.

Takeaway Messages. We summarize several key observations: (i) All of our self-defined author subscores are strongly associated with rating change, suggesting that well-supported and constructive rebuttals—especially those that are clear, direct, positive, and open—are particularly effective. (ii) *Review quality, reviewer open-mindedness*, and *severity of concerns* predict rating adjustments, indicating that significant shifts occur when reviewers are receptive and rebuttals directly address serious issues. (iii) Metadata features such as *title length, abstract length, and number of authors* show little relation to rating change. (iv) *Primary area* exhibits only weak dependence on rating change. (v) The *number of interactions* correlates with rating change, underscoring the importance of back-and-forth engagement in reviewer discussions. (vi) The average rating of other reviewers is strongly correlated with an individual reviewer's rating change, highlighting peer influence among reviewers. (vii) Finally, confidence ratings appear largely *independent* of rebuttal and reviewer attributes, suggesting they reflect intrinsic reviewer disposition rather than rebuttal content. (viii) Overall, these findings indicate that rebuttal effectiveness depends less on superficial characteristics and more on substantive qualities (*e.g., rigor, etc.*) combined with reviewer willingness to reconsider. This provides a strong empirical foundation for modeling rebuttal effectiveness within a causal framework and offers practical guidance for authors on how to craft more effective rebuttals.

3 Causal Formulation and Identifiability Theory

Let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]$ denote the observed data, where $\mathbf{x}_1$ represents the aggregated text from reviewers and $\mathbf{x}_2$ the aggregated text from authors. Let $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2]$ denote the corresponding latent variables, with $\mathbf{z}_1$ linked to reviewers and $\mathbf{z}_2$ to authors. We further introduce θ as a latent factor capturing changes across domains (e.g., primary areas).

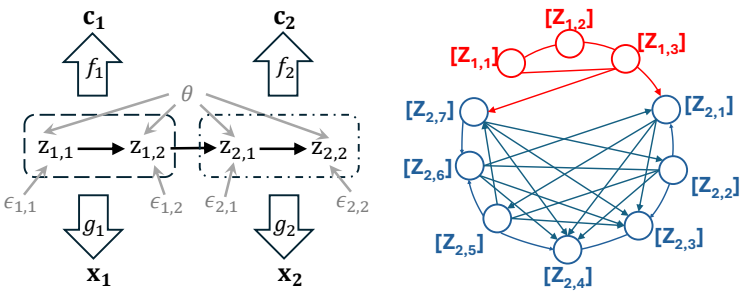


Figure 2: *Left:* True causal model. *Right:* Learned latent causal graph.

In addition to $\mathbf{x}$ and $\mathbf{z}$, we consider a set of human-interpretable concepts $\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2]$, referring to reviewers and authors respectively. These concepts capture qualitative attributes such as *rigor, evidence, or openness*. We assume $\mathbf{c}$ is a linear transformation of $\mathbf{z}$, thus providing interpretable, low-dimensional views of latent causes. Formally, the data generation process (Fig. 2) can be expressed as follows: (1) assume $m \in \{1, 2\}$ indexes reviewers and authors respectively. Each latent variable is generated by $z_{m,i} = h_{m,i}(\text{Pa}(z_{m,i}), \theta, \epsilon_{m,i})$, where $\text{Pa}(z_{m,i})$ denotes its parents in the latent causal graph $\mathcal{G}_{\mathbf{z}}$, $\epsilon_{m,i}$ are mutually independent exogenous noise variables, and θ represents domain-specific factors; (2) observations are generated by $\mathbf{x}_m = g_m(\mathbf{z}_m)$, with $g_m(\cdot)$ a nonlinear mixing function; (3) human-aligned concepts $\mathbf{c}_m$ are generated by applying an affine transformation to a point $\tilde{\mathbf{z}}_m$ from the affine subspace of $\mathbf{z}_m$: $\mathbf{c}_m = f_m(\tilde{\mathbf{z}}_m)$, where $f_m(\cdot)$ denotes a linear mapping. Our goal is to leverage multiple datasets across varying contexts θ to recover the latent variables $\mathbf{z}$ and the concepts $\mathbf{c}$ and their causal relationships (up to inherent indeterminacies). Achieving this is crucial for providing authors with causal insights into which rebuttal strategies most affect rating changes. We present the main theoretical results for our setting here and show the full results in App. A3.3.

Theorem 1. ***(Identifiability of Review Concepts)*** *Suppose we match the observations $\mathbf{x}_m$ across modalities (authors and reviewers), and the following conditions hold in the data-generating process:*

i *(Information Preservation): The functions g_1 and g_2 are differentiable and invertible.*

ii *(Primary Area Diversity): All entries of $v^\top B$ are non-zero, where $B_{i,j} = \frac{b_k^e}{\sigma^2}$ denotes the area–concept matrix.*

iii *(Thought Reflection): The latent components in $\mathbf{z}_1$ are causal parents of $\mathbf{z}_2$, but not vice versa.*

iv *(Distinctive Concept Alignment): There exists a set of linearly independent aligning vectors $\mathcal{C} = \{a_1, \ldots, a_n\}$ such that, for each concept C^e, the rows of the aligning matrix A^e lie in $\mathcal{C}$, i.e., $(A^e)^\top e_i \in \mathcal{C}$. Let S^e denote the indices of the subset of $\mathcal{C}$ that appear as rows of A^e. Every aligning vector in $\mathcal{C}$ appears in at least one area e (where an area corresponds to a concept-conditional distribution), that is,*

$$\bigcup_e S^e = [n].$$

Then the review concepts are identifiable as in Definition 1.

Discussion of Assumptions Assumption i requires that the latent space is recoverable from the observed data. Assumption ii further requires the presence of latent distribution shifts in the review concepts across different primary areas, ensuring variability in the underlying structure. Assumption iii reflects the natural process in which authors first read the reviews, then engage in reflection, and finally provide rebuttals. Finally, Assumption iv ensures that all concepts can be decomposed into a finite set of atomic components that remain distinct across primary areas, which is essential.

Experiments and Analysis. Building on our identifiability theory, we design a network architecture with corresponding loss functions (See App.A4). We first use the CRL framework to learn the latent variables $\mathbf{z}$, and then apply causal discovery methods to recover the causal graph among them (see Fig. 2 (Right)). We evaluate the resulting model on two dimensions: (i) predictive performance, i.e., its ability to predict rating change, and (ii) interpretability, i.e., alignment with human-understandable aspects of the review process. The learned latent variables can be interpreted as meaningful review concepts. In particular, $z_{1,1}$, $z_{1,2}$, and $z_{1,3}$ capture reviewer-related subscores, while $z_{2,1}$ through $z_{2,7}$ correspond to rebuttal-related subscores. Notably, $z_{1,3}$ exerts a direct influence on both $z_{2,1}$ and $z_{2,7}$, highlighting a causal pathway from reviewers to authors in the rebuttal process.

4 Conclusion

We introduced rebuttal effectiveness as a causal modeling problem and developed a weakly supervised causal representation framework using ICLR review data from OpenReview. By combining fine-grained annotations with independence tests, we found that substantive rebuttal qualities together with reviewer receptiveness strongly drive rating changes, while superficial factors like title length or author count have little impact. Theoretically, we established conditions under which human-aligned review concepts are identifiable, enabling interpretable causal analysis. Our findings provide actionable guidance for authors and broader implications for transparency and fairness in peer review, while opening directions for extending causal analyses to other venues and discourse-level features.

References

[1] Jonathan P Tennant. The state of the art in peer review. *FEMS Microbiology letters*, 365(19):fny204, 2018.

[2] Bruce Alberts, Brooks Hanson, and Katrina L Kelner. Reviewing peer review, 2008.

[3] Jonathan P Tennant and Tony Ross-Hellauer. The limitations to our understanding of peer review. *Research integrity and peer review*, 5(1):6, 2020.

[4] Stephen J Ceci and Douglas P Peters. Peer review: A study of reliability. *Change: The Magazine of Higher Learning*, 14(6):44–48, 1982.

[5] David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of openreview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*, 2020.

[6] Gang Wang, Qi Peng, Yanfeng Zhang, and Mingyang Zhang. What have we learned from openreview? *World Wide Web*, 26(2):683–708, 2023.

[7] Hao Sun, Yunyi Shen, and Mihaela van der Schaar. Openreview should be protected and leveraged as a community asset for research in the era of large language models. *arXiv preprint arXiv:2505.21537*, 2025.

[8] Tony Ross-Hellauer. What is open peer review? a systematic review. *F1000Research*, 6:588, 2017.

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

[10] John Langford and Mark Guzdial. The arbitrariness of reviews, and advice for school administrators. *Communications of the ACM*, 58(4):12–13, 2015.

[11] Azzurra Ragone, Katsiaryna Mirylenka, Fabio Casati, and Maurizio Marchese. On peer review in computer science: Analysis of its effectiveness and suggestions for improvement. *Scientometrics*, 97:317–356, 2013.

[12] Dietmar Wolfram, Peiling Wang, Adam Hembree, and Hyoungjoo Park. Open peer review: promoting transparency in open science. *Scientometrics*, 125(2):1033–1051, 2020.

[13] Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. 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, 2021.

[14] Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major nlp conference. *arXiv preprint arXiv:1903.11367*, 2019.

[15] Ryan Liu, Steven Jecmen, Vincent Conitzer, Fei Fang, and Nihar B Shah. Testing for reviewer anchoring in peer review: A randomized controlled trial. *PloS one*, 19(11):e0301111, 2024.

[16] Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the nips 2016 review process. *Journal of machine learning research*, 19(49):1–34, 2018.

[17] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[18] Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. *arXiv preprint arXiv:2409.02772*, 2024.

[19] Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. *arXiv preprint arXiv:2403.08335*, 2024.

[20] Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, et al. Causal representation learning from multimodal biomedical observations. *ArXiv*, pages arXiv–2411, 2025.

[21] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

[22] Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. From causal to concept-based representation learning. *Advances in Neural Information Processing Systems*, 37:101250–101296, 2024.

[23] Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models. *arXiv preprint arXiv:2402.09236*, 2024.

[24] Emanuele Marconato, Andrea Passerini, and Stefano Teso. Interpretability is in the mind of the beholder: A causal framework for human-interpretable representation learning. *Entropy*, 25(12):1574, 2023.

[25] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pages 6348–6359. PMLR, 2020.

[26] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

[27] Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

[28] Xiangchen Song, Zijian Li, Guangyi Chen, Yujia Zheng, Yewen Fan, Xinshuai Dong, and Kun Zhang. Causal temporal representation learning with nonstationary sparse transition. *Advances in Neural Information Processing Systems*, 37:77098–77131, 2024.

[29] Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. *Advances in Neural Information Processing Systems*, 36:8092–8113, 2023.

[30] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pages 372–407. PMLR, 2023.

[31] Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36:50254–50292, 2023.

[32] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

[33] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

[34] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

[35] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624, 2016.

[36] John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, 2020.

[37] Jing Yang. Paper copilot: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. *arXiv e-prints*, pages arXiv–2502, 2025.

[38] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

[39] Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):20180017, 2019.

[40] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[41] Ronald J Tallarida, Rodney B Murray, Ronald J Tallarida, and Rodney B Murray. Chi-square test. *Manual of pharmacologic calculations: with computer programs*, pages 140–142, 1987.

[42] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78, 2006.

[43] Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. Lazyreview a dataset for uncovering lazy thinking in nlp peer reviews. *arXiv preprint arXiv:2504.11042*, 2025.

[44] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.

[45] Sebastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi LE PRIOL, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 428–484. PMLR, 11–13 Apr 2022.

[46] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

[47] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[48] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

[49] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

[50] Michel Besserve, Naji Shajarisales, Bernhard Schölkopf, and Dominik Janzing. Group invariance principles for causal generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 557–565. PMLR, 2018.

[51] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*. ICLR, 2020.

[52] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022.

[53] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[54] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. Specter: Document-level representation learning using citation-informed transformers, 2020.

[55] Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*, 2019.

[56] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406, 2018.

[57] Jianyou Andre Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. Scientific document retrieval using multi-level aspect-based queries. *Advances in Neural Information Processing Systems*, 36:38404–38419, 2023.

[58] David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. Multivers: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, 2022.

[59] Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. Argument mining driven analysis of peer-reviews. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4758–4766, 2021.

Appendix for

"How Effective is Your Rebuttal? Identifying Causal Models from the OpenReview System"

Table of Contents:

A1 Details about Related Work **9**

A1.1 Peer Review Analysis 9

A1.2 Causal Representation Learning 10

A1.3 Natural Language Processing for Scientific Text 10

A2 Details about the Dataset and Analysis **10**

A2.1 Explanation of Subscores in Tab.1 10

A2.2 Explanation of Variables in Fig.1 10

A2.3 Dataset Analysis 10

A3 Learning Human-aligned Causal Representations **11**

A3.1 Basic Concept 11

A3.2 Definition of Identifiability 11

A3.3 Identifiability of Causal Models: Theorem and Proof 12

A4 Implementation Framework **13**

A4.1 Variational Inference and Model Architecture 13

A4.2 Learning Objectives for Causal Concept Discovery 14

A4.2.1 Causal Sparsity Loss 14

A4.2.2 Weak Supervision Loss 14

A4.3 Final Objective Function 14

A1 Details about Related Work

A1.1 Peer Review Analysis

Peer review in scientific publishing has been widely studied, with work addressing bias [9], consistency [10], and predictive validity [11]. The transparency of the OpenReview platform has further enabled analyses of reviewer behavior and decision-making [13, 14]. Recent studies provide complementary perspectives. [15] conducted a randomized controlled trial and found that reviewers are not strongly anchored to their initial scores, showing a willingness to revise after rebuttals, though the drivers of such changes remain unclear. The LazyReview dataset [43] addresses a different challenge by identifying low-effort or vague reviews, offering tools to improve review quality. By contrast, the effectiveness of rebuttals themselves has received relatively limited attention. [16] showed that rebuttals lead to score changes in about 25% of reviews, while [14] explored correlates of successful rebuttals without establishing causality. Our work extends these efforts by explicitly modeling the causal mechanisms underlying rebuttal effectiveness.

A1.2 Causal Representation Learning

Causal representation learning (CRL) seeks to uncover latent causal factors from high-dimensional data [17, 44], enabling reasoning about interventions and counterfactuals. Recent work has shown that CRL can learn disentangled representations capturing causal mechanisms [45, 46], making it particularly useful in domains where causal factors are latent or noisy, such as peer review. Unsupervised CRL methods face identifiability challenges [47], which researchers have attempted to address using temporal structure [48], sparsity assumptions [49], or group-theoretic frameworks [50]. However, such assumptions often fail in real-world settings. To overcome this, weak supervision and multi-environment data have been proposed to improve identifiability [25, 51]. Building on weakly supervised approaches [52] and concept-based representation learning [22], our work adapts these ideas to model rebuttal effectiveness.

A1.3 Natural Language Processing for Scientific Text

Analyzing rebuttals requires handling complex scientific text. Advances in natural language processing have enabled richer analysis of scientific documents [53, 54], supporting tasks such as classification, summarization, citation intent detection [55, 56], document retrieval [57], and fact-checking [58]. While less explored, rebuttals have been studied through argument mining [36, 59] and persuasive language [35], reflecting their persuasive nature in influencing reviewer opinions.

Our work connects these directions by applying causal representation learning to study rebuttal effectiveness in scientific peer review, focusing on the OpenReview system in machine learning conferences.

A2 Details about the Dataset and Analysis

A2.1 Explanation of Subscores in Tab.1

We consider ten variables capturing key aspects of rebuttals and reviews. *Clarity (CL)* reflects how clearly the rebuttal communicates its arguments, while *Directness (DI)* measures the extent to which it addresses reviewer concerns explicitly. *Attitude (AT)* captures the tone of the rebuttal, distinguishing professional and respectful responses from defensive ones. *Authors Openness (AO)* denotes the willingness of authors to acknowledge limitations or alternative perspectives. *Evidence (EV)* refers to the use of data, experiments, or citations to support claims, and *Rigor (RI)* evaluates the technical soundness and thoroughness of rebuttal arguments. *De-Escalation (DE)* reflects the ability to resolve misunderstandings and reduce conflict during the exchange. On the reviewer side, *Review Quality (RQ)* measures the specificity and constructiveness of feedback, *Reviewer Openness (RO)* captures the willingness of reviewers to revise their evaluation in light of rebuttals, and *Concern Severity (CS)* indicates the seriousness of the issues raised in the review.

A2.2 Explanation of Variables in Fig.1

We further include metadata and reviewer-provided variables. *Title Length* and *Abstract Length* measure the verbosity of the submission's title and abstract, respectively, while *Num Authors* captures the number of contributing authors. *Status* indicates the acceptance outcome (e.g., oral, poster, reject), and *Primary Area* records the main research domain of the paper. *Num Interactions* reflects the extent of back-and-forth exchanges between authors and reviewers. Reviewer scores are also considered: *Soundness* assesses methodological correctness, *Presentation* evaluates clarity of exposition, and *Contribution* reflects novelty and significance. Finally, *aoor_rating_diff* measures the average of other reviews's rating differences or changes for one reviewer; we define this variable in order to see how one reviewer can be influenced by other reviewers.

A2.3 Dataset Analysis

For fine-grained analysis, we annotate a 10% random sample of the dataset with interpretable labels capturing both rebuttal quality and reviewer behavior. Rebuttal-related dimensions include *Clarity, Directness in Addressing Reviewer Concerns, Positive Attitude, Willingness to Acknowledge Limitations, Strength of Evidence, Technical Convincingness and Rigor, and Handling of Misunderstandings and*

De-escalation, while reviewer-related dimensions include *Review Specificity and Constructiveness, Open-mindedness, and Severity of Concerns*. All labels are rated on a 5-point ordinal scale, with detailed guidelines provided in the annotation prompt (Appendix).

To construct the annotated subset, we manually labeled 20 review–rebuttal threads and used them to benchmark 10 LLMs. We then computed the L-2 distance between model predictions and human labels across dimensions. As shown in Table 1, DeepSeek-R1 achieved the closest alignment to human annotations and was chosen to label the full 10% set.

In addition to the annotated labels, we extract further labels from OpenReview, including metadata such as *Title Length, Abstract Length, Number of Authors, Status, Primary Area, and Number of Reviewer-Author(s) Interactions*, as well as reviewer-provided scores *Soundness, Presentation, Contribution, Initial Rating, Final Rating, Initial Confidence, and Final Confidence*. Using this subset, we conduct pairwise independence tests with Kernel-based Conditional Independence (KCI), Randomized Conditional Independence (RCSI), Hilbert-Schmidt Independence Criterion (HSIC), Chi-squared, and G-squared tests. Detailed results for each method are given in the Appendix. Figure A1 summarizes the findings, where each cell shows how many of the five tests failed to reject the null hypothesis.

The aggregated results in Figure A1 reveal several patterns. *Rating Difference* shows strong dependence with *Openness*, *Evidence*, and *Rigor*, suggesting that reviewers who initially gave low scores are more likely to revise them when faced with open, well-supported, and rigorous rebuttals. *Number of Interaction* is also dependent on *Rating Difference*, reflecting the role of back-and-forth communication in driving score changes. By contrast, *Clarity*, *Directness*, and *Attitude* show no dependence with *Rating Difference*, likely due to their skewed distribution (most rebuttals score highly, leaving little variability) or the selection bias of top-tier conference submissions, where both papers and reviews tend to be of consistently high quality. Interestingly, *Clarity* and *Attitude* do show dependence with *Initial Rating* and *Final Rating*, but not with *Rating Difference*, implying that they shape the overall impression of a paper without directly influencing score updates.

We also find that *Reviewer Openness* and *Severity of Concerns* are strongly associated with *Rating Difference*, indicating that large score changes occur when open-minded reviewers engage with rebuttals addressing serious issues. In contrast, metadata features such as *Title Length*, *Abstract Length*, *Number of Authors*, and *Primary Area* show no dependence on *Rating Difference*, suggesting they play only a minor role compared to content-based signals. The dependence of *Status* on *Initial Rating*, *Final Rating*, and *Rating Difference* is expected, as decisions (e.g., oral, poster) follow review scores. Finally, *Confidence* scores appear largely independent of other features, suggesting they are influenced by external factors.

A3 Learning Human-aligned Causal Representations

A3.1 Basic Concept

To connect abstract latent variables with human-understandable criteria, we model review subscores (e.g., soundness, clarity, novelty) as *concepts*. Formally, each concept is defined as a linear projection $A : \mathbb{R}^{d_z} \to \mathbb{R}^{d_C}$ of the latent rebuttal representation $\mathbf{z}$, with a valuation $b \in \mathbb{R}^{d_C}$ corresponding to the reviewer's assigned subscore (e.g., clarity = 4). Thus, rebuttals with the same subscore form a concept-conditional set in latent space. This formulation anchors the learned representations to interpretable axes aligned with reviewer evaluations.

A3.2 Definition of Identifiability

The key question is whether these human-aligned concepts can be uniquely recovered from data.

We can now state our primary learning problem. We are given an observational dataset (all reviews) and multiple concept-conditional datasets (subsets of reviews filtered by sub-scores). The fundamental question is whether we can uniquely recover the underlying concepts from this data. This is the problem of identifiability.

Definition 1. *Given observational and concept-conditional datasets, we say the concepts* $\{C^1, \ldots, C^m\}$ *with linear maps* $\{A^1, \ldots, A^m\}$ *are* ***identifiable*** *if for any other set of parameters* $(\tilde{f}, \tilde{A}^e, \tilde{b}^e)$ *that generates the same observed data distributions, there exists an invertible linear*

map T, a shift $w \in \mathbb{R}^{d_z}$, permutation matrices P^e, and invertible diagonal matrices Λ^e such that for all data points $\mathbf{x}$ and concepts e:

$$\tilde{A}^e \tilde{f}^{-1}(\mathbf{x}) = \Lambda^e P^e A^e (f^{-1}(\mathbf{x}) + w), \tag{1}$$

and the concept parameters are related by:

$$\tilde{A}^e = P^e A^e T^{-1}, \quad \tilde{b}^e = \Lambda^e P^e (b^e - A^e w). \tag{2}$$

Identifiability, in this context, means that we can recover the structure of the human-aligned concepts up to a set of acceptable ambiguities. We can learn the linear subspaces corresponding to concepts like 'soundness' and 'clarity' (Equation 2) and, crucially, we can learn to evaluate any rebuttal on these conceptual axes (Equation 1). The ambiguities—permutation (P^e), scaling (Λ^e), and a global linear transformation of the latent space (T)—are unavoidable as the latent space is never directly observed. However, they do not impede our goal. Recovering these concept evaluation maps $A^e f^{-1}$ is precisely what allows us to dissect a rebuttal, understand which of its latent characteristics causally drive reviewer perceptions along specific criteria, and ultimately provide the concrete, actionable guidance promised in our abstract. The theoretical conditions for achieving this identifiability, as outlined in [1], rely on the diversity of the available concept-conditional distributions, a condition naturally met by the rich sub-score data in OpenReview.

A3.3 Identifiability of Causal Models: Theorem and Proof

Theorem 1. ***(Identifiability of Review Concepts)*** *Suppose we match the observations $\mathbf{x}_m$ across modalities (authors and reviewers), and the following conditions hold in the data-generating process:*

- *i* *(Information Preservation): The functions g_1 and g_2 are differentiable and invertible.*
- *ii* *(Primary Area Diversity): All entries of $v^\top B$ are non-zero, where $B_{i,j} = \frac{b_k^e}{\sigma^2}$ denotes the area–concept matrix.*
- *iii* *(Thought Reflection): The latent components in $\mathbf{z}_1$ are causal parents of $\mathbf{z}_2$, but not vice versa.*
- *iv* *(Distinctive Concept Alignment): There exists a set of linearly independent aligning vectors $\mathcal{C} = \{a_1, \ldots, a_n\}$ such that, for each concept C^e, the rows of the aligning matrix A^e lie in $\mathcal{C}$, i.e., $(A^e)^\top e_i \in \mathcal{C}$. Let S^e denote the indices of the subset of $\mathcal{C}$ that appear as rows of A^e. Every aligning vector in $\mathcal{C}$ appears in at least one area e (where an area corresponds to a concept-conditional distribution), that is,*

$$\bigcup_e S^e = [n].$$

Then the review concepts are identifiable as in Definition 1.

Discussion of Assumptions Assumption i requires that the latent space is recoverable from the observed data. Assumption ii further requires the presence of latent distribution shifts in the review concepts across different primary areas, ensuring variability in the underlying structure. Assumption iii reflects the natural process in which authors first read the reviews, then engage in reflection, and finally provide rebuttals. Finally, Assumption iv ensures that all concepts can be decomposed into a finite set of atomic components that remain distinct across primary areas, which is essential for separating and identifying them.

Proof Sketch We first recover the latent space from the reviews and author responses by applying the inverse generating functions together with the fixed causal direction between the author and review modules. The presence of latent distribution shifts in the review concepts across different primary areas then provides additional variation, which allows us to identify each concept by comparing the concept spaces across environments. In this way, the atomic concepts can be causally inferred.

Proof. We outline the argument showing that the review concepts are identifiable under Assumptions i–iv.

Step 1: Recovering the latent space. By Assumption i, both generating functions g_1, g_2 are differentiable and invertible. Hence, we can consistently map the observed responses and reviews $(\mathbf{x}_1, \mathbf{x}_2)$

back into their latent representations $(\mathbf{z}_1, \mathbf{z}_2) = (g_1^{-1}(\mathbf{x}_1), g_2^{-1}(\mathbf{x}_2))$. Assumption iii enforces that the review latents $\mathbf{z}_2$ are causal parents of the response latents $\mathbf{z}_1$ (reviews influence responses, but not vice versa). Thus the recovered latent space retains a fixed causal ordering, eliminating confounding symmetries.

Step 2: Variation across primary areas. For each primary area e, the observed distribution corresponds to a concept-conditional dataset with parameters (A^e, b^e). Define the *area–concept matrix* $B \in \mathbb{R}^{m \times n}$ by

$$B_{i,j} = \frac{b_k^e}{\sigma^2} \quad \text{if } a_j \text{ is active in area } e, \qquad B_{i,j} = 0 \text{ otherwise.}$$

Assumption ii requires that $v^\top B$ has no zero entries. This ensures that across areas there is genuine distributional shift in the valuations b^e, preventing degenerate alignments of concepts. Intuitively, if two concepts always co-occur with the same valuation, they cannot be separated; diversity prevents this.

Step 3: Identifying atomic concepts. By Assumption iv, there exists a set of linearly independent aligning vectors $\mathcal{C} = \{a_1, \ldots, a_n\}$ such that the rows of each A^e are subsets of $\mathcal{C}$. Let $S^e \subseteq [n]$ be the indices of the atoms appearing in area e. The union condition $\bigcup_e S^e = [n]$ guarantees that every atomic concept is observed in at least one area. From the quadratic form of the concept-conditional densities,

$$\ln p(\mathbf{z}) - \ln p^e(\mathbf{z}) \;=\; \sum_{i=1}^{n} \left(\tfrac{1}{2} M_{ei} \langle a_i, \mathbf{z} \rangle^2 - B_{ei} \langle a_i, \mathbf{z} \rangle \right) + c_e,$$

where M is the environment–concept incidence matrix and B is the area–concept valuation matrix, we can solve for the atomic directions a_i up to permutation and scaling. This step mirrors the identifiability argument in Definition 1.

Step 4: Resolving symmetries. Because the latent space is not directly observed, solutions are unique only up to permutation and linear reparametrization. However, the combination of (i) invertibility (i), (ii) causal ordering (iii), (iii) area diversity (ii), and (iv) complete coverage of atomic alignments (iv) ensures that these are the only remaining symmetries. Thus the review concepts can be identified uniquely up to linear equivalence.

Putting together these steps, we recover all review concepts and their valuations from the observed multimodal data, completing the proof of Theorem 1. □

A4 Implementation Framework

Based on the theoretical identifiability conditions, we now introduce our practical implementation for learning the latent causal representations. The core challenge is to infer the posterior distribution of the latent variables $\mathbf{z}$ given the observed data $\mathbf{x}$, i.e., $p(\mathbf{z}|\mathbf{x})$. Since the generative process $\mathbf{x} = f(\mathbf{z})$ is assumed to be complex and nonlinear, this posterior is intractable to compute directly. Therefore, we employ variational inference to approximate it.

A4.1 Variational Inference and Model Architecture

We adopt a parametric implementation based on a Variational Autoencoder (VAE) architecture. This framework consists of an encoder that approximates the posterior distribution and a decoder that reconstructs the data.

- **Encoder:** We introduce an approximate posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$, parameterized by an encoder network with parameters ϕ. This network takes the high-dimensional observed data X (e.g., text embeddings of reviews and rebuttals) and outputs the parameters of a diagonal Gaussian distribution for each latent variable: $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_\phi(\mathbf{x}), \text{diag}(\sigma_\phi^2(\mathbf{x})))$.
- **Decoder:** The decoder network $p_\theta(\mathbf{x}|\mathbf{z})$, parameterized by θ, takes a sample from the latent space and aims to reconstruct the original input X.
- **Latent Causal Structure:** The causal relationships between the latent variables $\mathbf{z}$ are modeled according to a linear structural equation model: $\mathbf{z} = A^T\mathbf{z} + E$, where A is a learnable weighted adjacency matrix representing the causal graph $\mathcal{G}_z$, and E are exogenous noise variables. This causal structure is incorporated into the model's prior, $p(\mathbf{z})$.

The model is trained by maximizing the Evidence Lower Bound (ELBO), which is a lower bound on the log-likelihood of the data, $\log p(\mathbf{x})$. The ELBO is defined as:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{3}$$

The first term is the reconstruction log-likelihood, ensuring the latents capture the data's salient features. The second term is a KL divergence that regularizes the approximate posterior to be close to the prior over the latent variables.

A4.2 Learning Objectives for Causal Concept Discovery

While the ELBO is a standard objective for learning representations, our goal requires additional constraints to ensure the learned latents are not only representative but also causally structured and aligned with human concepts. We introduce two additional loss terms to achieve this.

A4.2.1 Causal Sparsity Loss

To ensure the learned causal graph $\mathcal{G}_Z$ is identifiable and interpretable, we must encourage sparsity. A dense graph of latent relationships would be difficult to analyze and prone to overfitting. We therefore impose an L_1 penalty on the adjacency matrix A, which promotes a sparse graph by driving many of the potential causal connection weights to zero. The sparsity loss is defined as:

$$\mathcal{L}_{\text{sparsity}} = \|A\|_1 = \sum_{i,j} |A_{ij}| \tag{4}$$

A4.2.2 Weak Supervision Loss

A central component of our framework is the alignment of the abstract latent space with the concrete evaluation criteria used by human reviewers. We leverage the review sub-scores (e.g., for soundness, clarity, novelty) as a form of weak supervision. We designate a subset of the latent variables, $\mathbf{z}_{\text{sup}} \subseteq \mathbf{z}$, to correspond to these concepts and penalize the deviation from the ground-truth scores, y. For the numerical scores in OpenReview, we use the Mean Squared Error (MSE):

$$\mathcal{L}_{\text{supervision}} = \frac{1}{m} \sum_{k=1}^{m} (\mathbf{z}_{\text{sup}}^{(k)} - y^{(k)})^2 \tag{5}$$

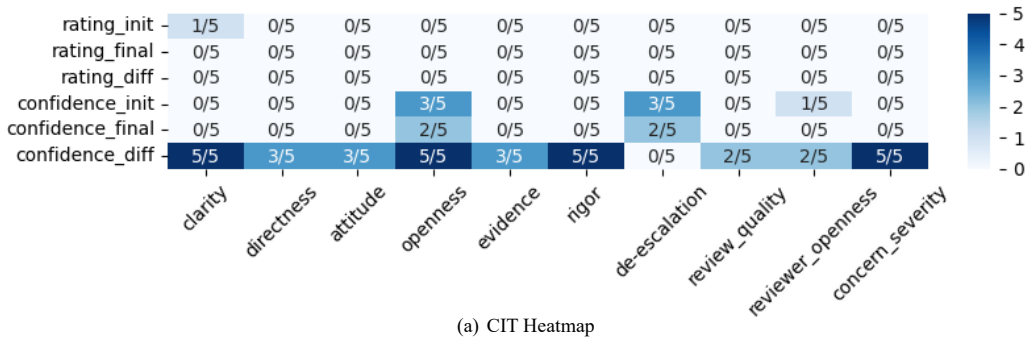
where m is the number of supervised concepts. This loss is crucial for grounding the representation and ensuring its practical utility.

A4.3 Final Objective Function

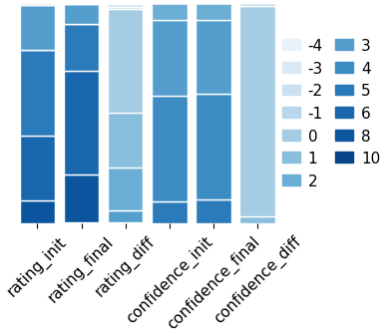
The model is trained end-to-end by minimizing a single, composite objective function that is a weighted sum of the three components described above. The final learning objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ELBO}} + \lambda_1 \mathcal{L}_{\text{sparsity}} + \lambda_2 \mathcal{L}_{\text{supervision}} \tag{6}$$

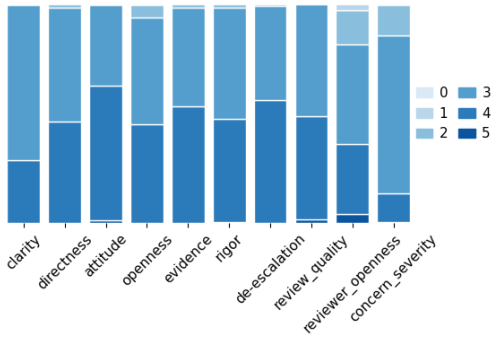
where λ_1 and λ_2 are hyperparameters that control the relative importance of enforcing causal sparsity and concept alignment against the primary objective of data reconstruction.



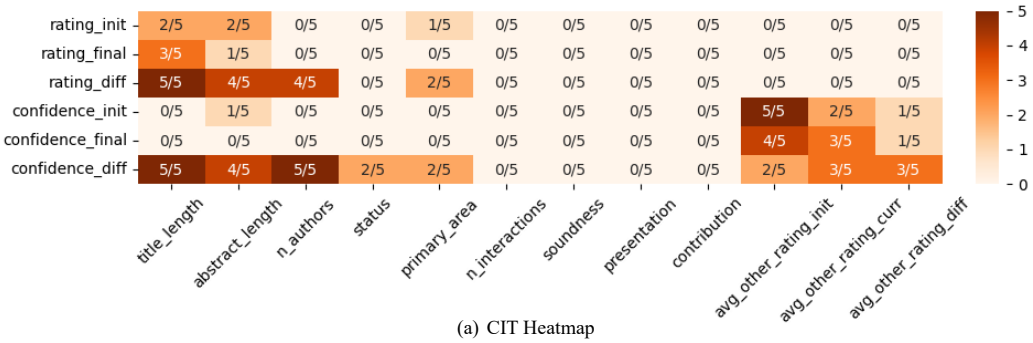
(a) CIT Heatmap



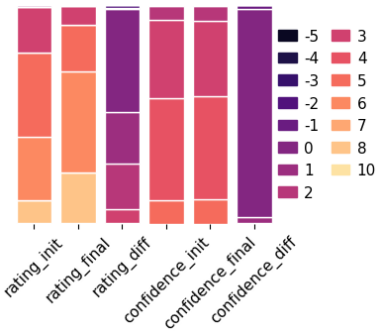
(b) Distribution of Rating and Confidence Scores.



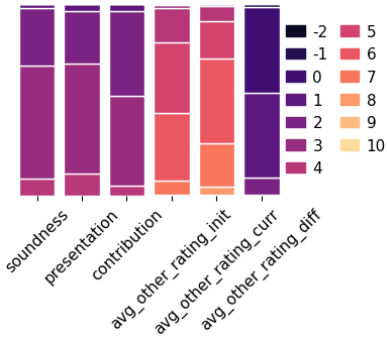
(c) Distribution of Labeled Scores.



(a) CIT Heatmap



(b) Distribution of Rating and Confidence Scores.



(c) Distribution of Labeled Scores.

Figure A1: Illustration of CIT results and data distribution for the 10% subset. (a) shows the aggregated CIT results. Each cell indicates how many tests failed to reject the null hypothesis of independence at $\alpha = 0.05$. A score of 0/5 (strong evidence of dependence) means all tests found significant dependence, while a score of 5/5 (strong evidence of independence) means that none did. (b) and (c) show the distribution of rating and confidence scores and other extracted/annotated sublabels, respectively.