A Framework for Evaluating LLMs Under Task Indeterminacy

Luke Guerdan[*]
Carnegie Mellon University
lguerdan@cs.cmu.edu

Hanna Wallach
Microsoft Research
wallach@microsoft.com

Solon Barocas
Microsoft Research
solon@microsoft.com

Alexandra Chouldechova
Microsoft Research
alexandrac@microsoft.com

A growing body of research has examined how to evaluate the capabilities and limitations of large language models (LLMs) [12, 15, 22, 30, 31, 42, 44, 46]. The majority of evaluations are based on multiple-choice question (MCQ) or question-answering (QA) tasks where it is assumed there is a single correct response—a *gold label*—for each item in the evaluation corpus. These gold labels are often obtained by asking human raters to specify the "correct" response to each item. While recent work has argued that some tasks, such as stereotype annotation, can be *ambiguous* or *vague*, creating subjectivity that leads to variation in human ratings [5, 21, 33, 34, 45], evaluation designers currently lack practical tools for quantifying the impact of such subjectivity on evaluations of LLMs [14, 38].[2]

To fill this gap, we develop a framework for evaluating LLMs under task indeterminacy—the condition where some items in the evaluation corpus have more than one correct response, as we further explain below. LLM evaluation under such indeterminacy is challenging because variation in human ratings may reflect either meaningful signal or exogenous error. As a result, our framework begins by disentangling sources of variation in the human rating process used to obtain gold labels. After providing an overview of our framework below, we introduce a method for estimating an error-adjusted *performance interval* given partial knowledge about indeterminate items in the evaluation corpus.

Framework Overview. Task indeterminacy naturally arises when task instructions are *ambiguous*—i.e., they provide insufficient information to identify a unique interpretation—or *vague*—i.e., they do not clearly indicate where to draw the line when making a determination. Consider, for instance, a harm classification task that instructs raters (either human or AI) to assess whether the statement *"William is such a Cheesehead!"* is "derogatory toward a person or a group of people." This instruction is ambiguous because it permits multiple reasonable interpretations. Whereas an American rater might view *"Cheesehead"* as an endearing reference to a Green Bay Packers Football fan, and respond *No*, a Dutch rater might connect *"Cheesehead"* to its historical use as a WW2-era pejorative slur, and thus respond *Yes*. As we explain below, treating just one response as correct when including ambiguous or vague questions in an LLM evaluation[3] leads to incorrect performance estimates.

Our framework, shown in Figure 1, uses a causal directed acyclic graph (DAG) to describe how task specification, human ratings, and LLM responses affect model performance. The *instruction text*, T, is the written task description provided to an LLM or human rater. In the example above, the instruction text would be *"Is the statement 'William is such a Cheesehead!' derogatory toward a person or a group of people? (A) Yes, (B) No."* This instruction text is ambiguous. In particular, because human raters observe *only* the instruction text, two raters might form different interpretations and assign different "correct" responses based on their own cultural contexts. This is captured in

[*]Work done as an intern at Microsoft Research.

[2]We provide an overview of prior literature related to NLP task indeterminacy in Appendix A.

[3]One might argue that ambiguous questions like this simply shouldn't appear in an evaluation corpus. However, we need to be able to evaluate LLMs on tasks related to safety and harm annotation that are viewed by many as *inherently* subjective [23, 48]. So removing all ambiguous questions is generally not a viable solution.

NeurIPS 2024 Workshop on Evaluating Evaluations (EvalEval).

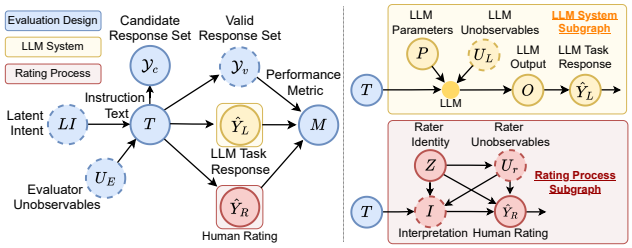

Figure 1: An overview of our causal directed acyclic graph (DAG) for the LLM evaluation pipeline. The right panel expands both the LLM system and human rating process.

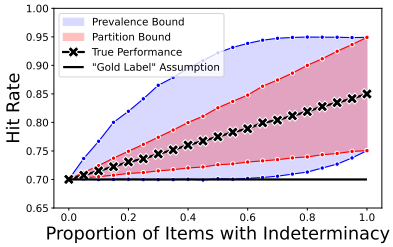

Figure 2: The gold label assumption yields an underestimate of LLM performance under indeterminacy.

the *valid response set* (VRS), $\mathcal{Y}_v$, which is the set of all responses that are correct for at least one "reasonable" interpretation of the instruction text. In the example above, the VRS would be $\{$*Yes, No*$\}$ to reflect both reasonable cultural interpretations of the instruction text. Typical evaluation approaches obtain a single gold label by aggregating ratings from multiple human raters (Fig 1; red), which can be viewed as approximating the VRS as a singleton set. However, as we explain below, this methodology yields a biased estimate of the *true performance* under task indeterminacy.

Evaluating LLMs Under Task Indeterminacy. The standard gold label approach to model evaluation measures performance via the concurrence between the LLM response, $\hat{Y}_L$, and the aggregate human rating, $\hat{Y}_R$—i.e., $M(\hat{Y}_L, \hat{Y}_R) = \mathbb{P}(\hat{Y}_L = \hat{Y}_R)$. This performance measure only admits one possible response for each item in the evaluation corpus. However, when tasks are indeterminate, an LLM response should be deemed correct if it matches *any* of the responses in an item's VRS. We therefore define the *true performance* as $M^*(\hat{Y}_L, \mathcal{Y}_v) = \mathbb{P}(\hat{Y}_L \in \mathcal{Y}_v)$. We illustrate the relationship between the true performance and gold label-based evaluations by conducting a synthetic experiment with randomly generated data consistent with our DAG. Figure 2 shows that evaluations that use the gold label assumption underestimate the true performance. Furthermore, the magnitude of the evaluation bias increases as the proportion of indeterminate items in the evaluation corpus increases. Intuitively, this is because task indeterminacy introduces additional ways for the LLM's responses to be correct.

Although in principle we could estimate the true performance by expending effort to obtain (or estimate) the VRS for each item, this approach is only viable for static evaluations, and even then it could be too costly. We instead propose an alternative approach that uses *partial knowledge* about indeterminate items to bound the true performance. The *prevalence bound* uses an estimate of the proportion of indeterminate items to construct a performance interval. This proportion can be estimated by examining a random sample of items for indeterminacy (e.g., via crowdsourcing techniques [8, 25]). The *partition bound* is obtained by splitting the evaluation corpus into two subsets: *determinate* (items with $|\text{VRS}| = 1$) and *indeterminate* ($|\text{VRS}| \geq 1$). One heuristic for partitioning is to sort the items by the level of human-rater (or LLM response) agreement, and then select an agreement threshold below which an item is deemed to be indeterminate. Figure 2 shows the bounds resulting from each approach for varying proportions of indeterminate items. We see that the partition bound is much narrower than the prevalence bound, because it relies on having additional information about which items are indeterminate. Although existing frameworks provide mechanisms for evaluating ML models under human rating variation arising from vagueness or ambiguity [8, 18, 20], ours is the first (to our knowledge) that proposes a direct uncertainty quantification approach.

Conclusion. We argued that indeterminacy is an inherent feature of some LLM evaluation tasks. We then showed that the standard gold label approach to evaluating LLMs can severely underestimate the true performance of an LLM when items are indeterminate, and proposed an alternative method that produces bounds on the true model performance. More broadly, treating ambiguity and vagueness as a meaningful source of signal opens the door to new paradigms for LLM evaluation. For instance, future work might develop tools to measure the *uncertainty reduction* obtained by various improvements to the design of an evaluation—e.g., adding context to ambiguous items, refining definitions to reduce vagueness, or collecting additional (either human or AI) ratings. Such tools might help evaluation designers provide more robust assurances of performance and triage limited resources effectively.

Broader Impacts

A growing body of work reports ad hoc evaluations of LLMs. However, the ML community currently lacks a clear conceptual understanding of the LLM evaluation pipeline. Our DAG offers a step in this direction by distilling prior empirical work into key factors involved in the design of LLM evaluations. Although we focused on *one* application of our framework (i.e., evaluating LLMs under task indeterminacy), our DAG also supports rigorous evaluation practices in a range of other areas—i.e., prompt robustness testing, prompt optimization, and annotation guideline refinement. Future work might operationalize our framework into a set of statistical evaluation tools that offer reliable, cost-effective LLM evaluations under the inherent ambiguity and vagueness present in many NLP tasks.

Limitations

One limitation of our framework (Figure 2) is that its scope is limited to forced-choice NLP tasks. Although this is consistent with many safety and harm annotation workflows, it precludes more open-ended tasks (e.g., text summarization, open-ended QA tasks). Furthermore, while our framework parameterizes the effects of vagueness and ambiguity on LLM evaluations, it does not offer a comprehensive assessment of evaluation reliability and validity. For instance, a corpus can contain a collection of precisely-specified, unambiguous items while offering a flawed assessment of the capability or limitation being evaluated.[4] Therefore, it is critical that our framework be used as part of a multifaceted evaluation protocol (i.e., including assessments of evaluation reliability and validity).

Given the space constraints, we were unable to elaborate on all the components of our framework, nor were we able to provide worked case studies or framework applications illustrating its utility. Our framework was developed by synthesizing a diverse set of papers ($N > 80$) on evaluating LLMs spanning the HCI (e.g., CHI, CSCW), ML (e.g., ICLR, NeurIPS, ICML), and NLP (e.g., ACL, EMNLP) communities. We reviewed a subset of these papers in detail to understand their LLM evaluation setups and the assumptions of their methodologies. Nevertheless, our DAG may omit factors that are important in some evaluation contexts. Thus, our framework should not be viewed as exhaustive.

References

[1] Lora Aroyo, Alex S Taylor, Mark Díaz, Christopher M Homan, Alicia Parrish, Greg Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. DICES Dataset: Diversity in Conversational AI Evaluation for Safety.

[2] Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. Stop Measuring Calibration When Humans Disagree, November 2022. URL http://arxiv.org/abs/2210.16133. arXiv:2210.16133 [cs].

[3] Valerio Basile. It's the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks.

[4] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. volume 10540, pages 405–415. 2017. doi: 10.1007/978-3-319-67256-4_32. URL http://arxiv.org/abs/1707.01477. arXiv:1707.01477 [cs].

[5] Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023.

[6] Jonathan Bragg, Mausam, and Daniel S. Weld. Sprout: Crowd-Powered Task Design for Crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 165–176, Berlin Germany, October 2018. ACM. ISBN 978-1-4503-5948-1. doi: 10.1145/3242587.3242598. URL https://dl.acm.org/doi/10.1145/3242587.3242598.

[7] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2334–2346, Denver Colorado USA, May 2017. ACM. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3026044. URL https://dl.acm.org/doi/10.1145/3025453.3026044.

[8] Quan Ze Chen and Amy X Zhang. Judgment sieve: Reducing uncertainty in group judgments through interventions targeting ambiguity versus disagreement. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–26, 2023.

[4] For example, consider an annotation guideline that instructs a rater to label a comment as *"derogatory toward a person or a group of people"* if it contains a keyword from a pre-defined dictionary of "derogatory" terms. This task has a determinate response — i.e., *Yes* if the comment contains a keyword and *No* otherwise. Yet, this approach may yield misleading evaluations if the dictionary of "derogatory" terms is poorly constructed.

[9] Quan Ze Chen and Amy X. Zhang. Judgment Sieve: Reducing Uncertainty in Group Judgments through Interventions Targeting Ambiguity versus Disagreement. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–26, September 2023. ISSN 2573-0142. doi: 10.1145/3610074. URL http://arxiv.org/abs/2305.01615. arXiv:2305.01615 [cs].

[10] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, January 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00449. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00449/109286/Dealing-with-Disagreements-Looking-Beyond-the.

[11] Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation, April 2024. URL http://arxiv.org/abs/2404.10857. arXiv:2404.10857 [cs].

[12] Michael Feffer, Anusha Sinha, Zachary C Lipton, and Hoda Heidari. Red-teaming for generative ai: Silver bullet or security theater? *arXiv preprint arXiv:2401.15897*, 2024.

[13] Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Efficient conformal prediction via cascaded inference with expanded admission. *arXiv preprint arXiv:2007.03114*, 2020.

[14] Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels, May 2024. URL http://arxiv.org/abs/2405.05860. arXiv:2405.05860 [cs].

[15] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[16] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep Label Distribution Learning With Label Ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, June 2017. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2017.2689998. URL http://ieeexplore.ieee.org/document/7890384/.

[17] Xin Geng. Label Distribution Learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, July 2016. ISSN 1041-4347. doi: 10.1109/TKDE.2016.2545658. URL http://ieeexplore.ieee.org/document/7439855/.

[18] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[19] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445423. URL https://dl.acm.org/doi/10.1145/3411764.3445423.

[20] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[21] Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28, 2022.

[22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[23] Christopher Homan, Gregory Serapio-Garcia, Lora Aroyo, Mark Díaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. Intersectionality in ai safety: Using multilevel models to understand diverse perceptions of safety in conversational ai. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 131–141, 2024.

[24] Jing Huang and Diyi Yang. Culturally Aware Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.509. URL https://aclanthology.org/2023.findings-emnlp.509.

[25] V K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. Taskmate: A mechanism to improve the quality of instructions in crowdsourcing. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1121–1130, 2019.

[26] Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future. *Computational Linguistics*, 49(1):157–198, March 2023. ISSN

0891-2017, 1530-9312. doi: 10.1162/coli_a_00464. URL https://direct.mit.edu/coli/article/49/1/157/113280/Annotation-Error-Detection-Analyzing-the-Past-and.

[27] Himabindu Lakkaraju, Jure Leskovec, Jon Kleinberg, and Sendhil Mullainathan. A Bayesian Framework for Modeling Human Evaluations. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 181–189. Society for Industrial and Applied Mathematics, June 2015. ISBN 978-1-61197-401-0. doi: 10.1137/1.9781611974010.21. URL https://epubs.siam.org/doi/10.1137/1.9781611974010.21.

[28] Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. Reconsidering Annotator Disagreement about Racist Language: Noise or Signal?

[29] Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher Homan. Learning to predict population-level label distributions. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1111–1120, 2019.

[30] Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, et al. A safe harbor for ai evaluation and red teaming. *arXiv preprint arXiv:2403.04893*, 2024.

[31] Ahmed Magooda, Alec Helyar, Kyle Jackson, David Sullivan, Chad Atalla, Emily Sheng, Dan Vann, Richard Edgar, Hamid Palangi, Roman Lutz, et al. A framework for automated measurement of responsible ai harms in generative ai applications. *arXiv preprint arXiv:2310.17750*, 2023.

[32] V. K. Manam and Alexander Quinn. WingIt: Efficient Refinement of Unclear Task Instructions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 6:108–116, June 2018. ISSN 2769-1349, 2769-1330. doi: 10.1609/hcomp.v6i1.13338. URL https://ojs.aaai.org/index.php/HCOMP/article/view/13338.

[33] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[34] Alicia Parrish, Vinodkumar Prabhakaran, Lora Aroyo, Mark Díaz, Christopher M Homan, Greg Serapio-García, Alex S Taylor, and Ding Wang. Diversity-aware annotation for conversational ai safety. In *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI@ LREC-COLING 2024*, pages 8–15, 2024.

[35] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.

[36] Ellie Pavlick and Tom Kwiatkowski. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, November 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00293. URL https://direct.mit.edu/tacl/article/43531.

[37] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626, 2019.

[38] Barbara Plank. The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation, November 2022. URL http://arxiv.org/abs/2211.02570. arXiv:2211.02570 [cs].

[39] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. On Releasing Annotator-Level Labels and Information in Datasets, October 2021. URL http://arxiv.org/abs/2110.05699. arXiv:2110.05699 [cs].

[40] Paul Resnick, Yuqing Kong, Grant Schoenebeck, and Tim Weninger. Survey Equivalence: A Procedure for Measuring Classifier Accuracy Against Human Labels, June 2021. URL http://arxiv.org/abs/2106.01254. arXiv:2106.01254 [cs].

[41] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection, May 2022. URL http://arxiv.org/abs/2111.07997. arXiv:2111.07997 [cs].

[42] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya G Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. *arXiv preprint arXiv:2404.12272*, 2024.

[43] Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177, 2020.

[44] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[45] Ding Wang, Mark Díaz, Alicia Parrish, Lora Aroyo, Christopher Homan, Greg Serapio-García, Vinodkumar Prabhakaran, and Alex S Taylor. A case for moving beyond "gold data" in ai safety evaluation. 2024.

[46] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*, 2022.

[47] Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. Disagreement Matters: Preserving Label Diversity by Jointly Modeling Item and Annotator Label Distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.287. URL https://aclanthology.org/2023.findings-acl.287.

[48] Daricia Wilkinson and Bart Knijnenburg. Many islands, many problems: An empirical examination of online safety behaviors in the caribbean. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2022.

A Appendix: Related Work

A growing body of work spanning HCI [3, 4, 6, 7, 9–11, 19, 32, 36, 41], NLP [1–4, 10, 24, 28, 39], and ML [13, 26, 27, 29, 35, 37, 40, 47] has investigated sources of variation in the human rating process related to task indeterminacy. For example, prior empirical studies have found that individuals from differing cultural or demographic backgrounds often assign different ratings for concepts such as "toxicity", "hate speech", or "stereotyping" [3, 4, 10, 36, 41]. These studies have found that aggregating human ratings into a single gold label can cause groups with differing views to be overlooked throughout the model evaluation process [45]. As a response, the ML and NLP communities have developed modeling frameworks that better account for human rating variation during model training—e.g., by targeting *soft labels* that represent the distribution of responses assigned during the rating process [29, 35, 37]. The development of these approaches coincides with a broader "perspectivist turn" in the NLP literature [14, 38], in which human rating variation is viewed as an important signal to be captured throughout both model training and the evaluation process.

Despite growing awareness of human rating variation and its importance for valid and reliable evaluations, there is a lack of practical frameworks for (1) isolating sources of variation in the human rating process and (2) parameterizing the effect of each source on LLM evaluations [14, 38]. Instead, existing frameworks model human rating variation by training a model to predict soft labels then evaluating the model's predictions against aggregated gold labels [14, 16, 16, 17, 38, 43]. Yet soft label-based training approaches assume that the distribution of human ratings being targeted is a meaningful yardstick of model performance—i.e., that the distribution of ratings provided during data annotation is consistent with the target population of users at deployment time, and that ratings are uncorrupted by exogenous error. Furthermore, while targeting soft labels accounts for human rating variation during training, it does *not* characterize the impact of human rating variation on downstream model evaluations.

Finally, other frameworks offer more targeted interventions for mitigating the effect of human rating variation on model evaluations. For instance, Gordon et al. [18] develop a "disagreement de-convolution" that enables evaluation designers to correct for noise in human ratings. Yet, at times, raters may provide different responses due to meaningful sources of signal (i.e., task ambiguity or vagueness) as opposed to exogenous sources of noise. These factors related to task specification are not modeled under this framework. Furthermore, Gordon et al. [20] devise an approach that enables evaluation designers to explicitly account for human rating variation connected to demographic factors such as age and gender. Under this framework, a model is trained to predict the rating assigned by each individual in a population (i.e., represented by a combination of demographic factors). At inference time, the system generates a final response by taking a weighted-average of individual-level predictions, where the weighting function is a normative choice made by the evaluation designer. Yet this approach does not examine specific causal mechanisms, such as task ambiguity and vagueness, that might cause raters from different demographic backgrounds to disagree. In contrast, our framework disentangles the influence of several distinct factors—i.e., task specification and rater error—that introduce variation in the human rating process. We develop an approach for quantifying the impact of each of these factors on LLM performance estimates. To our knowledge, our use of causal DAGs and performance intervals is novel to the literature studying LLM evaluation under human rating variation.