Enhancing Robustness of Foundation Model Representations under Provenance-related Distribution Shifts

Xiruo Ding[1*] **Zhecheng Sheng**[2] **Brian Hur**[1] **Feng Chen**[1]
Serguei Pakhomov[2] **Trevor Cohen**[1]
[1]University of Washington [2]University of Minnesota

Abstract

Foundation models are a current focus of attention in both industry and academia. While they have shown their capabilities in a variety of tasks, in-depth research is required to determine their robustness to distribution shift when used as a basis for supervised machine learning. This is especially important in the context of clinical data, with particular limitations related to data accessibility, lack of pretraining materials, and limited availability of high-quality annotations. In this work, we examine the stability of models based on representations from foundation models under distribution shift. We focus on *confounding by provenance*, a form of distribution shift that emerges in the context of multi-institutional datasets when there are differences in source-specific language use and class distributions. Using a sampling strategy that synthetically induces varying degrees of distribution shift, we evaluate the extent to which representations from foundation models result in predictions that are inherently robust to confounding by provenance. Additionally, we examine the effectiveness of a straightforward confounding adjustment method inspired by Pearl's conception of backdoor adjustment. Results indicate that while foundation models do show some out-of-the-box robustness to confounding-by-provenance related distribution shifts, this can be considerably improved through adjustment. These findings suggest a need for deliberate adjustment of predictive models using representations from foundation models in the context of source-specific distributional differences.

1 Introduction

Machine learning methods have been widely applied in biomedical and clinical research. Applications have included those at the molecular level, such as predicting protein structure [1, 2] and drug-drug interactions [3]; and those at the individual level, such as electric health record (EHR) phenotyping [4] and supporting participant enrollment for clinical trials [5]. In the field of natural language processing (NLP), the development of deep learning methods and techniques - notably the transformer architecture introduced by Vaswani et al. [6] and subsequent encoder-only (BERT [7]) and decoder-only models (GPT [8]) - has advanced performance across many NLP tasks, including those in biomedicine [9]. Most recently, generative transformer architectures pre-trained on large amounts unlabeled text, commonly referred to as Large Language Models (LLMs), have demonstrated impressive performance on many NLP tasks [10, 11]. LLMs' capabilities have been demonstrated using not only traditional natural language processing (NLP) benchmarks, but also in tests of human capabilities such as the SAT and LSAT [12]. Foundation models are models pretrained on vast amounts of data and can be adapted for multiple downstream tasks [2]. While model weights have

*xiruod@uw.edu

[2]`https://crfm.stanford.edu/`

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

been released for some foundation models, such as Llama [13] (and its successor Llama-2 [14]) and Bloom [15], end-to-end fine-tuning of these models requires computational resources beyond those in many academic and clinical settings.

Contrary to the large amount of pretraining materials collected from a variety of sources for LLMs, in biomedical research high quality annotated datasets are often very limited in size and diversity [16]. Consequently, researchers may choose to integrate data from multiple institutions. This practice can increase dataset size but also introduces a potential bias when both language use and class distribution differ across these institutions. We refer to this form of distribution shift as *confounding by provenance* [17]. The main concern is that of site-specific label distribution shift from training to testing/deployment time. For example, if one institution has a much higher proportion of positive examples at training time, but a much lower proportion at test time, the model may make erroneous positive predictions based on language use at this institution that is unrelated to the outcome of interest. Representations derived from LLMs encode linguistic information from outside the context of a labeled training set, and it is possible that this information may confer a degree of robustness to confounding by provenance, making a resulting model less sensitive to institution-specific linguistic differences. The current work is motivated by a desire to assess the extent of this robustness, if it is indeed conferred.

In this work, we first propose an evaluation framework for confounding by provenance. We use a BERT [7] variant (Sentence-BERT [18]) and Llama [13, 14] models in our experiments, two widely-used and publicly-available foundation models that exemplify the encoder-only and decoder-only approaches, respectively. We test their stability under different degrees of distribution shift within the framework, across a range of synthetically-induced shifts in provenance-specific class distribution. To preserve computational resources, we extract the contextual embeddings generated from these foundation models and test them under a regression framework, with and without a simple adjustment. This procedure involves extending a method of confounding adjustment originally developed for discrete representations [19], to representations from foundation models, and we assess its utility as a means to enhance their robustness to confounding shift.

2 Preliminaries and Methods

Evaluation Framework In this work, we focus on one specific form of distribution shift, *confounding shift* [19], where label distributions among subpopulations differ in the training and testing set for a text classification problems: $P_{train}(Y|Z) \neq P_{test}(Y|Z)$, where Y is the label and Z is the provenance variable. Note that this problem formulation does not include the distribution of predictor variables, X, which in our case is derived from language.

We build upon the approach for synthetic injection of confounding shift [19, 20], to develop an evaluation framework for binary classification with two subpopulations, assuming $Y \in \{0, 1\}$ and $Z \in \{0, 1\}$. The following parameters to construct a testing scenario are set:

$$P_{train}(y = 1|z = 0) \tag{1}$$

$$P_{train}(y = 1|z = 1) \tag{2}$$

$$P_{train}(z = 1) = P_{test}(z = 1) = C_z \tag{3}$$

$$P_{train}(y = 1) = P_{test}(y = 1) = C_y \tag{4}$$

where Equation (3) aims to eliminate a potential confounding factor where the proportion of training examples (irrespective of their class label) from each source is different at training and testing time, and Equation (4) is implicitly enforced to negate effects of different background positive rates in the train and test sets. The objective of these constraints is to focus on shifts related to provenance. In contrast to the work of Landeiro and Culotta [19, 20] where a relative difference of subtraction was used, we introduce two auxiliary variables for measuring differences in site-specific class distribution:

$$\alpha_{train} = \frac{P_{train}(y = 1|z = 1)}{P_{train}(y = 1|z = 0)},\ \alpha_{test} = \frac{P_{test}(y = 1|z = 1)}{P_{test}(y = 1|z = 0)} \tag{5}$$

During evaluation, we specify desired ranges for variables (1), (2), (3), and α_{test}. All combinations of these parameters are applied to govern selection of corresponding samples to construct multiple train/test set combinations with different degrees of confounding shift. Ultimately, the goal is to examine a model's robustness to these different degrees of distribution shift (measured by the

difference between α_{train} and α_{test}). To quantify robustness (or model stability), α_{test} is first log-transformed and a linear regression line is fit against a target evaluation metric (AUPRC value, Area Under the Precision-Recall Curve, in our case), inspired by Taori et al. [21]. This coefficient measures the slope of a line that relates changes in the performance metric of interest to changes in α_{test}. The lower the absolute value of the fitted coefficient, the more robust a model is to confounding shift, with a value of zero indicating equivalent performance irrespective of this shift.

Backdoor Adjustment Originally proposed by Pearl [22] (*Causality*, Equation 3.19), backdoor adjustment is a technique to make adjustments on predictions when confounding variable (z) exists:

$$P(y|x) = \sum_{z} P(y|x, z)P(z) \tag{6}$$

A similar approach was developed by Landeiro and Culotta [19] for text classification in the presence of confounding bias. Specifically, a logistic regression model is fit to estimate $P(y|X, z = c)$:

$$logit(y_c) = \beta_0 + \boldsymbol{\beta_1 X} + \boldsymbol{\beta_2 z_c} + \epsilon \tag{7}$$

where $\boldsymbol{z_c}$ is a one-hot matrix where the membership in a specific Z class is represented by a value v, a hyperparameter that controls the degree of adjustmnet [23, 20]. Estimates of $P(y_c|x, z_c)$ from Formula 7 can then be used in Formula 6 to get an adjusted estimate for $P(y|x)$.

Embeddings We use a simple distributional language model, **Binary Unigram**, as our baseline for representing natural language, as in previous work [19] for point of comparison. Additionally, we use the **Sentence-BERT** model, which has been optimized for generating semantically meaningful sentence embeddings [18], as a moderately-sized foundation model.

Llama Embeddings Three versions of the model are public, marked by the number of parameters as Llama 2-7B, Llama 2-13B, and Llama 2-70B. Our extension to Llama model beyond its generative ability is to extract embeddings, and then apply a classification head to them. We use the average embedding across all tokens to represent a unit of text. To investigate a potential relationship between robustness and model size, we derive representations from all three versions of Llama 2.

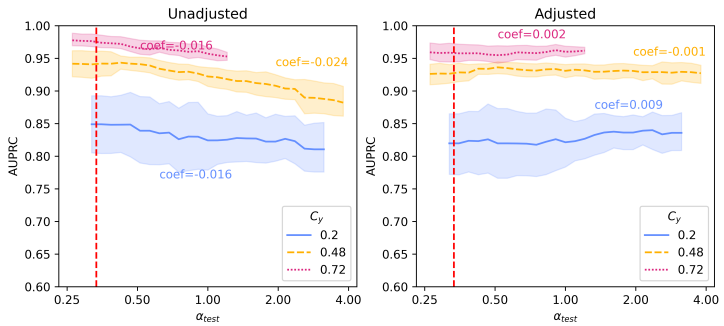
For consistency with prior work [19, 20], we use logistic regression as the classifier for all models, with and without applying backdoor adjustment.

3 Experiments

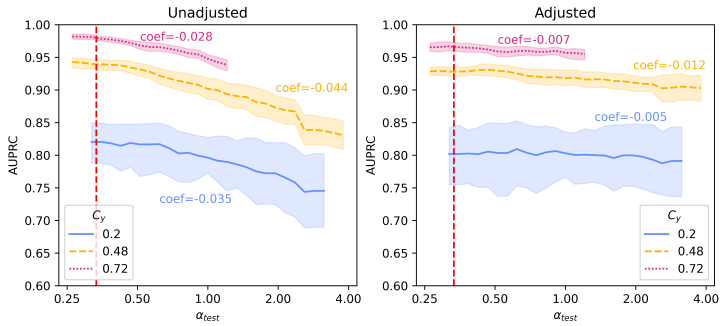
Datasets SHAC is a dataset of electronic health record notes annotated for social determinants of health (SDoH) [24, 25]. The notes were collected from two different sources: clinical notes of chronic pain patients from the University of Washington Medical Center, and discharge notes of intensive care unit patients from MIMIC-III [26]. Our goal for this work is to classify patient's social history sections for whether they show any sign of drug abuse. For Hate Speech Detection, we selected two publicly available hate speech detection datasets: (1) a dataset of hate speech entries generated through perturbation on publicly available datasets [27, 28]; (2) a dataset collected from a random set of posts extracted from a white supremacist forum, and labeled for hate speech [29].

Simulations Using the evaluation framework, we simulated different degrees of provenance-related distribution shift. Specifically, we set value ranges for $P_{train}(y = 1|z = 0), P_{train}(y = 1|z = 1), P_{train}(z = 1), \alpha_{train}$. For each simulated setting, an experiment was repeated five times for robustness of results. In addition, we varied C_y, the overall prevalance of the positive class irrespective of provenance (more positive examples lead to better performance). Three C_y values were selected to represent different background prevalence rates: 0.2, 0.48, and 0.72 for evaluations in the SHAC dataset, and 0.36, 0.44, 0.52 for Hate Speech Detection dataset. For further details of simulation configurations, refer to Appendix A.

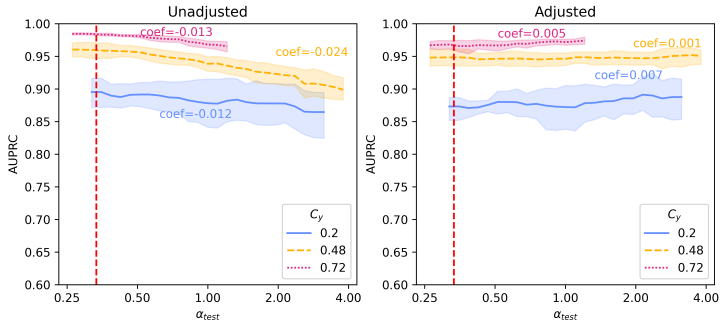
Results and Discussion Figure 1 shows results of the unadjusted (left column) and adjusted (right column) logistic regression models using Binary Unigram, Sentence-BERT, and Llama 2 embeddings when $\alpha_{train} = 0.33$. This α_{train} value indicates that we are shifting toward greater test set representation of positive examples from the second site ($z = 1$) as we move to the right



(a) Binary Unigram



(b) Sentence-BERT



(c) Llama 2-7B Average

Figure 1: Logistic regression and Backdoor Adjustment on SHAC. (a) Binary Unigrams; (b) Embeddings generated from Sentence-BERT; (c) Embeddings generated from Llama 2-7B, then averaged. Red vertical line indicates $\alpha_{train} = 0.33$, in which setting the model is trained.

along the x-axis from the dashed red line. An approximation of a mirror image of this figure can also be generated by setting $\alpha_{train} = 3$ (more positive examples from the first site at training time) and moving to the left to decrease representation of positive examples from this site ($z = 1$) (Figure A5 and Figure A6). With both Sentence-BERT and Llama, the unadjusted regression models (leftmost panels) decrease in performance as the provenance-specific class distribution moves toward over-representation of the second site ($z = 1$) relative to α_{train}, resulting in moderate negative slopes. However, the absolute values of the coefficients of regression lines fit to the performance when using Llama 2 embeddings is lower than those from Sentence-BERT (e.g., 0.013 vs 0.028 for $C_y = 0.72$, 0.024 vs 0.044 for $C_y = 0.48$), suggesting that the Llama2 embeddings may be innately more robust to confounding shift than those from Sentence-BERT. However, in comparison with baseline binary unigrams, the absolute values of the coefficients from Llama 2-7B embeddings are only slightly smaller, e.g., 0.012 vs 0.016 for $C_y = 0.2$, 0.013 vs 0.016 for $C_y = 0.72$, suggesting that representations from foundation models have no innate advantage in robustness to confounding shift as compared with unigram representations. Results from the Hate Speech Detection dataset

(Figure A4) show a similar trend, with the difference that those generated using foundation models have slightly smaller absolute coefficients than the unigram baseline.

With Backdoor Adjustment applied (rightmost panels), the line fit to performance is flattened, and the absolute values of the coefficients decrease (e.g. dropping from 0.035 to 0.005 for $C_y = 0.2$ using Sentence-BERT embeddings), demonstrating a marked increase in robustness to confounding shift. This technique also increases the robustness of models trained on Llama 2 embeddings (shown in Figure 1c right pane), though the difference between models with and without adjustment is relatively small. When comparing with baseline binary unigrams, absolute coefficients on results from both Sentence-BERT and Llama 2 embeddings do not drop, and increase in some scenarios. Results on the Hate Speech Detection dataset (Figure A4) show a similar trend. However, with respect to performance (rather than robustness) models using Sentence-BERT and Llama 2 embeddings both comfortably outperform baseline binary unigrams, measured directly by AUPRC (the height of AUPRC lines in general).

4 Conclusion

In this work, we investigate robustness of foundation models, from Sentence-BERT to different versions of Llama 2, for the task of text classification under a framework for provenance-related confounding shift. From empirical experiments on one biomedical and one general domain dataset, embeddings from foundation models show some out-of-the-box robustness to confounding shifts, demonstrated by their smaller absolute coefficients when no adjustment is made. Additionally, they increased baseline performances on one of our datasets. Employing Backdoor Adjustment within a logistic regression framework further enhances this robustness for both foundation models and baseline binary unigram models, although the latter gain more from this adjustment. One assumption is that the contextual embeddings, generated from both Sentence-BERT and Llama, could be so sensitive that robustness cannot be correctly modeled (or separated) as signals using a simple logistic regression classifier. In future work we will explore different classifiers, alternative methods for adjusting foundation models, to better enhance their robustness to distribution shifts of this nature.

Acknowledgments and Disclosure of Funding

This work was supported by U.S. National Library of Medicine Grant (R01LM014056).

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[2] Mehmet Akdel, Douglas EV Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L Good, Roman A Laskowski, Gabriele Pozzati, et al. A structural biology community assessment of alphafold2 applications. *Nature Structural & Molecular Biology*, 29(11):1056–1067, 2022.

[3] Feixiong Cheng and Zhongming Zhao. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association*, 21(e2):e278–e286, 2014.

[4] George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

[5] Xingyao Zhang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Deepenroll: patient-trial matching with deep embedding and entailment prediction. In *Proceedings of the web conference 2020*, pages 1029–1037, 2020.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[9] Shuang Zhang, Rui Fan, Yuti Liu, Shuang Chen, Qiao Liu, and Wanwen Zeng. Applications of transformer-based language models in bioinformatics: a survey. *Bioinformatics Advances*, 3(1): vbad001, 2023.

[10] Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, et al. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*, 2022.

[11] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

[12] OpenAI. Gpt-4 technical report, 2023.

[13] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[14] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[15] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[16] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598, 2020.

[17] Xiruo Ding, Zhecheng Sheng, Meliha Yetişgen, Serguei Pakhomov, and Trevor Cohen. Backdoor adjustment of confounding by provenance for robust text classification of multi-institutional clinical notes. *arXiv preprint arXiv:2310.02451*, 2023.

[18] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[19] Virgile Landeiro and Aron Culotta. Robust text classification in the presence of confounding bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[20] Virgile Landeiro and Aron Culotta. Robust text classification under confounding shift. *Journal of Artificial Intelligence Research*, 63:391–419, 2018.

[21] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[22] Judea Pearl. *Causality*. Cambridge university press, 2009.

[23] Charles Sutton, Michael Sindelar, and Andrew McCallum. Reducing weight undertraining in structured discriminative learning. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 89–95, 2006.

[24] Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of Biomedical Informatics*, 113:103631, 2021.

[25] Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. The 2022 n2c2/uw shared task on extracting social determinants of health. *arXiv preprint arXiv:2301.05571*, 2023.

[26] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[27] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300, 2020.

[28] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.132.

[29] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5102.

[30] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022.

[31] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

Appendix

A Simulation Configurations

In the simulation of different degrees of distribution shifts by provenance, 4 parameters are required to set up the framework: $P_{train}(y=1|z=0), P_{train}(y=1|z=1), P_{train}(z=1), \alpha_{test}$. The first three were sampled from 0 to 1 evenly in linear space, with a step size of 0.05 or 0.1, depending on scenarios.

Since α represents the ratio for positive rates from two sources and $\alpha = 1$ means same prevalence rates, we sampled α_{test} in a reciprocal manner while centering around 1.0. One such example is $\{..., 1/8, 1/4, 1/2, 1, 2, 4, 8, ...\}$. This can ensure a uniform distribution of α_{test} in the log space, as shown in Figure A1.

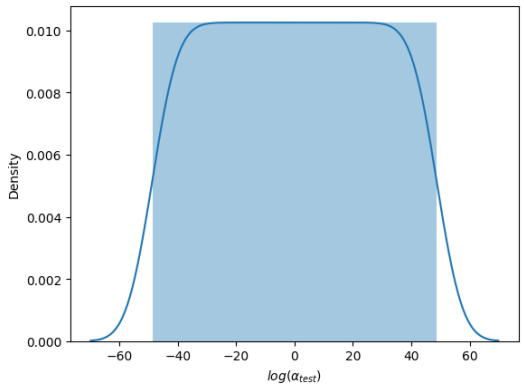


Figure A1: α_{test} distribution.

To break down α_{test} into detailed samplings of its two subpopulations, Figure A2 shows the theoretical sampling for the joint distribution of $P_{test}(y=1|z=0)$ and $P_{test}(y=1|z=1)$. It is worth noting that not all sampling configurations from Figure A2 can lead to a valid training/testing set, given how many positive/negative samples from both sources required for that setting.

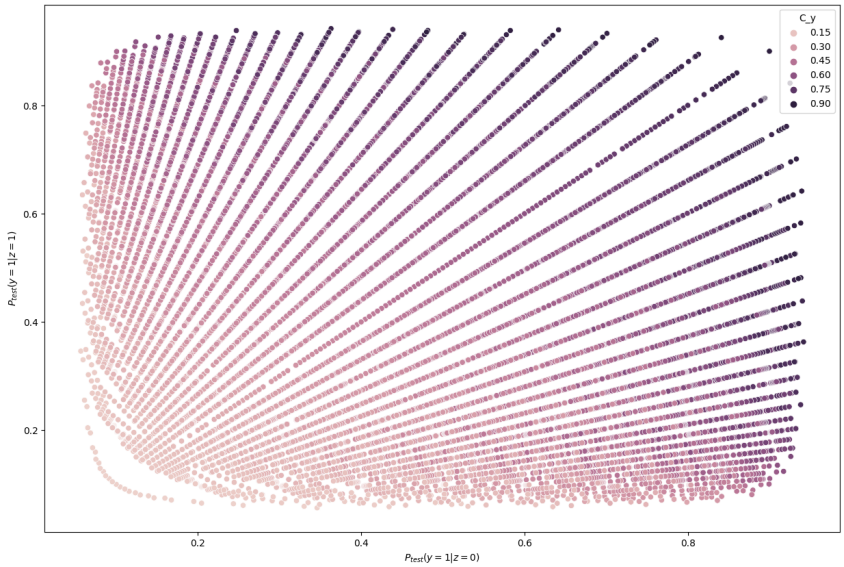


Figure A2: Theoretical sampling from the joint distribution of $P_{test}(y=1|z=0)$ and $P_{test}(y=1|z=1)$.

B Details on Datasets

SHAC SHAC is a dataset of electronic health record notes annotated for social determinants of health (SDoH) that provided a basis for the recent n2c2/UW SDoH Challenge [24, 25]. The Social History Annotation Corpus (SHAC) was designed for extracting Social Determinants of Health (SDOH) in clinical notes under an active learning framework. The notes are from two different sources: clinical notes of chronic pain patients from the University of Washington Medical Center (UW set), discharge notes of intensive care unit patients from MIMIC-III (MIMIC set). Its annotation guideline includes several event types: (1) substance use (alcohol, drug, tobacco); (2) employment; (3) living status. Among those, we selected a small section, Drug Abuse, as the classification target in this work. Summaries are shown in Table A1. To ensure enough samples for most testing scenarios, we set the training set size to 800 and testing set size to 200 for the SHAC dataset.

Table A1: SHAC dataset summary

	Total Number	Identified Drug Abuse Cases	Positive Rate
UW	2,528	1,040	41.1%
MIMIC	1,877	371	19.8%

Hate Speech We collected two publicly available datasets for hate speech detection. The first one (DynGEN set) is a dynamically generated dataset by Vidgen et al. [28], through four rounds of data creation. The first round collected synthetic texts, created by the annotation team to closely mimic real-world content. It was then followed by perturbations in the texts to create new examples for next three rounds. In the end, it results in a total of around 40,000 entries, with a positive rate of 45%.

The second dataset (WSF set) comes from a real-world white supremacist forum published between 2002 and 2017 [29]. Posts were collected from a subset of 22 sub-forums covering diverse topics and nationalities, segmented into sentences, and then manually labeled. Authors used 4 types for annotation: (1) HATE; (2) NOHATE; (3) RELATION, where the sentence itself does not convey any information and must be put in its context to be correctly identified, such as a reply to a hate speech comment; (4) SKIP, where HATE/NOHATE cannot be decided. To reduce noise, we only keep HATE and NOHATE label for our work, which is the majority of the texts. RELATION has 168 records and SKIP has 73.

Table A2 provides the summary of Hate Speech Detection dataset. For this dataset, we set training set size to 4000 and testing set size to 1000.

Table A2: Hate Speech dataset summary

	Total Number	Identified Hate Speech	Positive Rate
DynGEN	41,144	18,969	46.1%
WSF	10,703	1,196	12.2%

C Sentence-BERT Embeddings

To generate embeddings for sentences, we applied the Sentence-BERT model [18]. We adopted the publicly available version of the model `all-MiniLM-L6-v2` from the HuggingFace repository [3]. Sentence-BERT produces a single embedding for a given unit of text. The resulting sentence embeddings serve as predictor variables in the regression model.

D Llama 2 Embeddings

In this work, we extracted sentence embeddings from all three versions of the Llama 2 model (without additional training for dialog). Officially, they are named after how many parameters (roughly) each

[3] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

of them contains: Llama-7B, Llama-13B, and Llama-70B. Embedding dimensions from Llama-7B, 13B, and 70B model are 4096, 5120, 8192, respectively. The first two can be fit into one A100 GPU, while for the largest 70B version, we applied 8-bit quantization [30] before loading it into GPU. For each token, we only kept the outputs from the very last attention layer. From embeddings, we took average as the pooling function (as oppose to use the embedding from the last non-padding token) over each dimension for all tokens in one sentence. This was demonstrated with better performance in the preliminary work.

E Additional Results

In this section, we present baseline results on additional results on Llama 2-13B and quantized Llama 2-70B, for the SHAC dataset (Figure A3, coefficients presented in Table A3) and full results for the Hate Speech Detection dataset (Figure A4, coefficients presented in Table A4). Besides of results on $\alpha_{train} = 0.33$, Figure A5 and Figure A6 show additional results on $\alpha_{train} = 3$, as the reciprocal according to Simulation Configurations in Appendix A. It is noted that the value pair of 0.33 and 0.3 for α_{train} is randomly selected as an example and our interpretation holds true for other α_{test} values (not presented) with differences in degree of the effects. Different α_{test} ranges are shown for results on two datasets because empirical sampling may invalidate some of theoretical combinations.

For unadjusted models, embeddings from foundation models - Llama 2 models in particular - provide best innate robustness to confounding shifts. With respect to this robustness, Sentence-BERT lies at a similar level to the baseline binary unigrams. As for Llama 2 specifically, a larger model in size does not confer more robustness, as shown in our results on the two datasets where Llama 2-7B can produce lowest absolute coefficients in some scenarios (C_y levels).

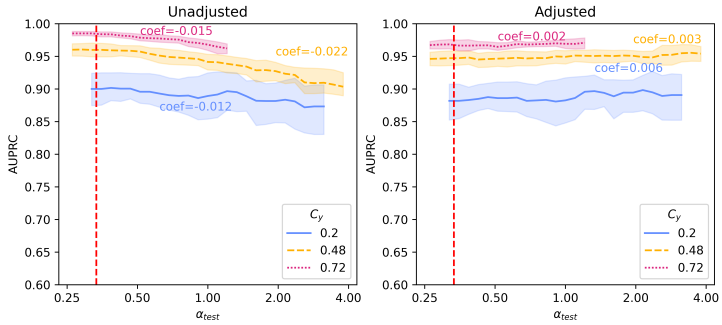
After applying Backdoor Adjustment to the logistic regression models, the absolute values of coefficients for all models drop, indicating better model robustness to provenance-related distribution shifts. Among these, the baseline binary unigrams exhibit the most significant improvements after the adjustment, with the lowest absolute coefficients. However, for results on both datasets, the AUPRC measures from this baseline model are typically lower and deteriorate rapidly in comparison with those using foundation model representations. This performance drop is more significant with the Hate Speech dataset (Figure A4).

When reviewing results for $\alpha_{train} = 3$ in Figure A5 and Figure A6, we observe similar trends. Due to the sampling strategy, only some C_y groups can be matched as in the case where $\alpha_{train} = 0.33$. One apparent difference is that the slope of the lines is now mostly positive, as opposed to the negative slopes with $\alpha_{train} = 0.33$. This indicates the training set can influence the model's baseline performance, in that when the testing set is more different from the training set (in terms of α values in our case), the performance usually drops. This matches the far right area when $\alpha_{train} = 0.33$ and far left area for $\alpha_{train} = 3$.

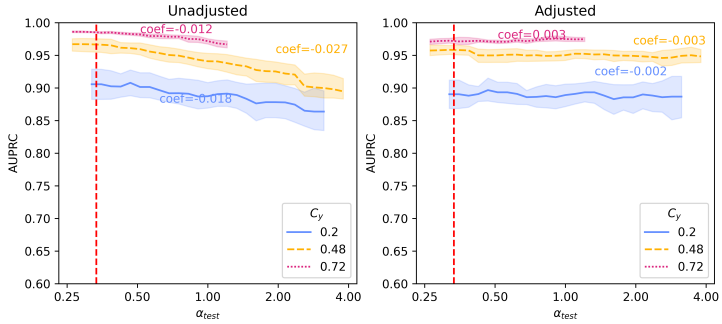
Table A3: Coefficients for regression-based models on SHAC Dataset

		C_y		
		0.20	0.48	0.72
Unadjusted LR	Binary Unigram	-0.016	-0.024	-0.016
	Sentence-BERT	-0.035	-0.044	-0.028
	Llama 2-7B	-0.012	-0.024	-0.013
	Llama 2-13B	-0.012	-0.022	-0.015
	Llama 2-70B	-0.018	-0.027	-0.012
Adjusted LR	Binary Unigram	0.009	-0.001	0.002
	Sentence-BERT	-0.005	-0.012	-0.007
	Llama 2-7B	0.007	0.001	0.005
	Llama 2-13B	0.006	0.003	0.002
	Llama 2-70B	-0.002	-0.003	0.003

* Logistic Regression. † Logistic regression with Backdoor Adjustment.



(a) Llama 2-13B Average



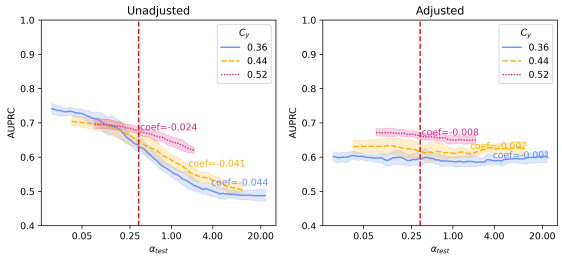
(b) Llama 2-70B Average

Figure A3: Logistic regression and Backdoor Adjustment on SHAC dataset, based on (a) Embeddings generated from Llama 2-13B, then averaged; (b) Embeddings generated from quantized Llama 2-70B, then averaged. Red vertical line indicates $\alpha_{train} = 0.33$, in which setting the model is trained.

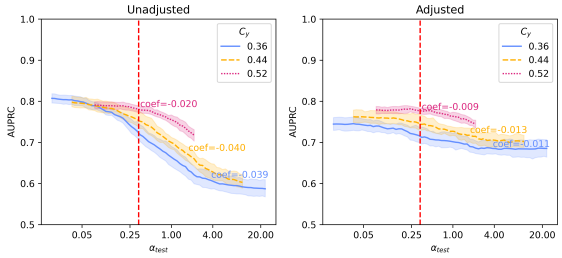
Table A4: Coefficients for regression-based models on Hate Speech Dataset

		C_y		
		0.36	0.44	0.52
Unadjusted LR*	Binary Unigram	-0.044	-0.041	-0.024
	Sentence-BERT	-0.039	-0.040	-0.020
	Llama 2-7B	-0.038	-0.039	-0.023
	Llama 2-13B	-0.039	-0.038	-0.019
	Llama 2-70B	-0.040	-0.040	-0.022
Adjusted LR†	Binary Unigram	-0.001	-0.002	-0.008
	Sentence-BERT	-0.011	-0.013	-0.009
	Llama 2-7B	-0.017	-0.017	-0.014
	Llama 2-13B	-0.016	-0.015	-0.009
	Llama 2-70B	-0.021	-0.021	-0.013

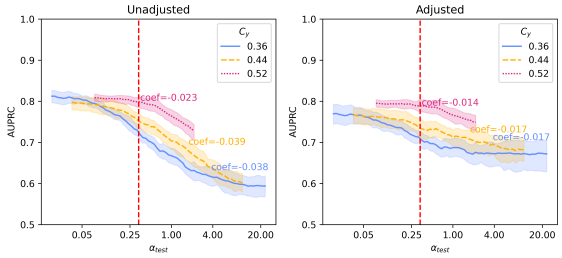
* Logistic Regression. † Logistic regression with Backdoor Adjustment.



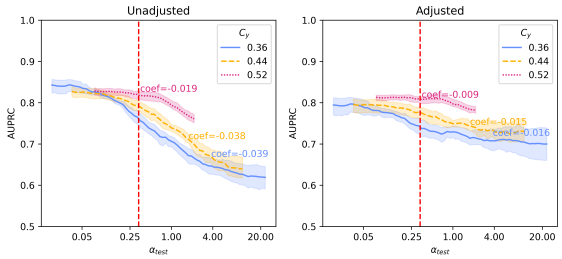
(a) Binary Unigram



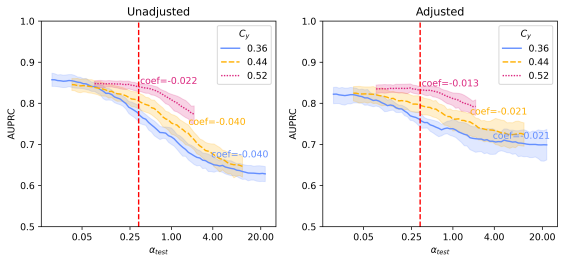
(b) Sentence-BERT



(c) Llama 2-7B Average

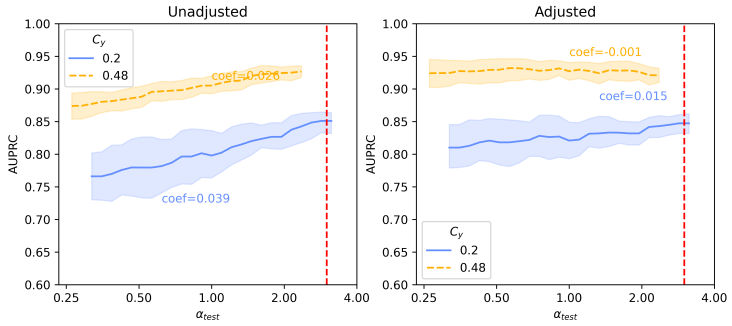


(d) Llama 2-13B Average

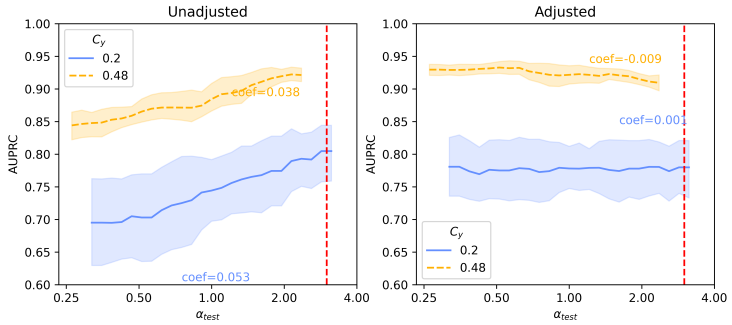


(e) Llama 2-70B Average

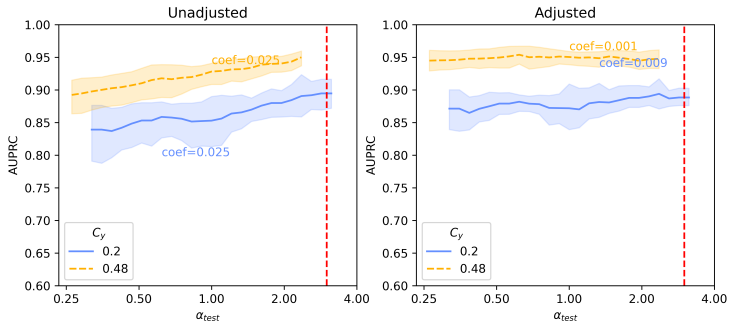
Figure A4: Logistic regression and Backdoor Adjustment on the **Hate Speech Detection** dataset. (a) Binary Unigrams; (b) Embeddings generated from Sentence-BERT; (c) Embeddings generated from Llama 2-7B, then averaged; (d) Embeddings generated from Llama 2-7B, then averaged; (e)Embeddings generated from Llama 2-7B, then averaged. Red vertical line indicates $\alpha_{train} = 0.33$, in which setting the model is trained.



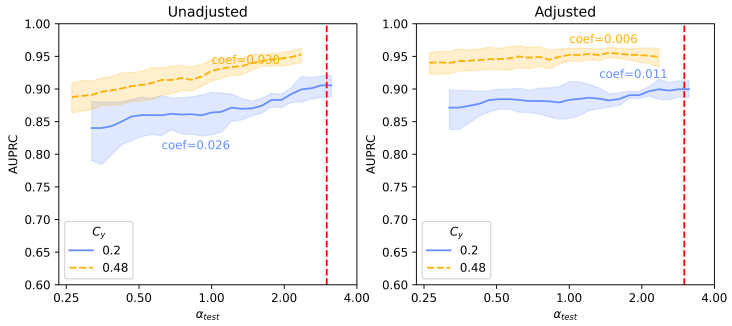
(a) Binary Unigram



(b) Sentence-BERT

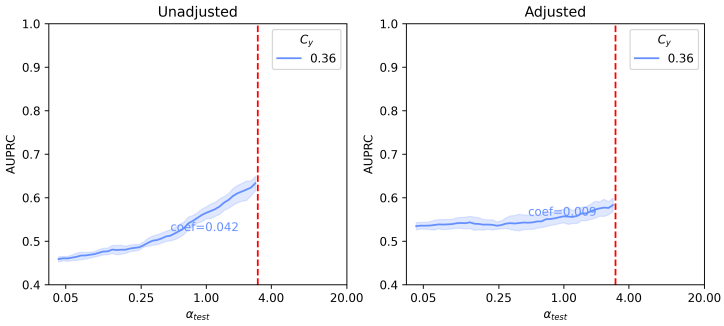


(c) Llama 2-7B Average

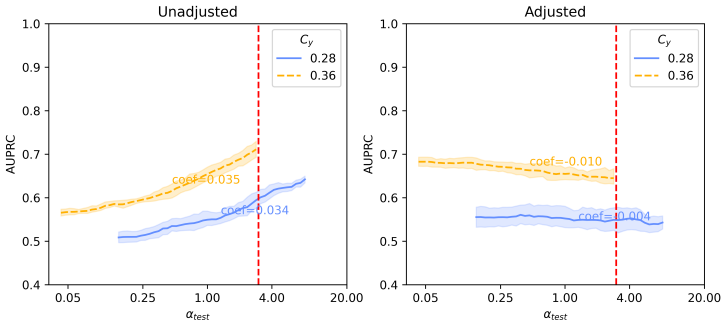


(d) Llama 2-13B Average

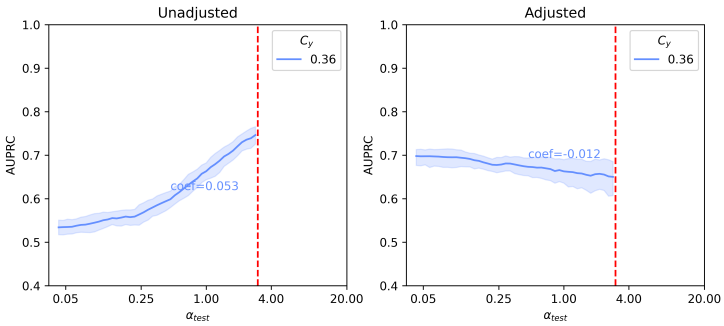
Figure A5: Logistic regression and Backdoor Adjustment on SHAC dataset, based on (a) Binary Unigrams; (b) Embeddings generated from Sentence-BERT; (c) Embeddings generated from Llama 2-7B, then averaged; (d) Embeddings generated from Llama 2-13B, then averaged. Red vertical line indicates $\alpha_{train} = \mathbf{3.0}$, in which setting the model is trained.



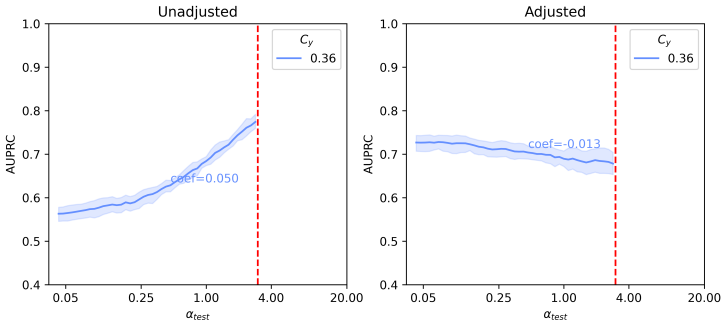
(a) Binary Unigram



(b) Sentence-BERT



(c) Llama 2-7B Average



(d) Llama 2-13B Average

Figure A6: Logistic regression and Backdoor Adjustment on Hate Speech Detection dataset, based on (a) Binary Unigrams; (b) Embeddings generated from Sentence-BERT; (c) Embeddings generated from Llama 2-7B, then averaged; (d) Embeddings generated from Llama 2-13B, then averaged. Red vertical line indicates $\alpha_{train} = \mathbf{3.0}$, in which setting the model is trained.

F Discussion

Overall, the empirical experiments on both datasets show consistent results. The logistic regression model, without any adjustment, is very sensitive to distribution shift, as evident in the higher absolute values of the coefficients of lines fit to the performance curves. However, different embedding methods show various degrees of robustness. When no adjustment is made, models trained on Llama 2 embeddings show best performance, including the smaller 7B or 13B versions, especially when the dataset is small. As with the SHAC dataset, out-of-the-box usage of embeddings from foundation models (e.g., Llama 2-7B) already shows some robustness to confounding shifts. However, without adjustment these representations do not provide significantly better robustness over the baseline unigram model (as in a larger Hate Speech Detection dataset). All these results suggest a need for adjustment to logistic regression models in the context of provenance-related confounding shift, irrespective of the choice of text representation technique.

Backdoor adjustment, when applied to the logistic regression, can significantly improve robustness to these provenance-related shifts. This holds true for different embedding methods on the two datasets evaluated. Among them, binary unigrams benefits most from the adjustment, as shown in Figure 1a when $C_y = 0.48$, especially for the larger Hate Speech Detection dataset in Figure A4a. However, this comes with a loss in performance of AUPRC, which in comparison is not significant for the SHAC dataset with very high baselines.

One reason that Backdoor Adjustment works relatively well on unigrams could be that foundation models generate highly contextual embeddings. In comparison, binary unigrams only store information at the unigram level from any corpus. These two ways of representing text capture different information, which may further affect the distribution of representations. Moreover, it should be noted that the embeddings from different models have different dimensions, which subsequently affect the size of a regression model ($\boldsymbol{\beta}$). As such, it could be argued that models trained on the resulting embeddings are not strictly comparable. A potential way of controlling for this would be to apply dimension reduction methods, to enforce the constraint that all inputs to the regression model must have the same dimension. However, this will inevitably cause loss of information, with effects that are as yet unclear. There is another hyper-parameter, v, applied to the one-hot matrix for Backdoor Adjustment. This serves as an added regulation term in regression models, and may have different effects with different numbers of parameters. In our work, v was set to 10 for all experiments. The optimal v value for models of different sizes remains to be determined. Beyond our empirical experiments, formal testing of these assumptions is left to future work.

G Limitations

In this work, we extended an evaluation framework for provenance-related confounding shift and examined one strategy for mitigating its effect on model stability. Several constraints were enforced during the development of the framework. They were originally set to isolate the effects of changes in site-specific class prevalence, but at the same time limit the scope of testing. Further work is also required to expand the framework to multiple classes and sources, other than binary cases discussed in this work. Moreover, it is left to future work to expand the framework to include other confounders, as in the case of the CivilComments dataset where a combination of 8 dimensions (e..g, gender, religion) collectively serve as the domain label [31].

The unprecedented size of foundation models (Llama 2 especially) limited our ability to utilize it fully in the context of available resources, for example by preventing us from fine-tuning it end-to-end. The path we took in this work is a computationally lightweight approach and only serves as a first step into exploring Llama 2's potential. It remains to be determined whether such foundation models can be more robust when tuned using a parameter-efficient approach, or even fully fine-tuned.

Our primary focus has been on the application of foundation models to classification tasks in multi-institutional datasets within the biomedical domain, though several observations from our biomedical set align well with results from the general domain dataset. Further validation is required to determine how these findings generalize to different scenarios, including those with larger training samples and different model architectures, as well as whether different biomedical subdomains will favor different strategies.