Influence of Gender-Specific Data Imbalance on scGPT Fine-Tuning for Single-Cell Genomics

Mohammad Aman Ullah Al Amin[1], **Daniil Filienko**[2], **Hong Qin**[1,3]
[1]Department of Computer Science, Old Dominion University, Norfolk, VA, USA
[2]School of Engineering & Technology, University of Washington, Tacoma, WA, USA
[3]School of Data Science, Old Dominion University, Norfolk, VA, USA
malam007@odu.edu, daniilf@uw.edu, hqin@odu.edu

Abstract

The transformer-based foundation model scGPT has demonstrated strong capabilities in analyzing high-dimensional single-cell RNA sequencing data. However, the impact of demographic factors, particularly gender, on model performance remains insufficiently understood. Gender is known to influence cell-type compositions in the immune system. Here, using the gender-sensitive cell-type composition in immune system, we comprehensively evaluated how the gender-sensitive imbalance of training data influences the performance of scGPT in cell-type predictions. We fine-tuned scGPT on male-only, female-only, and mixed-gender subsets from two large-scale datasets containing immune cells. We used a logit difference to measure the confidence gap between the true label and the actual model prediction. The confidence gap is zero for perfect classifications and negative for incorrect predictions. We observed that training and testing configurations with aligned gender distributions generally showed higher prediction confidence, while mismatched gender during training and testing, especially when training excludes one gender, leads to substantial confidence drops. We also found that training with mixed-gender data promoted more balanced generalization, but did not eliminate all biases. We conclude that gender-specific data imbalance, represented by immune cell-type subpopulation variation between women and men, can influence fine-tuning of scGPT and its performance in cell-type classification, highlighting the importance of addressing such demographic biases in biomedical AI models.

Introduction

Foundation models for single-cell genomics have become increasingly valuable in biological research, especially for analyzing high-dimensional data such as single-cell RNA sequencing (scRNA-seq) (Szałata et al. 2024). Among these, *single-cell Generative Pre-trained Transformer (scGPT)* stands out by adapting generative pre-training techniques from natural language processing (NLP) to learn meaningful representations of gene expression at the single-cell level, capturing relationships among genes, similar to how language models understand sentence structure, enabling it to accomplish complex tasks such as cell-type classification and biological inference (Vaswani et al. 2017; Cui et al. 2024).

Biological sex, often referred to as gender, can influence immune cell composition, gene expression, and disease risk. Recent single-cell studies of the human brain have revealed that persistent gender-biased gene expression patterns throughout development and adulthood affect both the neural and immune systems (Zelco and Joshi 2025). Similar gender-specific immune responses are observed in autoimmune diseases such as multiple sclerosis, where males and females exhibit distinct cellular dynamics and inflammatory profiles (Pan et al. 2024). These expression biases vary between cell-types and developmental stages, reflecting chromosomal, hormonal, and evolutionary influences (Darolti and Mank 2023). In addition, gender affects the regulatory architecture in tissues, suggesting widespread functional variation linked to gender-based biology (Jones et al. 2024).

Benchmarking of foundation models such as scGPT, scBERT, and Geneformer has shown that class imbalance is common in single-cell datasets impairs cell-type annotation performance, particularly for rare cell-types (Alsabbagh et al. 2023). On the Zheng68K dataset, scGPT's macro F1 score dropped from 0.725 (default) to 0.616 with random undersampling, while oversampling improved it to 0.968. In the MS dataset, macro F1 improved from 0.753 to 0.962 with oversampling (Alsabbagh et al. 2023).

To address class imbalance in bioinformatics, Abu Shanab et al. showed that combining feature selection with data sampling significantly improved minority-class classification in high-dimensional gene expression data (Shanab and Khoshgoftaar 2018). Similarly, Bugnon et al. explored deep neural networks for extreme class imbalance (e.g., microRNA classification with 1:2000 ratios), finding that unsupervised pretraining and hierarchical filtering outperformed traditional models (Bugnon et al. 2019). Dittman et al. evaluated sampling methods for Random Forests on imbalanced bioinformatics datasets, observing slight but statistically insignificant gains (Dittman, Khoshgoftaar, and Napolitano 2015). Alongside data level approaches, architecture and training level solutions have also been developed to improve rare population annotation. For example, scBalance (Cheng et al. 2023) used dropout during training to improve robustness across protocols and enhance rare cell-type detection.

Building on these insights, Zeng et al. introduced CellFM, a foundation model pretrained on transcriptomes from 100 million human cells. Benchmarks across intra and inter-dataset settings showed that CellFM outperformed models such as scGPT in cell-type annotation, perturbation prediction, and gene function inference (Zeng et al. 2025). These findings highlight the need to audit foundation models for fairness, as gender-specific and cell-type imbalances in training data could amplify bias.

In this study, we examined how gender-specific data imbalance affects scGPT's performance in cell-type classification. We fine-tuned the model using male-only, female-only, and mixed-gender configuration from two large-scale single-cell datasets: the Asian Immune Diversity Atlas (AIDA) and the Acute Myeloid Leukemia (AML) Atlas (Kock et al. 2025; Whittle et al. 2024). We quantified model confidence using the $\Delta\ell$ metric as confidence gap, explored how the composition of the training data influences fairness and generalization in foundational single-cell model (scGPT).

Materials & Methods

We fine-tuned scGPT on gender-specific subsets to evaluate the impact of demographic composition on model fairness and downstream performance.

Model Architecture and Fine-Tuning

scGPT is a GPT-style transformer for scRNA-seq that tokenizes genes and represents expression as either categorical bins or continuous values. It uses biologically inspired attention masks and a CLS token for cell-type classification. We fine-tuned a pretrained scGPT using cross-entropy objective under three configurations: **mixed** (equal number of male and female-donor cells), **male-only**, and **female-only**. All models were evaluated on the same gender-balanced mixed test set.

Datasets

We used two large human scRNA-seq resources: the Asian Immune Diversity Atlas (AIDA) and the Acute Myeloid Leukemia (AML) Atlas. AIDA contains $>$ 1M PBMCs from hundreds of healthy donors across multiple Asian countries(Chinese, Indian, Japanese, Korean, and Malay donors in Japan, Singapore, and South Korea) (Kock et al. 2025); AML comprises 748,679 cells from 159 AML patients and 44 healthy donors collected across 20 studies (Whittle et al. 2024).

Evaluation Metrics

Confidence gap. For each cell with true class y, let ℓ_j be the logit for class j. We define

$$\underbrace{\Delta\ell}_{\text{confidence gap}} \;=\; \ell_y \;-\; \max_j \ell_j \,.$$

Thus, $\Delta\ell = 0$ when the true-class logit equals the largest logit (correct with zero margin), and $\Delta\ell < 0$ otherwise. More negative values mean lower confidence (Weng et al. 2023). All values are reported in *logit* units.

To reduce overplotting, cells with $\Delta\ell = 0$ are *excluded* from the regression fits and the regression scatter plots; this exclusion does not affect the standard classification metrics.

Population percentage: it denotes the percentage (0–100) of a given cell-type in a specified split (train or test) and is used as a predictor in the regression analyses:

$$p_s(c) \;=\; 100 \times \frac{\#\{\text{cells of type } c \text{ in split } s\}}{\#\{\text{cells in split } s\}} \in [0, 100]\,.$$

For clarity, we use the terms *population percentage* and *cell-type percentage* interchangeably to refer to the proportion of each cell-type within a given training or testing dataset.

Preprocessing

All data were processed using the **scGPT Preprocessor** class. For **AIDA**, only expression binning was applied, as normalization was handled by the original providers. For **AML**, the top 1,402 highly variable genes were selected using Seurat v3, followed by expression binning. Cells with missing gender labels were excluded.

Sampling. *AIDA:* three configurations: **mixed** (balanced by donor gender; 38,061 male and 38,061 female; ; 90/10 train/test split, stratified by cell-type), **male only**(68,509 training cells), and **female only** (68,509 training cells); all evaluated on the same gender balanced test set (3,800 male and 3,800 female). *AML:* same configurations; evaluated on fixed, cell-type stratified test subsets (**female only** $n = 146$, **male only** $n = 150$, **mixed** $n = 296$).

To analyze the behavior of the model, we distinguish between the 'population' and 'population+gender' plots. Both use the same test set, but differ in regression inputs: the 'population' graphs use percentage of cell-types from the training data as the x-axis, while the 'population+gender' graphs use test-set cell-type percentage on the x-axis and include gender as a covariate. This setup allows us to isolate the impact of the composition of training population distribution of cell-type and gender on the confidence of the fixed test set.

Visualization and Analysis

We used several types of visualizations to understand model behavior and the effects of gender differences.

Regression models. For each cell i with true class y_i, we estimated the confidence gap ($\Delta\ell$). Population percentages $p_s(c)$ were estimated using a formula defined above. In the univariate model we used $p_{\text{train}}(c_i)$; in the multivariate model we use $p_{\text{test}}(c_i)$.

Univariate Regression (Population only):

$$\Delta\ell_i \;=\; \beta_0 \;+\; \beta_1\, p\big(c_i\big) \;+\; \varepsilon_i.$$

We reported β_1 (the slope vs. $p_{\text{train}}(c_i)$), R^2, and the Pearson correlation r with two-sided p-value.

Multivariate Regression (Population + Gender): Let $G_i = 1$ if *female* and $G_i = 0$ if *male* (male is the reference) (Hardy 1993). We fit

$$\Delta\ell_i \;=\; \beta_0 \;+\; \beta_1\, p\big(c_i\big) \;+\; \gamma\, G_i \;+\; \varepsilon_i.$$

Here β_1 is the slope vs. $p_{\text{test}}(c_i)$ controlling for G_i, and γ is the (female$-$male) shift in $\Delta\ell$ controlling for $p_{\text{test}}(c_i)$.

Scatter plots. Each point is a cell colored by gender. For panels (a–c) we use the population percentage on the x-axis and fit a single OLS line (no gender term). For panels (d–f) we use the population percentage; we draw one overall OLS line and compute a multiple OLS model with a female dummy to test the gender effect while controlling for population percentage. Cells with $\Delta\ell = 0$ (perfect predictions) are excluded from the plots and the fits. Summary statistics are in Tables 1 and 2.

In all six scatter plots, the y axis is the *confidence gap* ($\Delta\ell$), computed on the same mixed gender test set for each fine tuned model.

Top row (a to c). Each test-cell point is positioned by the percentage of its cell-type in the corresponding *training* dataset (female, male, or mixed), and the y value comes from the matching fine-tuned model. This shows whether the model was more confident for cell-types that were more common in its training data. For these panels, we fit a simple regression of the confidence gap on the training set cell-type percentage only, with no gender term.

Bottom row (d to f). The x-axis represents *test-set* cell-type percentages; the model varies by panel (female-trained, male-trained, mixed-trained). For these panels, we fit a multiple regression of the confidence gap on test-set population percentage and gender to test whether gender has any effect after adjusting for cell-type composition.

Due to limited sample sizes in AIDA and AML, stratified sampling did not always provide identical cell-type frequencies. As a result, the x-axis distributions differ between the top and bottom rows of scatter plots, which is considered when interpreting the results.

Population percentage heatmaps show how cell-type percentages vary between gender-specific training and testing datasets. These distributions help explain confidence patterns by showing which cell-types are more or less common, which can influence the model's behavior.

Confidence gap heatmaps show how model confidence changes across gender-aligned and cross-demographic training/testing configurations. The plots show which cell-types struggle when a model trained on one gender, is tested on the other.

Results and Discussion

We evaluated the influence of gender-specific data imbalance on scGPT performance by comparing models fine-tuned on female-only, male-only, and mixed-gender data. We examined how the confidence gap correlates with cell-type abundance and gender composition. For the population only plots, we fit a univariate regression of $\Delta\ell$ on the training set cell-type percentage. The predictor comes from the training data but is assigned to each *test* cell by matching its cell-type; thus all plotted points are test cells. To examine whether gender may have a 'direct' effect on model confidence, we fit a multiple linear regression in which each test cell's confidence gap was predicted using two factors: (i) the percentage of that cell's type in the set *test* (to reflect class imbalance) and (ii) the donor's gender, represented as a binary variable (male = 0, female = 1). The multiple regression model was specified as:

$$\Delta\ell \sim \text{CellTypePercentage} + \text{Gender}$$

The confidence gap ($\Delta\ell$) is modeled as a function of population percentage and gender.

Regression Analysis in AIDA

Here, we compared model confidence under three training setups: female-only, male-only, and mixed-gender as shown in Fig. 1 and Table 1. Each panel plots the confidence gap ($\Delta\ell$) against the testing cell-type population percentage using jittered scatterplots colored by gender, with linear regression overlays. The slope captures how strongly confidence varies with cell-type percentage. In multiple regression, we additionally included gender to assess whether confidence depends on both population and gender.

For the female-tuned model (Fig. 1(a)), we observed a significant positive association between confidence gap and cell-type percentage (slope $= 0.0406$, $p = 0.00372$, $R^2 = 0.0125$). The positive slope indicated that the model tends to be more confident on cell-types that were more abundant. Visual inspection also showed a wider spread for male donor cells, consistent with their absence during female-only training which resulted in poor generalization.

For the multiple regression of the female-tuned model (Fig. 1(d)), we found a statistically significant and positive correlation between confidence gap and cell-type percentage (Slope (pop) = 0.0362, $p = 0.00628$, $R^2 = 0.0113$). The gender variable was not found to be a significant factor in multiple regression.

For the male-tuned model, we observed a similar qualitative trend, but the associations were not statistically significant in either the univariate regression (Fig. 1(b)) or the multiple regression (Fig. 1(e)).

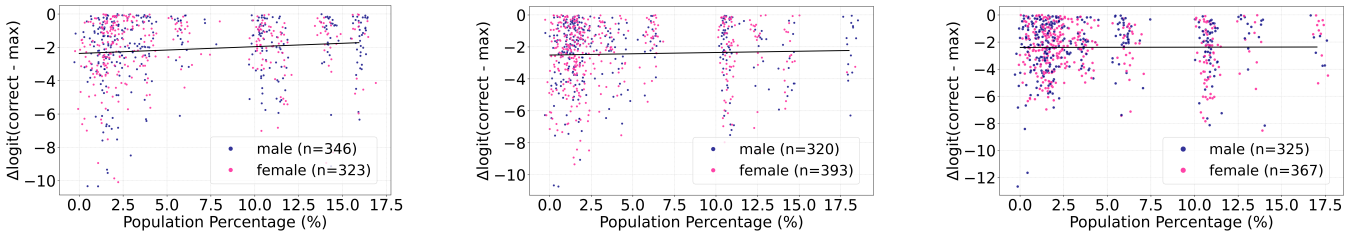
The mixed-tuned model was found to be the least-biased in both univariate and multiple-regression as shown in (Fig. 1(c), (f)); because correlation slopes between confidence gap and cell-type percentage were more than 10x smaller than those in male-tuned, and nearly 20x smaller than those in female-tuned models.

Overall, in the AIDA dataset, we found that mixed-gender tuned model was the least biased model. Single-gender models, especially the female-tuned model, show reduced confidence due to an imbalanced cell-type representation.

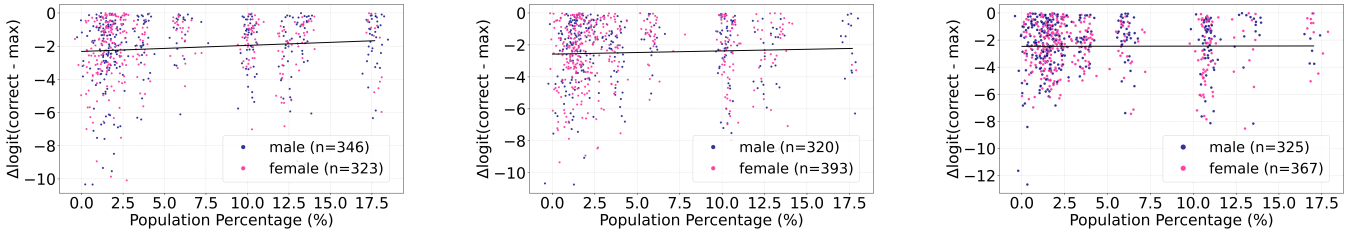
Heatmap Analysis in AIDA

Here, we used heatmaps to further examine the influence of gender-specific data imbalance on model performance. We illustrate the population percentage difference between the testing and training samples by cell-types in a heatmap as shown in Fig. 2(a).

Positive values (red) indicate the most common cell-types in the test set, and negative values (blue) indicate overrepresentation in the training set. This analysis revealed immune cell distribution differences in gender-mismatched train/test comparisons that align broadly with established immunological findings (Escrivà-Font, Cao, and Consiglio 2025).



(a) Female-tuned: $\Delta\ell \sim$ population (b) Male-tuned: $\Delta\ell \sim$ population (c) Mixed-tuned: $\Delta\ell \sim$ population



(d) Female-tuned: $\Delta\ell \sim$ pop + gender (e) Male-tuned: $\Delta\ell \sim$ pop + gender (f) Mixed-tuned: $\Delta\ell \sim$ pop + gender

Figure 1: Regression analysis of model performance across female, male, and mixed-tuned models in the AIDA dataset. Jittered scatterplots were used for better visualization. All regressions were performed on the same gender-balanced test set (3,800 male-donor cells and 3,800 female-donor cells).

Gender	Regression Model	R^2	p-value (population)	Slope (population)	p-value (gender)	Slope (gender)
Female	Population-only	0.0125	**0.00372**	0.0406	–	–
	Population + Gender	0.0113	**0.00628**	0.0362	0.752	-0.046
Male	Population-only	0.0015	0.30600	0.0158	–	–
	Population + Gender	0.0030	0.19700	0.0210	0.487	0.105
Mix	Population-only	0.0000	0.90900	0.0018	–	–
	Population + Gender	0.0016	0.90700	0.0019	0.295	0.147

Table 1: Summary of regression analyses in the AIDA dataset for correlation of confidence gap to cell-type percentages and gender. Gender influences scGPT performance indirectly through gender-specific data imbalance in cell-type representation.

Males tend to have higher frequencies of blood monocytes, whereas females often demonstrate enhanced adaptive and inflammatory signaling (Huang et al. 2021).

In line with these, we observed that $CD14^+$ monocytes tend to be more abundant in males, and $CD4^+$ T cell tend to be more abundant in females (albeit there is slight stochasticity due to bootstrapping during sample split). We also noted substantial differences in natural killer (NK) cells, consistent with other studies showing that males, generally exhibit higher NK cell percentages, likely influenced by hormonal regulation (Klein and Flanagan 2016).

For B cells, the observed differences were modest but non-trivial, mirroring large cohort studies noting gender-linked variations in B-cell biology (Gheitasi et al. 2025).

We would like to emphasize that although Red blood cells (RBCs) show the largest and most consistent drop in confidence in all configurations, RBCs are not part of intended immune populations and their low confidence likely stems from sampling artifacts rather than a gender imbalance.

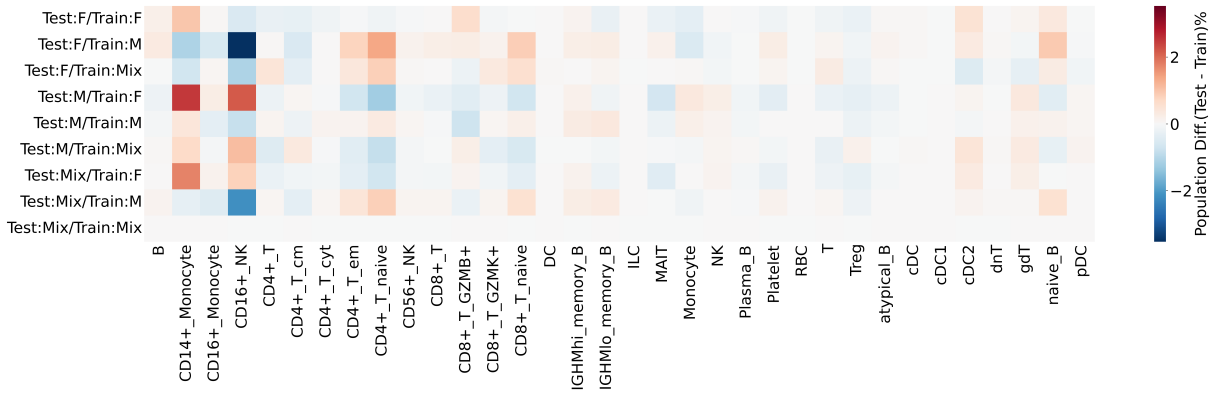
We estimated the average confidence gaps for individual cell-type in various fine-tuning and testing combinations, and visualized the results in a heatmap Fig. 2(b). Lighter shades indicate higher confidence, while darker shades indicate larger negative gaps, and hence poor model performance.

Comparison of the two heatmaps reaffirm the modest but significant correlation between confidence gap and data imbalance, as observed in Fig 1. For example, cDC1 populations are perfectly balanced in training and testing across all experimental settings, and its classification results show no biases. Cell subpopulation of cDC2 show imbalances across experimental settings and show biased classification across settings as well. It is also clear that data imbalance is just one factor for prediction biases, and other factors may include the single-cell expression complexities of some cell-types.

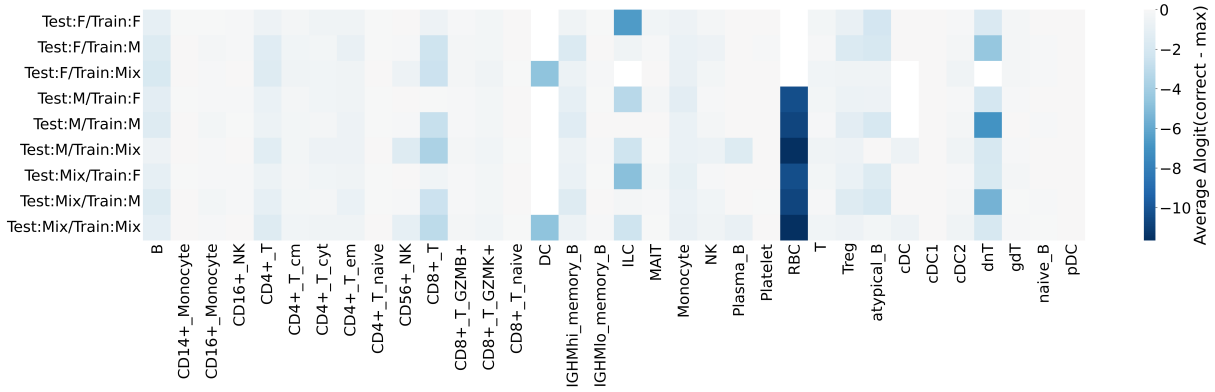
Overall, the heatmaps comparison confirm that prediction biases tend to worsen when training and testing datasets involve misaligned genders.

Regression Analysis of AML

To examine whether previous observations in AIDA would hold in different datasets, we applied the same regression



(a) Heatmap of cell-type population percentage differences between test and training datasets in AIDA. Positive values (red) indicate cell-types more common in the test set; negative values (blue) indicate overrepresentation in the training set.



(b) Heatmap of average confidence gap for each cell-type across gender-specific training and testing combinations in AIDA.

Figure 2: Examining model confidence by cell-type in the AIDA dataset across gender-specific training and testing setups. Note: RBCs are sampling artifacts and are not an indicator of gender-specific data imbalance.

analysis in an independent dataset AML. Figure 3 shows the regression analysis of confidence gaps versus population percentage and gender across female, male, and mixed-tuned models in AML. The top row shows the univariate regression and the bottom row shows the multiple regression results. The results of the regressions were summarized in Table 2.

For the female-tuned model, the confidence gap is negatively correlated with population percentage in both univariate regression (Fig. 3(a)) and multiple regression (Fig. 3(d)), with a p-value of 0.0516 and 0.0367, respectively. The negative slope here is opposite to the positive slope in AIDA, suggesting that as the population percentage increases, the confidence gap ($\Delta\ell$) decreases (becomes more negative), i.e., the model is less confident for more prevalent cell-types.

For the male-tuned model, both the univariate regression (Fig. 3(b)) and multiple regression (Fig. 3(e)) show that confidence gaps positively correlate with cell-type population at significant p-values of 0.00454 and 0.00599, respectively. This positive correlation here is consistent with the observation in AIDA dataset.

For the mixed-tuned model, in both univariate regression (Fig. 3(c)) and multiple regression (Fig. 3(f)), confidence gaps are not correlated with cell-type percentage with p-values of 0.884 and 0.777 respectively, which is consistent with the previous observation(AIDA) that mixed-gender fine-tuned model is the least-biased model.

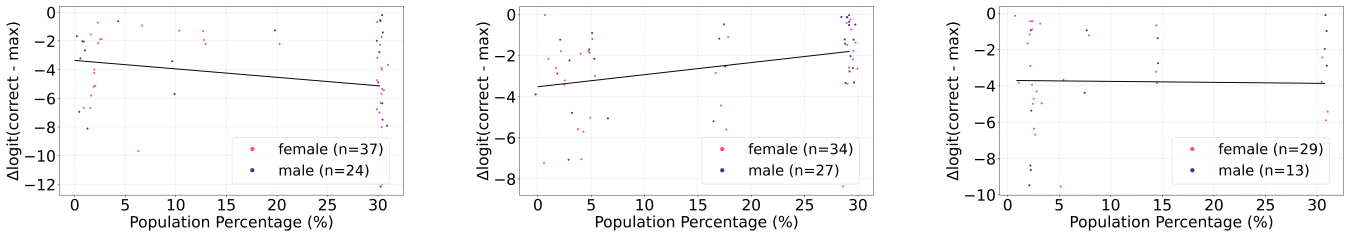
Similar to AIDA, the gender variable was not significantly correlated with confidence gaps, suggesting that gender-specific data imbalance is a key influence on prediction confidence.

Heatmap Analysis of AML

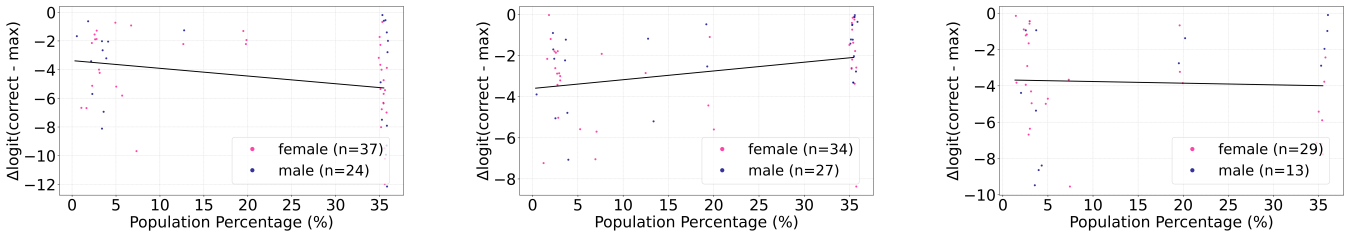
Similar to study in AIDA, we performed heatmap analysis in AML dataset to better understand the overall behavior of the fine-tuned models.

Figure 4(a) shows population percentage differences between test and training datasets.

Several deviations between training and testing datasets



(a) Female-tuned: $\Delta\ell \sim$ population (b) Male-tuned: $\Delta\ell \sim$ population (c) Mixed-tuned: $\Delta\ell \sim$ population



(d) Female-tuned: $\Delta\ell \sim$ pop + gender (e) Male-tuned: $\Delta\ell \sim$ pop + gender (f) Mixed-tuned: $\Delta\ell \sim$ pop + gender

Figure 3: Regression analysis of model performance across female, male, and mixed-trained models in the AML dataset. Jittered scatter plots were used for better visualization. All regressions were performed on the same mixed-gender test set (150 male-donor cells and 146 female-donor cells).

Gender	Regression Model	R^2	p-value (population)	Slope (population)	p-value (gender)	Slope (gender)
Female	Population-only	0.0627	0.05160	-0.0588	–	–
	Population + Gender	0.0729	**0.03670**	-0.0537	0.804	0.199
Male	Population-only	0.1286	**0.00454**	0.0589	–	–
	Population + Gender	0.1382	**0.00599**	0.0456	0.288	0.531
Mix	Population-only	0.0005	0.88400	-0.0053	–	–
	Population + Gender	0.0024	0.77700	-0.0088	0.911	0.106

Table 2: Regression analysis in the AML dataset suggests that gender influences scGPT indirectly through gender-specific imbalance of cell-type subpopulations. Similar to AIDA, the results here show that confidence gaps correlate significantly with cell-type percentage but not with gender itself.

align with previously known gender-specific immune cell subpopulation distributions (Huang et al. 2021; Klein and Flanagan 2016).

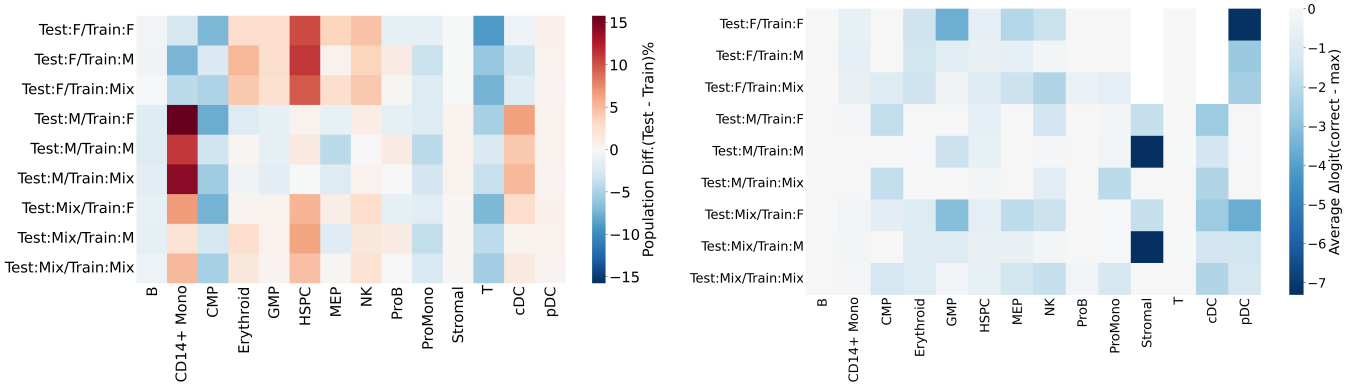
For example, CD14+ monocytes are substantially more common in male test sets (e.g., $+15.8\%$ in Test: male / Train: female; $+14.2\%$ in Test: male / Train: mixed), while HSPCs are enriched in female test sets ($+10.3\%$ to $+11.3\%$). In contrast, T cells and CMPs are consistently less frequent in several configurations (e.g., -9.0% in Test: male / Train: female; -7.8% in Test: female / Train: female).

Figure 4(b) presents the average confidence gap ($\Delta\ell$) for each cell-type across gender-specific training and testing configurations. Darker shades indicate larger confidence loss, while lighter shades reflect stronger model confidence. Similar to previous observation in AIDA, the largest drop occurs in mismatched gender setups, particularly in Test: male / Train: female configuration, indicating that gender mismatch consistently reduces model certainty. In contrast, aligned configurations (e.g., male/male, female/female) perform better, and mixed-gender training leads to more stable predictions across all test sets, suggesting improved generalizability.

Noticeably, Stromal cells and cDC cells show large confidence drops ($\Delta\ell$) in many configurations. MEP cells have nearly perfect prediction in one male-only test dataset in all three models, but struggles in the rest of the test dataset.

Interestingly, under mixed-trained conditions, pDCs and cDCs remain the least confidently predicted subtypes. However, in two of the three male-tested configurations (Test: male / Train: female; Test: male / Train: mixed), pDCs achieve nearly perfect predictions ($\Delta\ell = 0.00$), suggesting that these models capture clearer or more consistent transcriptional patterns for pDCs in male samples, but tend to struggle in the presence of female samples.

In contrast, prediction bias of CD14+ monocytes is moderate, even though they are among the most imbalanced subtype. Similarly, erythroid cells maintain high confidence in male-tested configurations but show declines in female-



(a) Heatmap of cell-type population percentage differences between test and training datasets in AML. Positive values (red) indicate cell-types more common in the test set; negative values (blue) indicate overrepresentation in the training set.

(b) Heatmap of average confidence gap for each cell-type across gender-specific training and testing combinations in AML.

Figure 4: Examining model confidence by cell-type in the AML dataset across gender-specific training and testing setups.

tested ones (e.g., $\Delta\ell = -1.45$ in Test: female / Train: male), indicating gender-specific data imbalance.

Overall, the AML results show that gender-specific data imbalance can influence scGPT prediction confidence. The results also paint a complicated picture. For example, biological complexities in gene expression of particular cell-type very often outweigh simple abundance.

Summary

Using two independent datasets, AIDA and AML, we found that scGPT model prediction confidence could be influenced by gender-specific data imbalance and cell-type complexity.

We found that training data from mixed-gender generally gave the least-biased results. Models trained only on one gender often struggled when tested on the opposite gender, especially for cell-type with gender-specific distributions. In many situations, we found that relatively small difference in cell-type abundance between training and testing sets could lead to pronounced drops in prediction confidence, suggesting that certain cell-type might have complicated gene expression patterns that were sensitive to gender mismatch during training and testing. For example, in the AIDA dataset, double-negative T cells (dnT), innate lymphoid cells (ILCs), dendritic cells (DCs), and CD8+ T cells consistently showed lower confidence, even when trained and tested on matching demographics. Similarly, in the AML dataset, stromal cells, dendritic cells (cDCs and pDCs), and CMPs also led to low-confidence predictions across setups.

Interestingly, pDCs in the AML dataset were almost perfectly predicted in all male-tested settings, even though they were rare. This might suggest that these male-donor cells have stronger or clear expression patterns.

Looking ahead, one way to improve scGPT model performance is to create synthetic cells for underrepresented groups, such as cell-type or gene expression patterns, using generative or statistical models. Prior work in health data shows that augmenting minority groups with synthetic samples can mitigate covariate bias and improve fairness under low–medium bias (Juwara, El-Hussuna, and Emam 2024).

In summary, our results show that gender-specific data imbalance can influence the fine-tuning of scGPT. Having diverse and balanced training data can mitigate some of these gender biases, but further research is required to better address data biases due to demographic factors such as gender.

Acknowledgments

HQ acknowledges the support of a catalyst award from US National Academy of Medicine, US NSF 2525493 and 2200138, a pilot Award from the University of Pennsylvania - Penn Artificial intelligence and Technology Collaboratory for Healthy Aging (PennAITech) with the NIH award P30AG073105, and internal support of the Old Dominion University. DF and HQ acknowledge the support from the National Institutes of Health (NIH) AIM-AHEAD Program Agreement NO. 1OT2OD032581. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

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*.

Bugnon, L. A.; Yones, C.; Milone, D. H.; and Stegmayer, G. 2019. Deep neural architectures for highly imbalanced data in bioinformatics. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8): 2857–2867.

Cheng, Y.; Fan, X.; Zhang, J.; and Li, Y. 2023. A scalable sparse neural network framework for rare cell type annotation of single-cell transcriptome data. *Communications Biology*, 6(1): 545.

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

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

Dittman, D. J.; Khoshgoftaar, T. M.; and Napolitano, A. 2015. The effect of data sampling when using random forest on imbalanced bioinformatics data. In *Proceedings of the 2015 IEEE International Conference on Information Reuse and Integration (IRI)*, 457–463. IEEE.

Escrivà-Font, J.; Cao, T.; and Consiglio, C. R. 2025. Decoding sex differences in human immunity through systems immunology. *Oxford Open Immunology*, 6(1): iqaf006.

Gheitasi, R.; Baumgart, S.; Roell, D.; Rose, N.; Watzl, C.; Dudziak, D.; Andreas, N.; Makarewicz, O.; Drube, S.; Schnizer, C.; Kamradt, T.; Weis, S.; Pletz, M. W.; and study group, C. 2025. Age- and sex-associated differences in immune cell populations. *iScience*, 28(8): 113092.

Hardy, M. A. 1993. *Regression with dummy variables*. 93. Sage.

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.

Jones, A. G.; Connelly, G. G.; Dalapati, T.; Wang, L.; Schott, B. H.; Roman, A. K. S.; and Ko, D. C. 2024. Biological sex affects functional variation across the human genome. *medRxiv*.

Juwara, L.; El-Hussuna, A.; and Emam, K. E. 2024. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns*, 5(4).

Klein, S. L.; and Flanagan, K. L. 2016. Sex differences in immune responses. *Nature Reviews Immunology*, 16(10): 626–638.

Kock, K. H.; Tan, L. M.; Han, K. Y.; Ando, Y.; Jevapatarakul, D.; Chatterjee, A.; Lin, Q. X. X.; Buyamin, E. V.; Sonthalia, R.; Rajagopalan, D.; et al. 2025. Asian diversity in human immune cells. *Cell*, 188(8): 2288–2306.e24.

Pan, L.; Wang, D.; Huang, Q.; et al. 2024. Single-cell landscape of sex differences in the progression of multiple sclerosis. *bioRxiv*.

Shanab, A. A.; and Khoshgoftaar, T. M. 2018. Is gene selection enough for imbalanced bioinformatics data? In *Proceedings of the 2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 346–353.

Szałata, A.; Hrovatin, K.; Becker, S.; Tejada-Lapuerta, A.; Cui, H.; Wang, B.; and Theis, F. J. 2024. Transformers in single-cell omics: a review and new perspectives. *Nature methods*, 21(8): 1430–1443.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 5998–6008.

Weng, J.; Luo, Z.; Li, S.; Sebe, N.; and Zhong, Z. 2023. Logit margin matters: Improving transferable targeted adversarial attack by logit calibration. *IEEE Transactions on Information Forensics and Security*, 18(6): 3561–3574.

Whittle, J.; Meyer, S.; Lacaud, G.; Baker, S. M.; and Iqbal, M. 2024. Single-Cell Atlas of AML Reveals Age-Related Gene Regulatory Networks in t(8;21) AML. *bioRxiv*.

Zelco, A.; and Joshi, A. 2025. Single-cell analysis of sex and gender differences in the human brain during development and disease. *Cellular and Molecular Neurobiology*, 45(1): 20.

Zeng, Y.; Xie, J.; Shangguan, N.; Wei, Z.; Li, W.; Su, Y.; Yang, S.; Zhang, C.; Zhang, J.; Fang, N.; et al. 2025. CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nature Communications*, 16(1): 1–17.