From Bias to Breakdown: Benchmarking Failure Mode Analysis of Single-cell RNA Sequencing Foundation Models in Acute Myeloid Leukemia

Amirreza Naziri[1,2,3], **Arash Asgari**[1,2], **Aijun An**[1,3], **Eleftherios Sachlos**[1,3], **Laleh Seyyed-Kalantari**[1,2,3]

[1]York University
[2]Vector Institute
[3]Connected Minds

naziriam@yorku.ca, arashasg@yorku.ca, aan@yorku.ca, sachlos@yorku.ca, lsk@yorku.ca

Abstract

Foundation models (FMs) trained on large-scale single-cell RNA-seq (scRNA-seq) data have shown strong performance across various biological tasks. These performances are often reported across a large set of test benchmarks across all samples. However, the pretraining data of these models are often highly imbalanced across disease types, patients' conditions, and demographics. For instance, disease samples are rarer and more challenging to collect, and the pretraining sets contain many more healthy cells. Such imbalances can hurt performance on underrepresented disease cases and the equality of the model outcome. To evaluate this hypothesis, we benchmark off-the-shelf scRNA-seq foundation models for cell-type classification in acute myeloid leukemia (AML), a rare but clinically important disease that represents low-prevalence settings. Here, besides overall performance, we conduct subgroup analysis of the outcome across cell types and disease conditions (clinical timepoints). Our results suggest that despite high overall F1 scores in cell-type classification, performance drops in disease conditions and varies across cell types. These findings highlight a limitation of current scRNA-seq foundation models and motivate more balanced pretraining and failure mode analysis rather than an overall performance report.

Introduction

Single-cell RNA sequence (scRNA-seq) has revolutionized our understanding of cellular heterogeneity by measuring transcriptomes at the resolution of individual cells (Kolodziejczyk et al. 2015; Saliba et al. 2014). scRNA-seq has enabled the discovery and annotation of cell types (Van de Sande et al. 2023; Jovic et al. 2022; Hedlund and Deng 2018). scRNA-seq datasets contain a large, sparse matrix including information about gene activities across different cells. The gene activity values can be from 0 to large floating-point numbers, indicating how the gene is expressed. The higher the value is, the more expression is witnessed of that gene.

To analyze and extract patterns from scRNA-seq data, transformer-based foundation models are arising, which are large neural networks pretrained in an unsupervised fashion on massive unlabeled data and finetuned for diverse downstream tasks, like cell type classification. For example, scBERT (Yang et al. 2022), an encoder-only FM, adapts the BERT masked-language framework to gene expression by learning contextualized embeddings of genes and cells through token-level reconstruction on discretized counts. scGPT (Cui et al. 2024a), a decoder-only FM, learns joint embeddings of cells and genes via masked language modelling on multi-omic scRNA-seq data. scFoundation (Hao et al. 2024) scales to 100 M parameters by incorporating read-depth–aware objectives over 50 M cells. The core idea in scFoundation is skipping zero and masked tokens in encoder layers, which can significantly reduce the computing needs without losing performance. Geneformer (Cui et al. 2024b), on the other hand, leverages transfer learning from a pretrained context-aware, attention-based model trained on a corpus of approximately 30 million single-cell transcriptomes, enabling context-specific network-biology predictions even with limited task-specific data.

Pretraining data for the foundation models often contain biases and imbalances. In single-cell RNA-seq (scRNA-seq), these appear in several ways that can distort biological interpretation and downstream analysis. Cell-type imbalance is very common (Maan et al. 2024, 2022). Disease-related imbalances also arise, as immune responses and cellular populations differ between infected, healthy, and cancerous tissues. At the patient level, demographic differences (age, sex, ethnicity, ancestry) can affect the datasets and biological patterns in different diseases (Darolti and Mank 2023; Huang et al. 2021). Technical factors, including batch effects and platform-specific biases, add further variability (Maan et al. 2024, 2022). Together, these imbalances may lead to unequal performance and poor generalization, especially for rare or complex cases. For example, Acute Myeloid Leukemia (AML) represents a complex, aggressive, diverse, and extremely rare group of blood cancers. Evaluating the performance of foundation models on this dataset will help us to reveal potential bias or misperformance issues in the underrepresented data.

As a related work, a recent study (Alsabbagh et al. 2023) created skewed single-cell datasets to benchmark three foundation models (scGPT, scBERT, and Geneformer (Cui et al. 2024b; Yang et al. 2022; Cui et al. 2024a)), showing that all models exhibited reduced performance on rare cell types when trained on biased finetuning data. This demonstrated

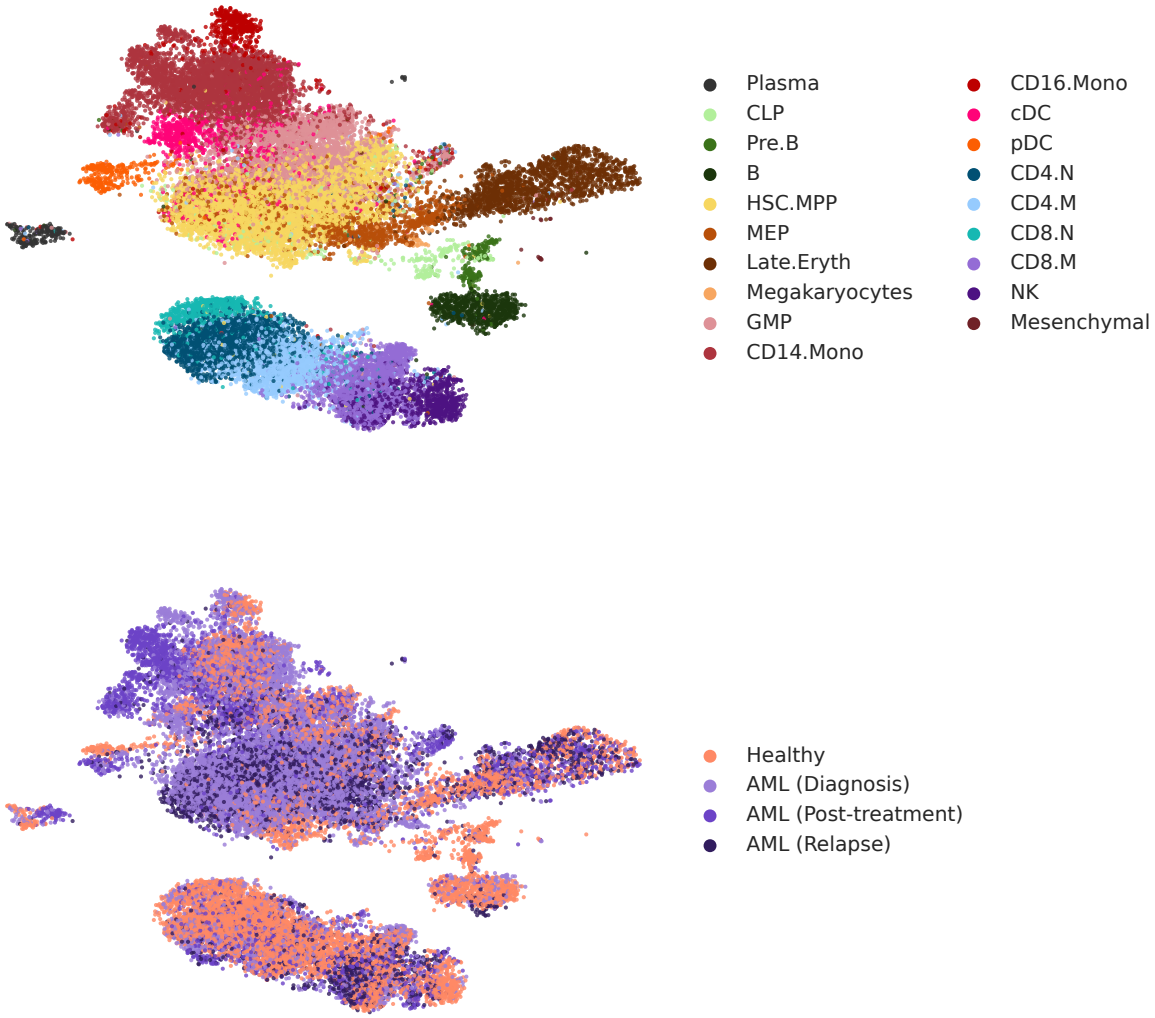

Figure 1: UMAP visualization of Hematopoietic Niche scRNA-seq (Acute Myeloid Leukemia) (Ennis et al. 2023). Each point represents a single cell colored by (top) cell type and (bottom) clinical timepoint.

that dataset imbalance during finetuning can also directly impact model performance. In contrast, our study reveals that bias can still arise from pretraining, even when finetuning data are not skewed.

Overall, in this study, we analyzed the performance of recent and widely used foundation models (Geneformer, scBERT, scFoundation, and scGPT (Cui et al. 2024b; Yang et al. 2022; Hao et al. 2024; Cui et al. 2024a)), finetuned on the AML dataset, for their ability in cell-type classification.

We only explored the cell type classification as the downstream task, due to data availability. Rather than reporting only overall performance, we examined how imbalances of original pretrained data lead to systematic performance gaps across models and disease states in a rare disease like AML.

We observed that even though our finetune/test dataset is relatively balanced through different disease conditions (clinical timepoint), the foundation models perform stronger on healthy cases. This shows that the foundation models are more biased towards healthy cases and more general cell types compared to rare and underrepresented cases.

Method

Dataset

In this study, we used the publicly available Hematopoietic Niche scRNA-seq (Acute Myeloid Leukemia) dataset (Ennis et al. 2023). It contains $\sim$350,000 cells from $\sim$50 human donors, each profiled across $\sim$16,000 genes. The dataset includes cell-type labels and clinical timepoints (Fig. 1). We sampled $\sim$250,000 cells from the dataset to have balanced distribution of clinical timepoints (approximately 25% Healthy, 30% AML (Diagnosis), 22% AML (Post-treatment), and 22% AML (Relapse)). It is worth mentioning that the raw data is compiled from multiple datasets, each with different donors and disease conditions. For more

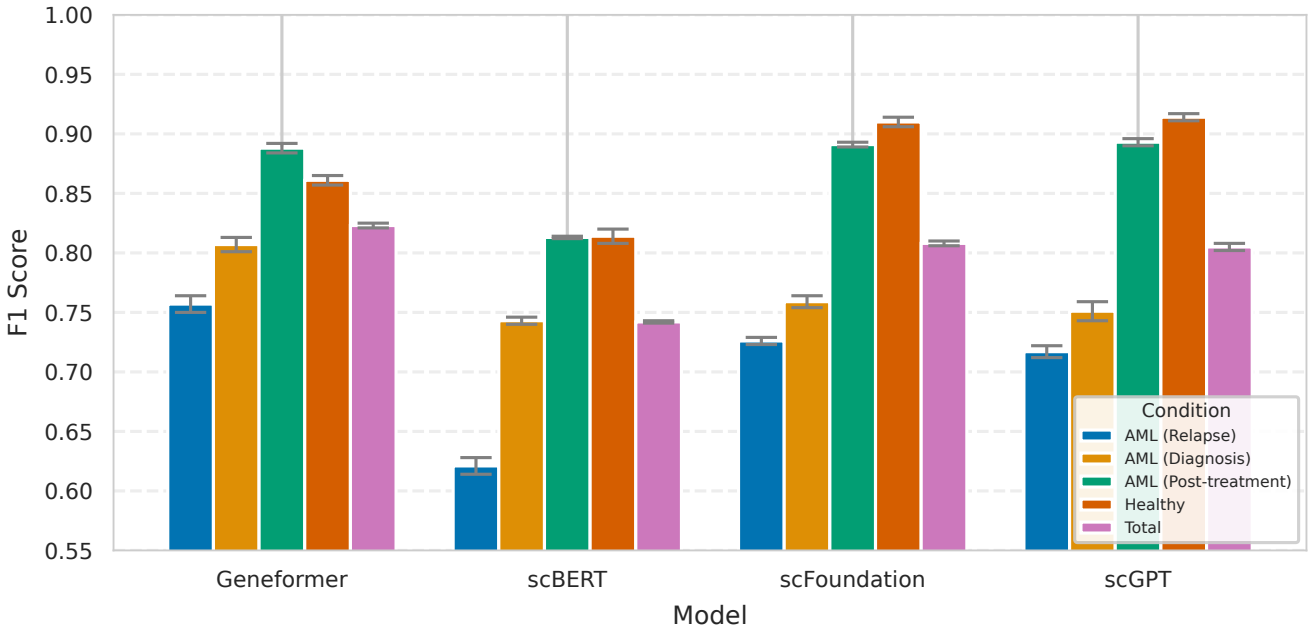

Figure 2: Performance of four foundation models (Geneformer, scBERT, scFoundation, scGPT) stratified by clinical timepoint. While models perform well on Healthy samples, performance drops in AML, with the largest decline observed in relapse cases. This pattern highlights a systematic gap: foundation models pretrained on mostly healthy single-cell data generalize poorly to clinically challenging disease states, even though our evaluation dataset was evenly balanced across all timepoints.

details, please refer to the original study (Ennis et al. 2023).

Model

We evaluated cell type classification performance of recent, widely used foundation models for scRNA-seq. A brief description of each model is given below:

- **scBERT** (encoder-only): adapts BERT's masked-token objective to discretized gene counts, learning contextual embeddings of genes and cells for classification (Yang et al. 2022).
- **scGPT** (decoder-only): uses masked language modeling to learn joint cell–gene embeddings across large multi-omic corpora (Cui et al. 2024a).
- **scFoundation**: ∼100M parameters with read-depth–aware objectives trained on ∼50M cells; skips zero/-masked tokens in the encoder to cut compute without accuracy loss (Hao et al. 2024).
- **Geneformer**: transfers from a large, context-aware attention model pretrained on ∼30M transcriptomes, enabling strong performance with limited task data (Cui et al. 2024b).

Experiments

First, we split the data into train (80%), test (10%), and validation (10%) sets using stratified splitting, ensuring that all cell types and timepoints (disease condition: healthy, relapse, diagnosis, and post-treatment) were represented across splits and no patients ' samples were distributed across splits.

We then added a linear head to the top of each foundation models and finetuned them on cell type classification labels using the train and validation sets. Performance was evaluated on the held-out test split. Since each FM has its own preprocessing pipeline, we applied the corresponding procedures to ensure that inputs were aligned with the models' requirements. For finetuning, we used the publicly released model checkpoints and followed the default hyperparameters proposed by each method. To confirm robustness, results are reported from 5-fold cross-validation on the test set, with each fold held out once.

Results

Performance disparity across clinical timepoints (Disease Status)

To assess how foundation models perform across health and disease conditions, we stratified performance by clinical timepoints. When aggregating results across all four models (Geneformer, scBERT, scFoundation, and scGPT (Cui et al. 2024b; Yang et al. 2022; Hao et al. 2024; Cui et al. 2024a)), we observed consistently higher F1 scores for Healthy donors compared to AML cases (Fig. 2). Within the AML cases, relapse samples posed the greatest challenge, with performance dropping more sharply than in primary AML samples.

For instance, scGPT maintained F1 scores above 0.90 in healthy subsets but fell below 0.75 in relapse AML samples,

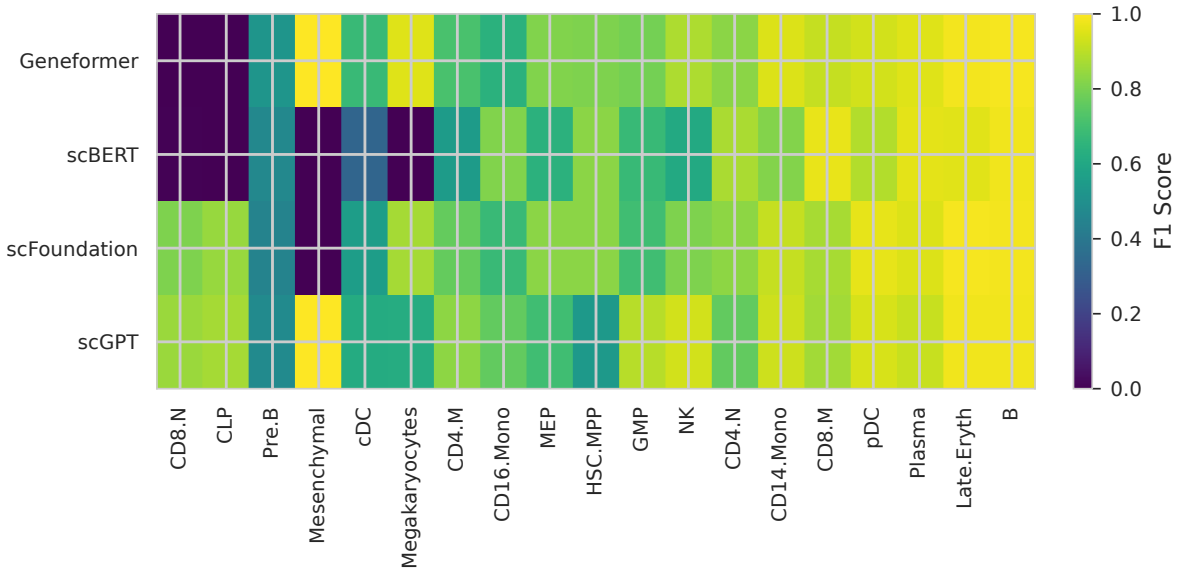

Figure 3: Heatmap of F1 scores by cell type across four foundation models. Performance is uneven: common and well-represented populations such as B cells show strong accuracy, while others like CLP, Pre.B exhibit markedly lower scores. In addition, some cell types are totally mistaken by foundation models, resulting in zero in their F1 Score.

with similar trends observed for scBERT and scFoundation. These patterns suggest that model generalization is not uniform and that disease-specific complexity, especially under relapse conditions, creates systematic performance disparities.

Performance disparity across cell types

In order to assess how foundation models perform on different cell types, we analyzed their performance per cell type individually. As shown in Fig. 3, we observe that all foundation models struggle to correctly predict certain important cell types, such as CLP and Pre.B, which contribute to the immune deficiency commonly seen in AML patients (Khaldoyanidi et al. 2021). On the other hand, all models could perform very well on common, general B cells. This imbalance underscores how under-represented cell types drive systematic errors, limiting the reliability of model predictions in clinically relevant contexts.

Discussion

Our results show failures and limitations of current scRNA-seq foundation models that are observed by subgroup analysis rather than reporting overall performance. Such biases may emphasize the need for more balanced pretraining data for training foundation models and the importance of bias analyses. To mitigate such biases on the data side, utilizing balanced representation of real-world data is essential.

Approaches such as targeted augmentation and synthetic oversampling (Bej et al. 2021), or transfer learning (Mieth et al. 2019) from related disease datasets can help strengthen rare or undersampled cell populations. This would give models better exposure to the populations where they perform poorly. On the modeling side, applying bias mitigation methods, like adversarial learning, may reduce systematic errors (Zhang, Lemoine, and Mitchell 2018).

From a benchmarking perspective, current practices often highlight only aggregate performance, which cannot reveal critical subgroup failures. We suggest benchmarking the outcome on subgroup-level and cell type–specific metrics as a standard substitute approach to give a better picture of the failure mode of the foundation models. This makes disparities visible and allows the community to track progress toward more equitable foundation models.

Limitations and Future Work

Our evaluation was limited in several ways. First, we focused only on cell type classification, as it was the only available label in our dataset. In future work, upon data availability, we would like to extend the evaluation to a broader range of downstream tasks and demographic features of patients, such as race and ancestry.

Conclusion

Our study shows that strong overall performance does not always mean fair and equal performance across all subgroups. Analyzing the failure mode supports the detection of foundation model biases. A potential actionable target for mitigating such biases in biomedical AI can be a better representation of rare or disease-specific cell types in pretrained data. We suggest bias investigation in failure mode across features to be a standard part of evaluating these models, not an afterthought.

Acknowledgments

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant, the Connected Minds initiative funded by the Canada First Research Excellence Fund (CFREF), the Center for AI & Society at York University (AI-EGS), and the Vector Institute for providing access to high-performance computing resources.

References

Alsabbagh, A. R.; de Infante, A. M. R.; Gomez-Cabrero, D.; Kiani, N. A.; Khan, S. A.; and Tegnér, J. N. 2023. Foundation Models Meet Imbalanced Single-Cell Data When Learning Cell Type Annotations. *bioRxiv*.

Bej, S.; Galow, A.-M.; David, R.; Wolfien, M.; and Wolkenhauer, O. 2021. Automated annotation of rare-cell types from single-cell RNA-sequencing data through synthetic oversampling. *BMC Bioinformatics*, 22(1): 557.

Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Duan, N.; and Wang, B. 2024a. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8): 1470–1480.

Cui, Z.; Xu, T.; Wang, J.; Liao, Y.; and Wang, Y. 2024b. Geneformer: Learned gene compression using transformer-based context modeling. 8035–8039.

Darolti, I.; and Mank, J. E. 2023. Sex-biased gene expression at single-cell resolution: cause and consequence of sexual dimorphism. *Evol. Lett.*, 7(3): 148–156.

Ennis, S.; Conforte, A.; O'Reilly, E.; Takanlu, J. S.; Cichocka, T.; Dhami, S. P.; Nicholson, P.; Krebs, P.; Broin, P. Ó.; and Szegezdi, E. 2023. Cell-cell interactome of the hematopoietic niche and its changes in acute myeloid leukemia. *Iscience*, 26(6).

Hao, M.; Gong, J.; Zeng, X.; Liu, C.; Guo, Y.; Cheng, X.; Wang, T.; Ma, J.; Zhang, X.; and Song, L. 2024. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8): 1481–1491.

Hedlund, E.; and Deng, Q. 2018. Single-cell RNA sequencing: technical advancements and biological applications. *Molecular aspects of medicine*, 59: 36–46.

Huang, Z.; Chen, B.; Liu, X.; Li, H.; Xie, L.; Gao, Y.; Duan, R.; Li, Z.; Zhang, J.; Zheng, Y.; and Su, W. 2021. Effects of sex and aging on the immune cell landscape as assessed by single-cell transcriptomic analysis. *Proceedings of the National Academy of Sciences*, 118(33): e2023216118.

Jovic, D.; Liang, X.; Zeng, H.; Lin, L.; Xu, F.; and Luo, Y. 2022. Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and translational medicine*, 12(3): e694.

Khaldoyanidi, S.; Nagorsen, D.; Stein, A.; Ossenkoppele, G.; and Subklewe, M. 2021. Immune biology of acute myeloid leukemia: Implications for immunotherapy. *J. Clin. Oncol.*, 39(5): 419–432.

Kolodziejczyk, A. A.; Kim, J. K.; Svensson, V.; Marioni, J. C.; and Teichmann, S. A. 2015. The technology and biology of single-cell RNA sequencing. *Molecular cell*, 58(4): 610–620.

Maan, H.; Zhang, L.; Yu, C.; Geuenich, M.; Campbell, K. R.; and Wang, B. 2022. The differential impacts of dataset imbalance in single-cell data integration. *bioRxiv*.

Maan, H.; Zhang, L.; Yu, C.; Geuenich, M. J.; Campbell, K. R.; and Wang, B. 2024. Characterizing the impacts of dataset imbalance on single-cell data integration. *Nat. Biotechnol.*, 42(12): 1899–1908.

Mieth, B.; Hockley, J. R. F.; Görnitz, N.; Vidovic, M. M.-C.; Müller, K.-R.; Gutteridge, A.; and Ziemek, D. 2019. Using transfer learning from prior reference knowledge to improve the clustering of single-cell RNA-Seq data. *Sci. Rep.*, 9(1): 20353.

Saliba, A.-E.; Westermann, A. J.; Gorski, S. A.; and Vogel, J. 2014. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, 42(14): 8845–8860.

Van de Sande, B.; Lee, J. S.; Mutasa-Gottgens, E.; Naughton, B.; Bacon, W.; Manning, J.; Wang, Y.; Pollard, J.; Mendez, M.; Hill, J.; et al. 2023. Applications of single-cell RNA sequencing in drug discovery and development. *Nature Reviews Drug Discovery*, 22(6): 496–520.

Yang, F.; Wang, W.; Wang, F.; Fang, Y.; Tang, D.; Huang, J.; Lu, H.; and Yao, J. 2022. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10): 852–866.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating Unwanted Biases with Adversarial Learning. arXiv:1801.07593.