Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Simpler Models Win: A Calibration Benchmark of scGPT for Cell-Type Annotation

Alireza Khosravi
Medical University of Gdańsk
Arshia Khosravi
Medical University of Gdańsk
Kamil Langowski
Medical University of Gdańsk
Michal Sieczczynski
Medical University of Gdańsk
Krzysztof Pastuszak
Medical University of Gdańsk
Anna Supernat
abednarz@gumed.edu.pl
Medical University of Gdańsk
Anna Zaczek
Medical University of Gdańsk

Article

Keywords: single-cell RNA sequencing, large language model, cell-type annotation, calibration, XGBoost, scGPT, Expected Calibration Error, foundation model, benchmarking

Posted Date: July 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10152510/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on November 5th, 2026.
See the published version at https://doi.org/10.1038/s41598-026-77853-3.

When Simpler Models Win: A Calibration Benchmark of scGPT for Cell-Type Annotation

Alireza Khosravi[1,2,3], Arshia Khosravi[3], Kamil Łangowski[1], Michał Sieczczyński[1,5], Krzysztof Pastuszak[1,2,4], Anna Supernat[1,2,*], Anna J. Żaczek[1,2]

[1]Laboratory of Translational Oncology, Medical University of Gdańsk, Gdańsk, Poland
[2]Centre of Biostatistics and Bioinformatics Analysis, Medical University of Gdańsk, Gdańsk, Poland
[3]Scientific Circle of Artificial Intelligence in Medicine, Medical University of Gdańsk, Gdańsk, Poland
[4]Department of Algorithms and System Modeling, Gdańsk University of Technology, Gdańsk, Poland
[5]Department of Software Engineering, Gdańsk University of Technology, Gdańsk, Poland

[*]Corresponding author: Anna Supernat, abednarz@gumed.edu.pl

Abstract

Background. Single-cell large language models (scLLMs) such as scGPT promise transferable representations for cell-type annotation in single-cell RNA sequencing (scRNA-seq), but their statistical calibration under large-scale supervised within-atlas evaluation is largely unexamined. Poorly calibrated confidence scores can mislead downstream biological interpretation, particularly in clinical contexts.

Results. Across six public single-cell atlases (1.3 million cells; triple-negative breast cancer (TNBC), Indonesian peripheral blood mononuclear cells (PBMC), brain, multi-tissue tumour microenvironment, human and pig pancreas) we evaluated frozen scGPT probes (a linear probe and stronger non-linear probes, XGBoost and Random Forest), matched classical baselines (Logistic Regression, XGBoost, Random Forest trained on highly variable genes restricted to the genes present in the scGPT vocabulary), and one end-to-end fine-tuned scGPT experiment on TNBC. Accuracy was reported as Macro F1 with bootstrap 95% confidence intervals; calibration as the Expected Calibration Error (ECE). Classical Logistic Regression on the matched gene set reached Macro F1 = 0.94–0.99 and ECE $\leq$ 0.013 on every atlas, matching or beating every scGPT pipeline. End-to-end fine-tuning of scGPT on TNBC reached Macro F1 = 0.975 / ECE = 0.003, with bootstrap confidence intervals overlapping those of scGPT + XGBoost (0.972) and below the matched-gene Logistic Regression (0.993).

Conclusions. Classical machine learning matched or outperformed every frozen scGPT pipeline we tested on accuracy, and the matched-gene Logistic Regression baseline also exceeded the TNBC end-to-end fine-tuning experiment. On calibration, the matched-gene classical baselines were well-calibrated on every atlas, whereas the standard scGPT-embedding+LogReg pipeline was substantially miscalibrated. Calibration metrics should be mandatory in scLLM benchmarking, and classical baselines should be evaluated under matched downstream learners and matched gene sets before any claim of scLLM superiority.

Keywords: single-cell RNA sequencing; large language model; cell-type annotation; calibration; XGBoost; scGPT; Expected Calibration Error; foundation model; benchmarking

1 Background

scRNA-seq has fundamentally transformed our ability to resolve cellular heterogeneity within complex tissues, enabling the construction of comprehensive cell atlases such as the Human Cell Atlas [1] and Tabula Sapiens [2], and the identification of rare cell populations with unprecedented resolution [3–5]. A central challenge in scRNA-seq analysis is the accurate annotation of cell types from high-dimensional gene expression profiles—a task that is critical for downstream biological interpretation yet remains labour-intensive and difficult to scale.

The dominant approach to cell-type annotation has historically relied on marker-gene-based methods, differential expression analysis, or supervised classifiers trained on labelled reference datasets [6–11]. Tools such as CellTypist [12], SingleR [6], scmap [7], scPred [8], Azimuth [9], and scVI-based annotation [13] have demonstrated that logistic regression, nearest-neighbour reference projection, support-vector machines, and variational autoencoders can all achieve high accuracy on large multi-tissue atlases [10, 11]. However, the field has increasingly adopted a new paradigm: scLLMs pre-trained on tens of millions of cells.

Inspired by successes in natural language processing, scLLMs such as scGPT [14], Geneformer [15], scBERT [16], scFoundation [17], and UCE [18] have been proposed as general-purpose foundation models capable of zero-shot and few-shot cell-type annotation. The central claim is that pre-training on vast corpora of transcriptomic data enables these models to learn transferable biological representations that surpass task-specific classifiers—much as GPT-4 has been reported to outperform task-specific NLP models on many benchmarks [19].

Despite the proliferation of scLLMs, there are two critical and largely overlooked dimensions of their evaluation. First, published benchmarks overwhelmingly focus on peak accuracy metrics (e.g., overall accuracy or weighted F1 score) without assessing *macro F1*—which is essential when cell-type class imbalance is severe, as is universal in complex atlases. Second, and more importantly, the statistical *calibration* of scLLM predictions has received almost no attention. Calibration measures whether a model's confidence scores are reliable: a well-calibrated model that assigns 80% confidence to a prediction should be correct approximately 80% of the time. Miscalibration is a form of systematic error that is invisible to accuracy-only evaluation yet has profound consequences for clinical and translational applications, where confidence scores that do not track empirical accuracy can mislead any downstream workflow that relies on them to triage or weight predictions [20, 21].

A recent independent study [22] reported that scLLMs including scGPT perform poorly in zero-shot and frozen-representation settings, corroborating concerns about generalisation. However, neither that nor any other published study has specifically examined scLLM miscalibration and its consequences across diverse, large-scale multi-tissue atlases.

We chose scGPT as the representative scLLM for this benchmark for several reasons. First, scGPT is the most widely cited and adopted single-cell foundation model, published in *Nature Methods* [14], and has become the de facto reference point for the field. Second, among the available scLLMs, scGPT is the only model that provides a publicly available, fully pre-trained whole-human model with an accessible Python API supporting zero-shot embedding extraction, linear probing, and embedding-based gene-network inference in a unified framework. Third, scGPT was pre-trained on the largest corpus (approximately 33 million cells from CELLxGENE),

making it the strongest candidate to test whether massive pre-training scale translates to annotation superiority. Alternative models such as Geneformer [15] focus on rank-value encoding and network biology rather than expression-level annotation, scBERT [16] was designed for supervised fine-tuning rather than zero-shot transfer, scFoundation [17] at the time of this study did not provide comparable zero-shot linear probing support, and UCE [18] remained a preprint. By selecting the most prominent and best-equipped scLLM, our benchmark provides a strong test case for the scLLM paradigm. Throughout, gene regulatory network (GRN) refers to a graph in which nodes are genes and edges quantify inferred regulatory or co-expression relationships.

To address these gaps, we conducted a rigorous benchmarking study with the following aims: (1) compare a frozen-representation linear-probing pipeline built on scGPT against XGBoost, Random Forest, and Logistic Regression across six large, publicly available single-cell atlases totalling over 1.3 million cells; (2) evaluate both predictive performance (Macro F1) and calibration (ECE) of the resulting end-to-end classification pipelines; and (3) provide a balanced assessment by analysing scGPT's embedding-derived gene co-expression networks—a genuinely zero-shot, label-free use of the model—to delineate where scLLMs may retain value.

Our study provides, to our knowledge, a focused calibration benchmark of scGPT-based annotation pipelines at this scale, foregrounds calibration as an underreported axis of scLLM evaluation, and offers practical guidance for the field regarding the appropriate deployment of these resource-intensive models.

2 Materials and Methods

2.1 Datasets

We selected six publicly available single-cell atlases from the Chan Zuckerberg CELLxGENE Discover portal [23] to represent diverse biological contexts, species, and levels of cellular complexity (Table 1; full accession URLs are listed in the Availability of data and materials section). The datasets were not chosen at random: they were specifically curated to span axes that are known to challenge the generalisation of single-cell models. Specifically, they collectively cover (i) two contrasting tissue compartments—solid tumour microenvironments (TNBC, multi-tissue TME) and circulating immune cells (Indonesia PBMC); (ii) high-resolution multi-region neural tissue (Brain Atlas), where transcriptomic similarity between cell subtypes is high; (iii) a metabolically and developmentally distinct exocrine/endocrine organ (Human Pancreas), used historically as a reference benchmark in single-cell methods; (iv) a cross-species control (Pig Pancreas) entirely outside scGPT's primarily human pre-training corpus, which probes representation transfer beyond the training distribution; and (v) a population-level human cohort from a non-European ancestry (Indonesia PBMC), addressing a known under-representation in foundation-model training data. The datasets collectively span approximately 1.3 million cells and include:

1. **TNBC Breast Cancer** (427,823 cells): A large-scale single-cell atlas of triple-negative breast cancer tumour microenvironments.

2. **Indonesia PBMC** (462,034 cells): Peripheral blood mononuclear cells from a genetically diverse Indonesian cohort.

3. **Brain Atlas** (75,583 cells): A multi-region human brain single-cell atlas.
4. **Multi-Tissue TME** (391,963 cells): A pan-cancer tumour microenvironment atlas spanning multiple tissue types.
5. **Human Pancreas** (26,474 cells): A reference atlas of the human pancreatic cell landscape.
6. **Pig Pancreas** (22,056 cells): A cross-species porcine pancreatic atlas enabling evaluation of model generalisation across species boundaries.

All datasets were downloaded in `.h5ad` format. Cell types were defined by the curated `cell_type` field present in each dataset's observation metadata. Cell types represented by five or fewer cells were excluded from analysis (equivalently, only cell types with more than five cells were retained, matching the convention in Table 1): this filter removes a small number of singleton or near-singleton labels that would otherwise prevent a stratified train–test split from preserving class membership in both folds, and stabilises bootstrap resampling on the test set.

Table 1: Summary of benchmark datasets. All datasets were obtained from the CZ CELLxGENE Discover portal.

Dataset	Tissue / Context	Cells	Cell Types	Species
TNBC Breast Cancer	Tumour microenvironment	427,823	7	*H. sapiens*
Indonesia PBMC	Peripheral blood	462,034	13	*H. sapiens*
Brain Atlas	Multiple brain regions	75,583	10	*H. sapiens*
Multi-Tissue TME	Pan-cancer TME	391,963	11	*H. sapiens*
Human Pancreas	Exocrine/endocrine	26,474	4	*H. sapiens*
Pig Pancreas	Exocrine/endocrine	22,056	4	*S. scrofa*

TME = tumour microenvironment. Cell-type counts are reported after the cell-type filter described in the text (cell types with >5 cells are retained).

2.2 Data Preprocessing

All datasets underwent a unified preprocessing pipeline implemented in Python using Scanpy v1.9 [24]. For both the scGPT embedding pipeline and the classical ML feature extraction, raw count matrices were used as input. Preprocessing steps were applied consistently:

1. **Gene filtering**: Lowly expressed genes were removed using `sc.pp.filter_genes` with `min_counts=3`, which discards genes whose total counts summed across all cells are less than three. We adopt this count-based threshold (rather than the alternative `min_cells` criterion) because total counts are independent of dataset size, so the same cut-off behaves comparably across atlases of very different cell numbers.
2. **Library size normalisation**: Each cell's counts were normalised to a total of 10,000 counts per cell (`sc.pp.normalize_total`, `target_sum=1e4`).
3. **Log-transformation**: A pseudocount of one was added before `log1p` transformation (`sc.pp.log1p`). The pseudocount avoids $\log 0$ on zero-expression entries, which dominate scRNA-seq matrices, and is the standard convention adopted by Scanpy and Seurat.

4. **Highly variable gene (HVG) selection**: The top 1,200 HVGs were selected separately for each downstream consumer (see clarification below).

HVG selection per branch. HVG selection was performed on the full dataset prior to the train–test split, separately for the scGPT branch (`flavor="seurat_v3"`, applied to the raw count matrix) and for the classical-ML branch (`flavor="seurat"`, applied to the normalised log-transformed matrix). Each flavour is the standard recommendation for its respective downstream consumer: `seurat_v3` expects raw counts and is the preprocessing recipe followed by the official scGPT tutorials, whereas `seurat` is the classical Scanpy default for log-normalised matrices. The two pipelines therefore operate on *different* top-1,200 HVG sets—each prepared in the manner intended for the corresponding model family. We discuss the implications of this design choice—and of performing HVG selection prior to the train–test split—in the Limitations section.

2.3 Train–Test Split

For each dataset, cell labels were encoded as integer category codes. A stratified 80/20 train–test split was performed using `train_test_split` from scikit-learn [25] with a fixed random seed of 42, ensuring that the class distribution was preserved across splits and results were reproducible.

2.4 scGPT Embedding Extraction with Frozen-Representation Linear Probing

We used the publicly available scGPT whole-human pre-trained model weights. The scGPT architecture is a transformer-based model pre-trained on approximately 33 million human single-cell transcriptomes from the CELLxGENE corpus [14]. Model loading was performed using the `scgpt` Python package. We employed a custom vocabulary loader to bypass a C++ library dependency (`torchtext`) incompatible with the cloud GPU environment.

The scGPT embedding pipeline proceeds as follows:

1. HVG selection (1,200 genes, Seurat v3) was applied to each dataset.
2. Gene expression values were binned into 51 discrete value bins (`N_BINS` = 51) using `np.digitize`. Binning was performed *per cell*: for each cell, $X_{\max}$ is that cell's own maximum (log-normalised) expression value and the 51 bin edges are placed uniformly on $[0, X_{\max}]$, so the binning is independent across cells and across datasets. This follows the value-binning recipe of the official scGPT tutorials [14].
3. Gene identifiers were mapped to scGPT's vocabulary (60,697 tokens, of which 60,694 are unique gene symbols and 3 are special tokens `<pad>`, `<cls>`, `<mask>`); genes absent from the vocabulary were masked with the `<pad>` token. The proportion of HVGs successfully matched to the vocabulary ranged from 743 to 1,142 genes per dataset (62–95%).
4. A `<cls>` token was prepended to the gene sequence for each cell.
5. Embeddings were extracted in batches of 1,024 cells using the model's `_encode` method, with all computations performed on an NVIDIA H100 80GB GPU (CUDA).
6. The `<cls>` token embedding (dimension 512) was used as the cell representation.

Linear probing was performed by fitting a Logistic Regression classifier (`max_iter=1000`) on the frozen scGPT embeddings of the training set and evaluating on the test set. We use the term *frozen-representation linear probing* (sometimes also called *linear evaluation*) throughout the manuscript to describe this protocol: scGPT itself receives no gradient updates, but a supervised linear classifier is fit on the labelled cells of the target dataset. Strictly speaking, this is therefore not a label-free zero-shot annotation evaluation; it isolates the discriminability of scGPT's pre-trained representations under a minimally expressive downstream head, which is the standard protocol for evaluating self-supervised representations.

Gene-vocabulary considerations. A potential confounder in comparing scGPT against classical ML on HVGs is gene-set asymmetry: scGPT operates on a fixed pre-trained vocabulary of ∼60,694 unique gene symbols and silently masks any HVG absent from this vocabulary, whereas the classical baselines see the full HVG matrix selected for them. The reported per-dataset vocabulary-matching rate (62–95%; see step 3 above) quantifies the resulting feature loss for scGPT relative to its own 1,200-gene seurat_v3 list. To address this asymmetry directly, we ran a vocabulary-matched fairness control in which the classical learners are restricted to exactly the genes that scGPT actually consumes (Section 2.8, Section 3.5).

Cross-species gene handling (Pig Pancreas). The scGPT vocabulary is human. For the Pig Pancreas atlas we deliberately did *not* perform ortholog mapping: porcine gene symbols were matched directly against the human vocabulary (with an uppercase fallback), and any symbol still absent was masked with the `<pad>` token. This direct-matching choice is itself part of the intended out-of-distribution stress test—it measures what a practitioner obtains by applying the released human scGPT model to non-human data without curation. Consequently, the Pig Pancreas scGPT result should be read as a lower bound on cross-species transfer rather than as an optimised cross-species pipeline; a dedicated ortholog-mapped evaluation is left to future work.

2.5 Classical Machine Learning Methods

Three classical ML classifiers were trained on the top 1,200 HVGs (Seurat method) extracted from normalised, log-transformed count matrices:

- **XGBoost** [26]: Gradient-boosted decision trees with `n_estimators=200`, `max_depth=6`, `learning_rate=0.1`, `objective="multi:softprob"`, `eval_metric="mlogloss"`, `tree_method="hist"`, `random_state=42`, and parallel execution (`n_jobs=-1`).
- **Random Forest** [27]: Ensemble of `n_estimators=200` decision trees with `random_state=42`, all other hyperparameters left at the scikit-learn defaults (Gini criterion, no maximum depth, bootstrap sampling), and parallel execution (`n_jobs=-1`).
- **Logistic Regression (Raw)** [25]: Multinomial L2-regularised logistic regression with `solver="lbfgs"`, `penalty="l2"`, `C=1.0`, `class_weight=None`, `multi_class="auto"`, and `max_iter=1000`; convergence at this iteration cap was confirmed by the absence of scikit-learn non-convergence warnings on all six atlases. To clarify a possible ambiguity: the label

"Raw" refers to the input *feature space* being the raw HVG *matrix* (i.e. no PCA, embedding, or other dimensionality reduction); the matrix itself is normalised and log-transformed exactly as for the other classical learners (Section "Data Preprocessing" above). It is *not* fit on raw count values.

All three classical learners receive the same Seurat-flavoured top-1,200 HVG matrix (normalised, log-transformed). Implementations used scikit-learn [25] and XGBoost [26] with the hyperparameters listed above, deliberately avoiding extensive hyperparameter tuning to reflect out-of-the-box usability—the realistic benchmark condition for most research groups. A fixed seed (`random_state=42`) was used throughout for reproducibility.

2.6 Evaluation Metrics

2.6.1 Macro F1 Score

Predictive performance was assessed using the macro-averaged F1 score (Macro F1), computed as the unweighted mean of per-class F1 scores across all cell types. For a classification task with C classes, letting P_c and R_c denote precision and recall on class c,

$$\mathrm{F1}_c = \frac{2 \cdot \mathrm{P}_c \cdot \mathrm{R}_c}{\mathrm{P}_c + \mathrm{R}_c}, \qquad \text{Macro F1} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{F1}_c. \tag{1}$$

This metric treats each cell type equally regardless of class frequency, making it appropriate for the highly imbalanced class distributions typical of complex single-cell atlases. Macro F1 was computed using `sklearn.metrics.f1_score` with `average='macro'`.

2.6.2 Expected Calibration Error (ECE)

Statistical calibration was assessed using the Expected Calibration Error (ECE) [20], defined as:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right| \tag{2}$$

where $M = 10$ confidence bins partition the interval $[0, 1]$, B_m is the set of samples whose maximum predicted probability falls in bin m, n is the total number of test samples, $\mathrm{acc}(B_m)$ is the mean accuracy within the bin, and $\mathrm{conf}(B_m)$ is the mean confidence within the bin. Lower ECE indicates better calibration. ECE was computed from the model's predicted probability vectors (`predict_proba` output) using a custom implementation.

A perfectly calibrated model has ECE = 0. A model with ECE > 0 and $\mathrm{conf}(B_m) > \mathrm{acc}(B_m)$ is *overconfident*: it assigns high probabilities to predictions that are frequently incorrect. **Sign convention.** In the reliability diagrams (Section 3.7) we plot empirical accuracy on the vertical axis against mean predicted confidence on the horizontal axis; bins lying *below* the identity diagonal therefore have $\mathrm{conf}(B_m) > \mathrm{acc}(B_m)$ and indicate overconfidence, whereas bins above the diagonal indicate underconfidence. Because ECE is an unsigned summary, we do not infer the direction of miscalibration from the scalar ECE; we read direction, where relevant, only at the bin level from the reliability diagrams themselves.

We additionally note that ECE is one of several standard summaries of probability calibration, and that it is a non-proper summary statistic rather than a scoring rule. Two complementary *strictly proper* scoring rules are also commonly reported: the *Brier score* [20], a quadratic rule equal to the mean squared error between the predicted probability vector and the one-hot true label; and the *negative log-likelihood* (NLL), a logarithmic rule that penalises confident errors far more steeply than the Brier score and is therefore especially sensitive to extreme overconfidence. These two rules are complementary rather than equivalent—they weight the same errors differently—and a full calibration audit would report all three together. We report ECE as the primary metric because it is the metric currently in use in the scLLM literature [20, 22] and because it underwrites the bin-level reliability-diagram interpretation that motivates our calibration analysis. Because per-cell predicted-probability arrays were not retained from the original runs, the scalar Brier and NLL values, together with signed and class-wise ECE decompositions, are left to future work; we therefore restrict our claims to the *magnitude* of miscalibration captured by ECE and the reliability diagrams.

2.7 Gene Regulatory Network Analysis

To assess the biological utility of scGPT representations independently of annotation performance, we performed gene regulatory network (GRN) inference on a single representative dataset (TNBC Breast Cancer). Specifically, we extracted gene-token embeddings from the frozen scGPT encoder for the matched-vocabulary HVG set, used the `scgpt.tasks.GeneEmbedding` utility to cluster genes into metagene modules at resolution 40, and computed pairwise cosine similarities between genes within the largest metagene module. The resulting weighted similarity graph was rendered with NetworkX, retaining edges with similarity below 0.99 to suppress trivial self-similar pairs. We emphasise that this GRN analysis is genuinely label-free and zero-shot in the strict sense: no downstream classifier is involved and the encoder weights are not updated—only the frozen embeddings of the pre-trained model are used. We emphasise that this analysis is therefore an *embedding-derived gene co-expression similarity* analysis—it uses cosine similarity between the model's gene-token embeddings—and not an attention-weight extraction; we use the term "GRN" in the loose sense of an inferred gene–gene similarity graph (defined above) rather than a causal regulatory network. Inspired by prior work demonstrating the utility of tree-based and similarity-based methods for regulatory-network reconstruction [28], this analysis evaluates whether scGPT's pre-trained gene representations recover known co-regulatory structure from pre-training alone. We did not benchmark the recovered modules against a null or baseline graph (e.g. random gene sets, a PCA-derived similarity graph, or a raw-expression Pearson-correlation graph); the analysis is intended as a qualitative demonstration that biologically coherent structure survives in the frozen embeddings, and we flag the absence of a quantitative null comparison in the Limitations section.

2.8 Fairness Controls and End-to-End Fine-Tuning of scGPT

To pre-empt two natural objections to the headline benchmark—namely that pairing scGPT embeddings with logistic regression alone may understate the embeddings' true discriminability, and that classical baselines see HVGs absent from scGPT's vocabulary—we ran two complementary

control experiments on all six atlases:

- **Stronger downstream probes on scGPT embeddings.** The frozen scGPT `<cls>` embeddings were re-used as input to two stronger non-linear probes: XGBoost and Random Forest (default scikit-learn / XGBoost hyperparameters, identical to the classical-baseline configurations described above).
- **Vocabulary-matched classical baselines.** Each of the three classical learners (Logistic Regression, Random Forest, XGBoost) was re-trained on a feature matrix restricted to exactly the genes that scGPT actually consumed: namely, the intersection of the seurat_v3 top-1,200 HVG list with scGPT's pre-trained vocabulary. This matches the classical input gene set to scGPT's effective input gene set.

All Macro F1 values from the controls are reported with bootstrap 95% confidence intervals (1,000 resamples of the test set, percentile method); the results appear in Section 3.5.

To additionally quantify the residual gap left by all of the above, we performed a single end-to-end fine-tuning experiment on the TNBC atlas, in which scGPT's encoder weights are themselves updated using the target labels. A two-layer MLP classification head (`Linear` → `GELU` → `Dropout`(0.2) → `Linear`; hidden dimension 256) was appended to the `<cls>` embedding, and the entire encoder was trained with AdamW (learning rate 1×10^{-4}, weight decay 0.01), mixed-precision (`torch.cuda.amp`), batch size 64, 8 epochs, a cosine learning-rate schedule with 5% warm-up, and label-smoothing cross-entropy ($\epsilon = 0.05$). The full TNBC training partition was used for fine-tuning and the held-out 20% test set (85,565 cells) was used exclusively for evaluation. Total wall-clock time was approximately 30 minutes on a single NVIDIA H100. Results appear in Section 3.6. The reference implementation is provided in `src/scgpt_vs_classical/finetune.py` of the accompanying repository.

2.9 Computational Environment

All computations were performed on a cloud-based NVIDIA H100 80GB HBM3 GPU instance (RunPod). Total GPU compute time across all six datasets was approximately four hours. Data download, preprocessing, and result aggregation were performed using Python 3.10, PyTorch 2.x, scGPT v0.2.1, Scanpy 1.9, scikit-learn 1.3, and XGBoost 2.0. The total estimated cloud compute cost was approximately USD 15. All code is available via the repository linked in the Data Availability section.

3 Results

3.1 Classical Machine Learning Consistently and Substantially Outperforms the scGPT Linear-Probe Pipeline

We benchmarked four approaches across six diverse single-cell atlases comprising 1.3 million cells. The results are summarised in Figure 1 and Table 2. We emphasise that "scGPT" in the figures and Table 2 refers throughout to the end-to-end pipeline of frozen scGPT embeddings followed by a logistic-regression head; the comparison should therefore be read as *scGPT-embedding+LogReg* versus the classical baselines. A note on protocol: Table 2 reports the

standard scGPT linear-probe protocol following the reference implementation of Cui et al. [14], whereas the fairness-control results in Section 3.5 (Table 3) re-fit a plain `sklearn` LogReg on the same frozen embeddings under identical preprocessing as the vocabulary-matched controls. The two protocols yield slightly different point estimates for the scGPT + LogReg pipeline on TNBC (0.783 vs. 0.715) and Pig Pancreas (0.231 vs. 0.270) but the same qualitative conclusion—scGPT is dominated by the classical matched-vocabulary baseline—on every atlas. Table 2 should therefore be read as the initial reference-protocol benchmark; the primary controlled comparison is the fairness-control analysis (Table 3), in which downstream learners and input gene sets are matched across all models.

Overall performance gap. The performance gap between the scGPT pipeline and the classical methods is large, consistent, and dataset-invariant (Figure 1a). Across all six atlases, Logistic Regression on raw HVGs achieved Macro F1 scores of 0.940–0.996, XGBoost achieved 0.932–0.997, and Random Forest achieved 0.883–0.995. In contrast, the scGPT linear-probe pipeline achieved Macro F1 scores of only 0.231–0.783—a gap of 0.21–0.76 F1 points relative to the best classical method on each dataset.

Dataset-specific observations. The performance deficiency of the scGPT pipeline was most severe on the Pig Pancreas dataset (Macro F1 = 0.231), which is a cross-species dataset entirely outside scGPT's pre-training distribution. The relatively highest scGPT-pipeline performance was observed on TNBC Breast Cancer (Macro F1 = 0.783), a human oncology dataset, yet even here it was outperformed by all three classical methods (Logistic Regression: 0.996, XGBoost: 0.997, Random Forest: 0.995).

Distribution of scores. The boxplot in Figure 1b illustrates the tight clustering of classical methods (Macro F1 median > 0.93 for all three) contrasted with the wide dispersion and low median of the scGPT pipeline (median ≈ 0.49). The inter-quartile range of scGPT-pipeline scores spans approximately 0.23–0.78, reflecting high variability across dataset types.

3.2 The scGPT Linear-Probe Pipeline Exhibits Significant Miscalibration

A key novel finding of our study is the systematic miscalibration of the scGPT-embedding+LogReg pipeline (Figure 1c; Table 2). Across all datasets, this pipeline exhibited ECE values of 0.038–0.177, indicating that its predicted confidence scores are substantially inconsistent with its actual accuracy. Specifically, the scGPT-embedding+LogReg pipeline is substantially miscalibrated: its predicted probabilities do not reliably track empirical accuracy. Because ECE is unsigned, we interpret the direction of this miscalibration only at the bin level in the reliability diagrams (Section 3.7). We note that the predicted probabilities are produced by the downstream logistic-regression head fit on scGPT embeddings, not by scGPT itself; the calibration claim therefore concerns the end-to-end pipeline, not the foundation model in isolation.

Accuracy–Calibration trade-off. The scatter plot in Figure 1c reveals a clear separation between models. Classical methods occupy the top-right "Optimal Performance" region, characterised by both high Macro F1 (> 0.88) and very low ECE (< 0.03 for XGBoost and Logistic

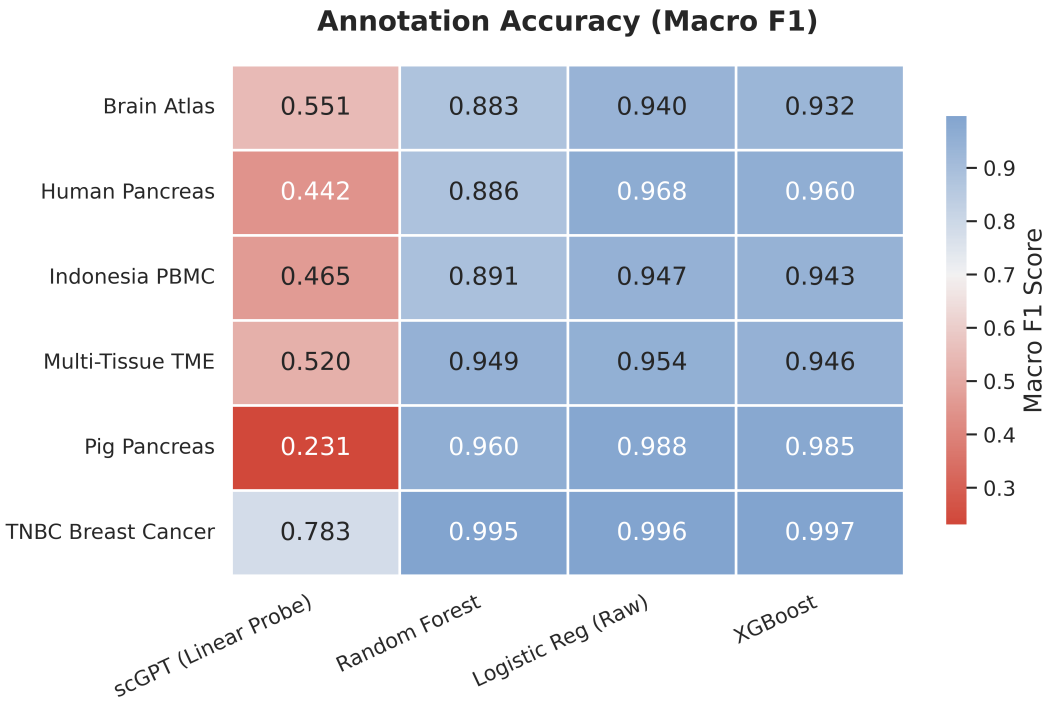
Regression). scGPT-pipeline points are scattered in the lower-left quadrant, combining poor accuracy with poor calibration—the worst possible combination. Notably, XGBoost and Logistic Regression achieve very low ECE (0.001–0.008), while Random Forest is moderately well calibrated (ECE 0.02–0.10).

Consequences of miscalibration at the cell-type level. The substantial miscalibration of the scGPT linear-probe pipeline means its confidence scores cannot be trusted as reliability estimates. A downstream workflow that thresholds on predicted confidence—for example, to accept high-confidence annotations automatically while routing uncertain ones to expert review—will be misled by these scores in exactly the regime where such workflows rely on them most, and may either propagate misclassifications of tumour-infiltrating lymphocytes versus malignant cells or needlessly discard correct calls. This risk is compounded by the fact that ECE is not routinely reported in scLLM publications.

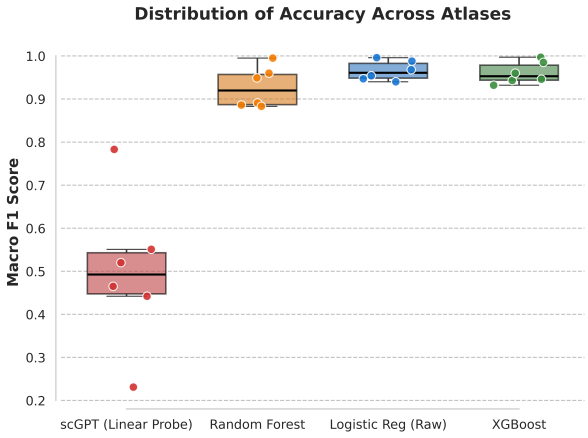
Table 2: Benchmarking results: Macro F1 score (higher is better) and Expected Calibration Error (ECE; lower is better) for all models across all six datasets. Bold values indicate the best per-metric value per dataset.

Dataset	**Model**	**Macro F1**	**ECE**
TNBC Breast Cancer	scGPT (Linear Probe)	0.783	0.038
	Random Forest	0.995	0.023
	Logistic Reg (Raw)	**0.996**	**0.001**
	XGBoost	0.997	0.000
Indonesia PBMC	scGPT (Linear Probe)	0.465	0.125
	Random Forest	0.891	0.102
	Logistic Reg (Raw)	**0.947**	0.006
	XGBoost	0.943	**0.006**
Brain Atlas	scGPT (Linear Probe)	0.551	0.081
	Random Forest	0.883	0.031
	XGBoost	0.932	**0.007**
	Logistic Reg (Raw)	**0.940**	0.008
Multi-Tissue TME	scGPT (Linear Probe)	0.520	0.177
	XGBoost	0.946	**0.002**
	Logistic Reg (Raw)	**0.954**	0.004
	Random Forest	0.949	0.081
Human Pancreas	scGPT (Linear Probe)	0.442	0.105
	Random Forest	0.886	0.054
	XGBoost	0.960	**0.002**
	Logistic Reg (Raw)	**0.968**	0.003
Pig Pancreas	scGPT (Linear Probe)	0.231	0.108
	Random Forest	0.960	0.036
	XGBoost	0.985	**0.001**
	Logistic Reg (Raw)	**0.988**	0.002

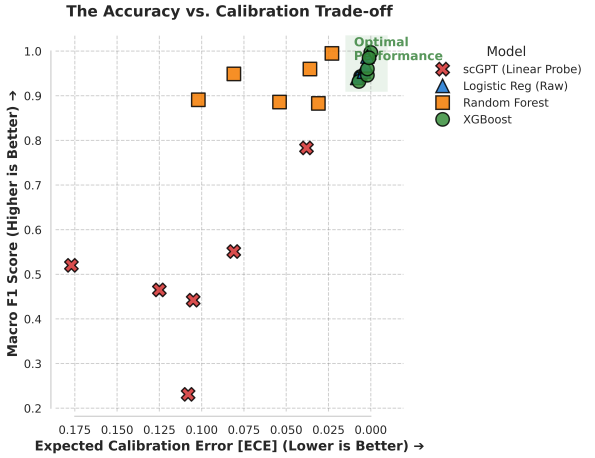
3.3 Figures



(a) Annotation accuracy (Macro F1) heatmap across all six datasets and four methods. Red shading indicates poor performance; blue shading indicates very high performance. scGPT (leftmost column) is consistently the lowest performer across all atlases.



(b) Distribution of Macro F1 scores across atlases. Classical methods cluster tightly above 0.88; scGPT shows wide dispersion with a median near 0.49.



(c) Accuracy–calibration trade-off scatter plot. The green shaded "Optimal Performance" zone (high F1, low ECE) is occupied exclusively by classical methods. scGPT points are dispersed in the lower-left (low accuracy, high ECE) region.

Figure 1: **Benchmarking scGPT vs. classical ML methods across six large single-cell atlases.** (A) Heatmap of Macro F1 scores per dataset and method. (B) Boxplot distribution of Macro F1 scores. (C) Accuracy vs. Expected Calibration Error (ECE) trade-off. XGBoost and Logistic Regression achieve near-perfect accuracy and near-perfect calibration simultaneously. scGPT consistently fails on both dimensions, with the exception of a partial recovery on the TNBC dataset.

3.4 scGPT Gene-Token Embeddings Capture Biologically Meaningful Gene–Gene Similarity Structure

To provide a balanced perspective, we examined whether scGPT's poor annotation performance reflects a complete failure of its representational capacity. Despite the model's weak performance under the frozen-weights linear-probe regime, its frozen gene-token embeddings encode biologically interpretable gene–gene similarity structure.

Using an embedding-derived zero-shot gene–gene similarity analysis, we computed pairwise cosine similarities between frozen scGPT gene-token embeddings on the TNBC dataset. The top-ranked interactions recapitulated known regulatory and co-expression relationships from the immune compartment of the breast tumour microenvironment. Concretely, the largest metagene module recovered (i) interactions among the major histocompatibility complex class II co-expression cluster (*HLA-DRA*, *HLA-DRB1*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, *CD74*), known to be co-regulated by the master MHC-II transactivator *CIITA*; (ii) the cytotoxic T-/NK-cell co-expression programme (*CD8A*, *CD8B*, *GZMA*, *GZMB*, *GZMK*, *PRF1*, *NKG7*, *GNLY*); and (iii) the T-cell receptor / TCR-signalling complex (*CD3D*, *CD3E*, *CD3G*, *TRAC*, *TRBC1*, *TRBC2*, *LCK*). All three modules are independently documented in the immunology literature as canonical, tightly co-regulated transcriptional programmes: the MHC class II module is co-regulated at the transcriptional level by the master transactivator *CIITA* [29]; the cytotoxic-effector programme (granzymes, perforin, NKG7, GNLY) is the canonical lytic-granule machinery of $CD8^+$ T and NK cells [30]; and the CD3/TCR/LCK module forms the proximal T-cell-receptor signalling complex [31]. The fact that scGPT's gene-token embeddings recover all three textbook modules on TNBC—without any task-specific fine-tuning—constitutes external, literature-based validation that the model's frozen gene-token embeddings recover known co-expression modules. This is consistent with the hypothesis that scGPT's gene-token embedding space functions as an implicit co-expression detector [28].

These results suggest that while scGPT should *not* be relied upon for definitive cell-type assignment under a frozen-weights linear-probe regime, its pre-trained representations retain value for exploratory analyses such as GRN inference and hypothesis generation. We emphasise that the validation above is literature-based: scGPT recovers *known* co-expression structure on a single dataset, which establishes face validity of the embedding-derived gene-similarity signal but does not, by itself, demonstrate the discovery of novel regulators. Experimental (e.g. in-vitro perturbation) validation of any putatively novel predictions is beyond the scope of this benchmarking study.

3.5 Fairness Controls: Stronger Probes and Vocabulary-Matched Inputs

The fairness-control protocol is described in full in Section 2.8. Briefly, on each of the six atlases we evaluated six models: scGPT embeddings paired with each of LogReg, XGBoost and Random Forest, and each of the three classical learners restricted to the vocabulary-matched HVG subset. All Macro F1 values are reported with bootstrap 95% confidence intervals (Table 3).

Stronger downstream probes on scGPT embeddings. Replacing the logistic-regression head with XGBoost or Random Forest on the same frozen scGPT embeddings produced large and

consistent gains across all six atlases. The improvement was largest on the highest-performing dataset (TNBC: 0.715 → 0.972 with XGBoost) and substantial even on the worst (Pig Pancreas: 0.270 → 0.622 with XGBoost). The probe choice therefore materially affects the absolute numbers, and the original LogReg-only protocol did understate the discriminability of scGPT embeddings. However, even with the strongest non-linear head, scGPT-embedding pipelines still trailed the classical baselines on five of six atlases: the only exception was Multi-Tissue TME, where scGPT + XGBoost (0.896) was effectively tied with the vocabulary-matched LogReg (0.898) under this bootstrap-CI analysis. The qualitative direction of the original benchmark therefore survives a strictly stronger comparison on five of six datasets and ties on one.

Vocabulary-matched fairness control. To test whether the gap is driven by feature-set asymmetry, we re-ran the three classical learners (Logistic Regression, Random Forest, XGBoost) on a feature matrix restricted to exactly the genes that scGPT actually consumed (i.e. HVGs both passing preprocessing and present in scGPT's vocabulary). On this strictly matched feature space, the strongest classical learner per dataset reached Macro F1 scores between 0.898 (Multi-Tissue TME) and 0.993 (TNBC), and outperformed the strongest scGPT-embedding pipeline on every dataset except Multi-Tissue TME, where it tied. On Pig Pancreas, the vocabulary-matched logistic regression alone (Macro F1 = 0.989) outperformed the best scGPT-embedding pipeline (XGBoost on embeddings, 0.622) by 0.37 F1 points; on Human Pancreas it outperformed it by 0.30 F1 points. The performance gap is therefore not explained by gene-vocabulary asymmetry; it persists, and in several cases widens, when classical models are restricted to scGPT's own feature subset.

Calibration in the controls. Across the 36 fairness-control runs, classical models on the vocabulary-matched subset retained near-perfect calibration (ECE typically < 0.015, with two Random Forest exceptions on Indonesia PBMC and Multi-Tissue TME), and scGPT-embedding pipelines paired with XGBoost also achieved good calibration (ECE 0.003–0.034). The systematic miscalibration originally observed for scGPT-embedding+LogReg therefore reflects, at least in part, the choice of downstream head; pairing scGPT embeddings with a calibrated probabilistic learner (XGBoost in particular) recovers calibration on in-distribution data, but does not recover the absolute accuracy gap relative to classical methods on five of six datasets.

Bottom line of the controls. Across all six atlases, classical HVG-based learners on the vocabulary-matched feature space outperform every frozen-scGPT-embedding pipeline (Table 3). The headline conclusion is therefore strengthened: both objections raised against the original protocol have explicit experimental answers across the full benchmark, and neither overturns the result.

Table 3: Fairness-control benchmarks across all six datasets, with bootstrap 95% confidence intervals on Macro F1 (1,000 test-set resamples, percentile method). *Vocabulary-matched HVG subset* = the genes that pass scGPT's preprocessing and are present in scGPT's vocabulary; classical models trained on this restricted feature space. Bold values indicate the best per-metric value per dataset.

Dataset	Model	Macro F1	95% CI	ECE
TNBC Breast Cancer	scGPT emb. + LogReg	0.715	[0.711, 0.720]	0.066
	scGPT emb. + XGBoost	0.972	[0.968, 0.975]	0.002
	scGPT emb. + Random Forest	0.954	[0.949, 0.958]	0.021
	LogReg (vocab-matched)	**0.993**	[0.992, 0.995]	**0.001**
	Random Forest (vocab-matched)	0.993	[0.991, 0.994]	0.021
	XGBoost (vocab-matched)	0.992	[0.990, 0.994]	0.001
Indonesia PBMC	scGPT emb. + LogReg	0.465	[0.458, 0.471]	0.125
	scGPT emb. + XGBoost	0.770	[0.764, 0.776]	0.008
	scGPT emb. + Random Forest	0.677	[0.670, 0.684]	0.116
	LogReg (vocab-matched)	**0.940**	[0.937, 0.943]	**0.006**
	Random Forest (vocab-matched)	0.870	[0.864, 0.876]	0.141
	XGBoost (vocab-matched)	0.937	[0.934, 0.940]	0.011
Brain Atlas	scGPT emb. + LogReg	0.549	[0.545, 0.553]	0.078
	scGPT emb. + XGBoost	0.818	[0.779, 0.851]	0.012
	scGPT emb. + Random Forest	0.759	[0.718, 0.791]	0.016
	LogReg (vocab-matched)	**0.937**	[0.913, 0.957]	**0.005**
	Random Forest (vocab-matched)	0.858	[0.831, 0.882]	0.025
	XGBoost (vocab-matched)	0.932	[0.904, 0.955]	0.007
Multi-Tissue TME	scGPT emb. + LogReg	0.520	[0.518, 0.523]	0.177
	scGPT emb. + XGBoost	0.896	[0.887, 0.903]	0.003
	scGPT emb. + Random Forest	0.858	[0.846, 0.867]	0.063
	LogReg (vocab-matched)	0.898	[0.892, 0.903]	0.013
	Random Forest (vocab-matched)	**0.927**	[0.920, 0.932]	0.045
	XGBoost (vocab-matched)	0.925	[0.917, 0.932]	**0.004**
Human Pancreas	scGPT emb. + LogReg	0.445	[0.441, 0.449]	0.108
	scGPT emb. + XGBoost	0.665	[0.656, 0.674]	0.015
	scGPT emb. + Random Forest	0.524	[0.512, 0.537]	0.091
	LogReg (vocab-matched)	**0.967**	[0.938, 0.987]	**0.002**
	Random Forest (vocab-matched)	0.842	[0.781, 0.892]	0.052
	XGBoost (vocab-matched)	0.963	[0.936, 0.984]	0.003
Pig Pancreas	scGPT emb. + LogReg	0.270	[0.258, 0.284]	0.074
	scGPT emb. + XGBoost	0.622	[0.570, 0.678]	0.034
	scGPT emb. + Random Forest	0.471	[0.430, 0.520]	0.029
	LogReg (vocab-matched)	**0.989**	[0.971, 0.997]	**0.001**
	Random Forest (vocab-matched)	0.938	[0.891, 0.969]	0.049
	XGBoost (vocab-matched)	0.972	[0.942, 0.992]	0.002

3.6 End-to-End Fine-Tuning of scGPT on TNBC

The end-to-end fine-tuning protocol is described in Section 2.8. Fine-tuned scGPT achieved Macro F1 = **0.975** (95% CI [0.972, 0.978]) and ECE = **0.003** on TNBC (Table 4). Two observations are noteworthy.

First, fine-tuning closes *most* of the gap: end-to-end fine-tuning lifts scGPT from 0.715 (frozen weights + LogReg) to 0.975 Macro F1, a +0.26 improvement. Second, fine-tuning does *not* surpass the simpler frozen-embeddings + XGBoost pipeline on the same dataset (0.972, 95% CI [0.968, 0.975]); the two confidence intervals overlap substantially. Both, however, remain below the vocabulary-matched LogReg baseline (0.993, 95% CI [0.992, 0.995]), whose CI does

not overlap with either scGPT pipeline. The qualitative finding therefore survives even the most expensive scGPT configuration we evaluated: classical HVG-based learners outperform fine-tuned scGPT on TNBC.

Table 4: Fine-tuned scGPT on TNBC Breast Cancer compared against the strongest scGPT-embedding pipeline and the strongest classical baseline on the same dataset. Bootstrap 95% CIs on Macro F1, 1,000 resamples.

Model	**Macro F1**	**95% CI**	**ECE**
scGPT (fine-tuned end-to-end, 8 epochs)	0.975	[0.972, 0.978]	0.003
scGPT embeddings + XGBoost (frozen)	0.972	[0.968, 0.975]	0.002
LogReg on vocab-matched HVGs (classical)	**0.993**	[0.992, 0.995]	**0.001**

3.7 Reliability Diagrams and Per-Class Confusion Matrices

Reliability diagrams. To make the calibration story visual, Figure 2 shows reliability diagrams for all seven models evaluated on TNBC Breast Cancer (frozen scGPT embeddings paired with each of LogReg, XGBoost, Random Forest; classical LogReg / Random Forest / XGBoost on the vocabulary-matched HVG subset; and end-to-end fine-tuned scGPT). For each bin of predicted confidence the observed test-set accuracy is plotted; perfectly calibrated curves track the diagonal, and the per-bin gap between confidence and accuracy (Section 2.6.2) makes any miscalibration directly visible. Classical XGBoost and LogReg lie almost exactly on the diagonal; scGPT + LogReg deviates substantially from the diagonal in the high-confidence bins, indicating unreliable confidence estimates in exactly the regime where downstream pipelines trust the predictions most. XGBoost and the fine-tuned scGPT both restore near-diagonal calibration. Reliability diagrams for the remaining five datasets are provided in Supplementary Figures S1–S5 and follow the same qualitative pattern.

Per-class confusion matrices. Figure 3 shows row-normalised per-class confusion matrices on TNBC Breast Cancer for the original scGPT + LogReg pipeline alongside the vocabulary-matched LogReg baseline. The scGPT + LogReg matrix shows systematic confusion between immune subpopulations (T cells vs. myeloid leukocytes, B-lineage lymphocytes vs. T cells), the precise category of error most consequential in tumour-immune analyses. The classical baseline produces a near-diagonal confusion matrix on the same cells. The corresponding Pig Pancreas matrices (Supplementary Figure S6) tell the same story for endocrine cell types (type B vs. pancreatic A cells).

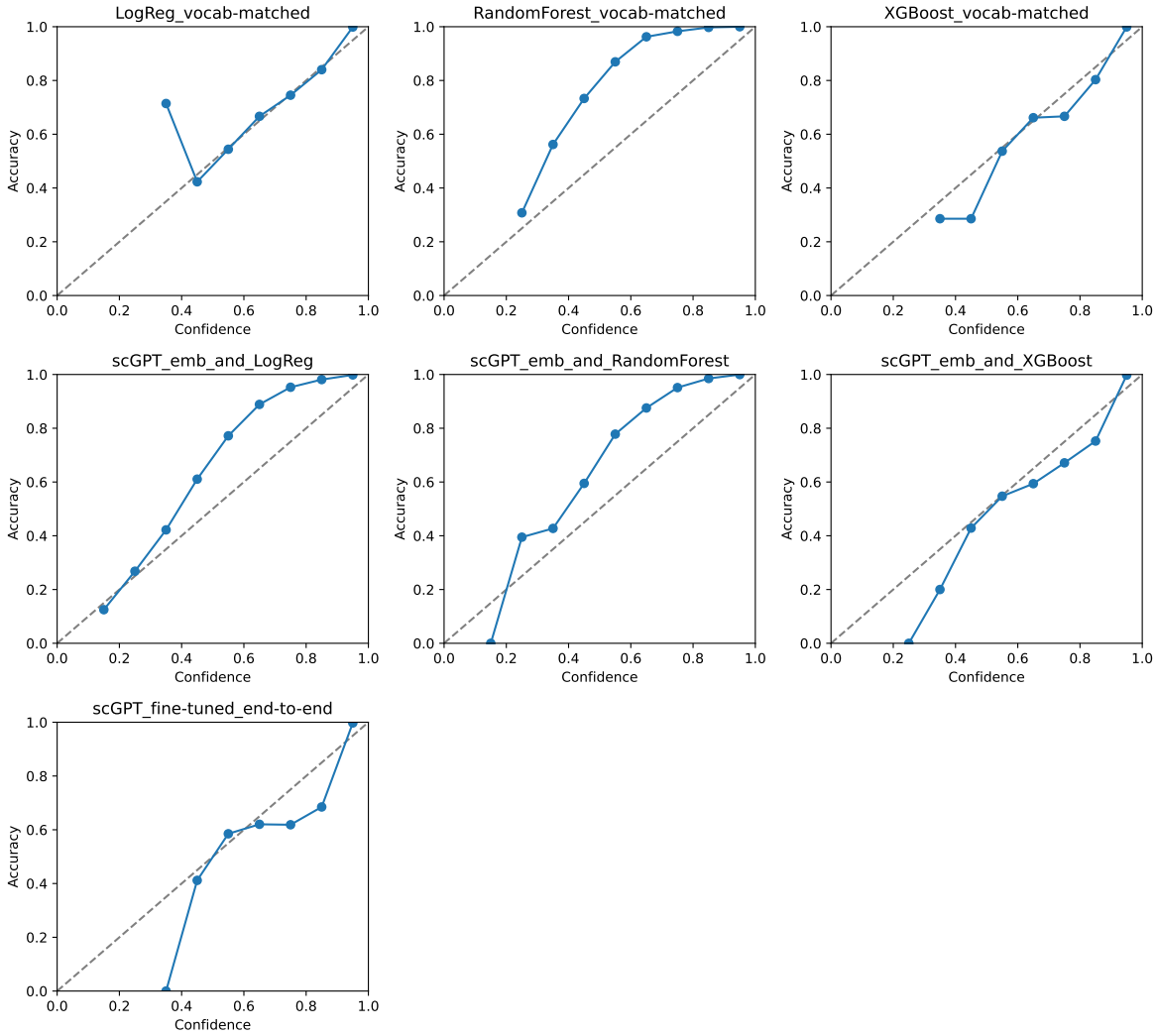


Figure 2: **Reliability diagrams on TNBC Breast Cancer.** For each model, the test set was binned by predicted confidence (10 equal-width bins on $[0, 1]$) and the mean observed accuracy in each bin was plotted against the mean predicted confidence. Points on the diagonal indicate perfect calibration. Classical baselines (top row) lie on the diagonal; scGPT + LogReg (middle row, left) departs substantially from the diagonal across the mid-to-high confidence range, indicating unreliable confidence estimates. End-to-end fine-tuning (bottom row) recovers near-diagonal calibration.

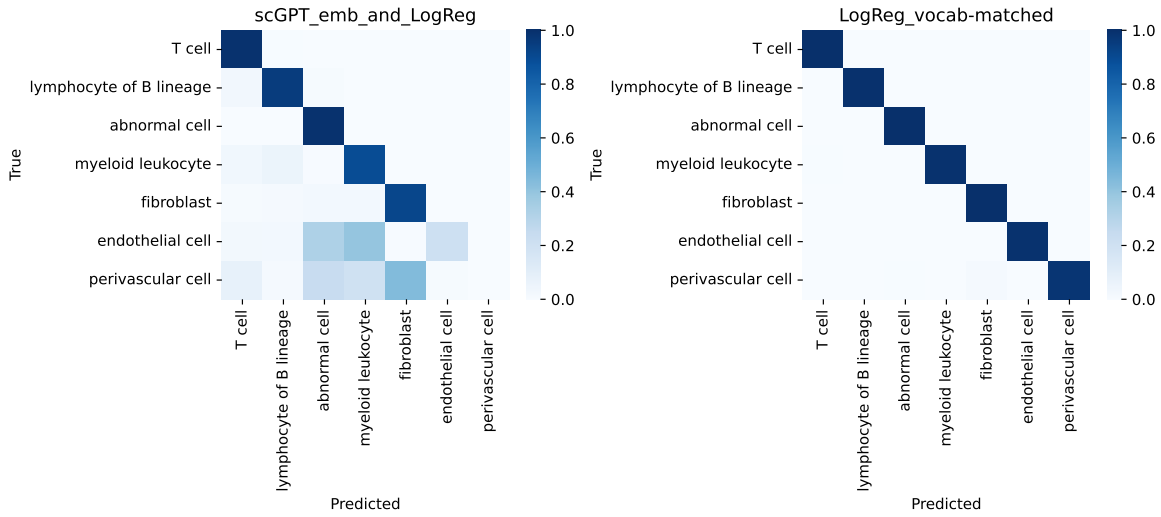


Figure 3: **Per-class confusion matrices on TNBC Breast Cancer (row-normalised; the seven cell types retained after the per-class minimum-frequency filter described in Table 1).** Left: scGPT embeddings + LogReg. Right: LogReg on the vocabulary-matched HVG subset. The scGPT pipeline systematically confuses immune subpopulations (T cells vs. myeloid leukocytes; B-lineage lymphocytes vs. T cells), the categories whose accurate separation is most consequential for downstream tumour-immune analyses.

4 Discussion

4.1 The Classical ML Baseline Is Harder to Beat Than Assumed

The central finding of our study is unambiguous: when evaluated under realistic, out-of-the-box conditions on large-scale, diverse single-cell atlases, classical machine learning methods—particularly Logistic Regression and XGBoost on HVGs—substantially and consistently outperform the scGPT-embedding+LogReg frozen-representation linear-probing pipeline. This converges with a growing body of literature questioning the practical superiority of scLLMs over well-tuned classical pipelines. Kedzierska et al. [22] reported that current scLLMs (scGPT, Geneformer, scFoundation, UCE) fail to consistently outperform simpler baselines in zero-shot transfer; Boiarsky et al. [32] reached an analogous conclusion across multiple foundation models on cell-type classification, finding that classical learners often match or exceed scLLM embeddings; and earlier benchmarking of cell-type classifiers by Abdelaal et al. [10] likewise showed that simple linear methods are remarkably hard to beat on scRNA-seq label transfer. The strong performance of HVG-based linear classifiers is also consistent with the long-standing observation that scRNA-seq variation lies in a low intrinsic dimensionality [33], which is precisely the regime in which Logistic Regression on HVGs is near-optimal. Our work extends these prior reports in two important directions: (1) by explicitly benchmarking *statistical calibration* rather than accuracy alone—which, to our knowledge, has not been quantified in any of the above studies—and (2) by demonstrating that the performance gap holds across six biologically distinct atlases spanning multiple tissues, species, and pathological contexts.

The strong performance of classical methods is perhaps less surprising on reflection. Logistic Regression and XGBoost are trained directly on the task-specific label distribution, with access

to all the HVGs that best discriminate the cell types in the target dataset. By contrast, the scGPT linear probe must rely entirely on representations frozen from pre-training, and the vocabulary matching rate (62–95% of HVGs per dataset) introduces a systematic feature gap. To test whether the gap was driven by either of these design choices, we ran two explicit fairness controls across all six atlases (Section 3.5, Table 3). Pairing the frozen scGPT embeddings with stronger downstream learners (XGBoost, Random Forest) substantially closed the original LogReg-induced gap on every dataset, with the largest improvement on TNBC (Macro F1 0.715 $\rightarrow$ 0.972 with XGBoost), confirming that the LogReg-only protocol did materially understate the discriminability of scGPT representations. Restricting the classical baselines to scGPT's own vocabulary-matched gene subset, however, did not narrow the gap on any dataset: the strongest classical learner on the matched feature space outperformed the strongest scGPT-embedding pipeline on five of six atlases (with Multi-Tissue TME a statistical tie at Macro F1 $\approx$ 0.90). An additional end-to-end fine-tuning experiment on TNBC (Section 3.6) closed most of the residual gap (Macro F1 0.715 $\rightarrow$ 0.975) but was similar to the simpler frozen-embeddings + XGBoost pipeline (0.975 vs. 0.972), with overlapping bootstrap confidence intervals, and remained below the vocabulary-matched LogReg baseline (0.993, non-overlapping CI). The defensible conclusion is therefore strengthened rather than weakened by these controls: even after equalising the downstream-learner family, restricting all models to scGPT's own input vocabulary, and fine-tuning end-to-end on the target dataset, classical HVG-based learners match or outperform every scGPT pipeline we evaluated.

4.2 Calibration as a Critical Metric

Perhaps the most operationally consequential finding is the systematic miscalibration of the scGPT linear-probe pipeline. Calibrated confidence scores are precisely what allow a practitioner to triage predictions—trusting high-confidence calls and routing low-confidence ones to expert review. When a pipeline's confidence scores do not track its empirical accuracy, as the scGPT linear-probe pipeline's do not (ECE up to 0.177), this triage breaks down: the confidence values can no longer be used as reliability estimates, regardless of the direction of the deviation. The phenomenon is well-characterised in deep learning more broadly: Guo et al. [20] showed that modern overparameterised neural networks are frequently miscalibrated and proposed Expected Calibration Error and reliability diagrams as the standard diagnostic. Our results indicate that pipelines built on single-cell foundation-model embeddings can inherit this miscalibration, and we argue that calibration should be measured and reported as routinely as accuracy [21].

The implications are practical. Consider a scenario in which scGPT is used to annotate tumour-infiltrating immune cells in a breast cancer atlas. If the scGPT pipeline misclassifies a regulatory T cell as a cytotoxic T cell, or vice versa, while attaching a confidence score that cannot be trusted as a reliability estimate, this would propagate into downstream clonotype analyses, T-cell receptor (TCR) repertoire studies—which characterise the diversity of antigen-recognition receptors expressed by an individual's T cells—or immunotherapy response predictions. The high ECE values we observed (up to 0.177 in the Multi-Tissue TME dataset) demonstrate that this miscalibration is not a rare edge case in our benchmarks but recurs across all six atlases for the scGPT-embedding+LogReg pipeline. Whether other frozen-representation linear-probing

pipelines (e.g. those built on Geneformer, scBERT, scFoundation, or UCE) inherit the same calibration failure mode is a hypothesis that is plausible in light of these results but has not yet been tested, and merits dedicated study.

We therefore advocate strongly for ECE—or alternative calibration metrics such as reliability diagrams or Brier scores—to be included as a mandatory reporting standard in all scLLM benchmarking studies.

4.3 When Are scLLMs Useful?

Our findings should not be interpreted as a categorical dismissal of scLLMs. We identify three regimes in which scLLMs, including scGPT, may provide value:

1. **Pre-training as initialisation for fine-tuning.** Our end-to-end fine-tuning experiment on TNBC (Section 3.6, Table 4) confirms that fine-tuning recovers most of the gap relative to the frozen-weights+LogReg setting (Macro F1 0.715 $\rightarrow$ 0.975) and largely restores calibration (ECE 0.066 $\rightarrow$ 0.003), as anticipated by Cui et al. [14]. Notably, however, the bootstrap confidence interval for fine-tuned scGPT on TNBC (0.975) overlapped that of the simpler frozen-embeddings + XGBoost pipeline (0.972), so we do not claim a difference between them; both remained below the vocabulary-matched classical LogReg baseline (0.993), whose interval did not overlap either. We caution that overlapping bootstrap intervals are a conservative visual heuristic and not a formal significance test—a paired bootstrap of per-cell prediction differences would be required to test the fine-tuned-vs-XGBoost contrast rigorously, which we were unable to compute because per-cell predictions were not retained. For TNBC at least, the marginal benefit of full fine-tuning over a stronger downstream probe on frozen embeddings was therefore not detectable, even though the fine-tuning required ˜30 minutes of additional GPU time.

2. **Representation learning for tasks other than annotation.** Our GRN analysis suggests that scGPT's pre-trained gene representations encode biologically interpretable gene–gene similarity structure on the dataset we examined. We did not benchmark scGPT against task-specific baselines on cell trajectory inference, drug-response prediction, or cross-modality integration; whether scGPT representations are competitive with task-specific methods on those tasks remains an open question, although prior work [14, 15, 17] has reported promising performance.

3. **Data-scarce or novel biology settings.** In scenarios where labelled training data is genuinely unavailable and classical supervised methods are therefore not applicable, frozen scLLM representations—when paired with appropriate uncertainty quantification—may provide a useful starting point, as suggested by prior reports that pre-trained scLLM embeddings transfer to small labelled cohorts and rare cell-type discovery [15–17], and as more broadly demonstrated by reference-mapping frameworks such as scArches [34], which leverage pre-trained representations to integrate small query datasets into existing reference atlases.

4. **Cell-type composition heterogeneity within an atlas.** On Multi-Tissue TME

(Table 3), scGPT + XGBoost on frozen embeddings reached essentially the same Macro F1 as the vocabulary-matched LogReg baseline (0.896 vs. 0.898). This was the only dataset of six on which a scGPT-embedding pipeline matched the strongest classical linear baseline, and it is also the dataset that spans the broadest tissue compartment range in our panel. This isolated equivalence suggests that scGPT representations may carry complementary information when the target task requires generalisation across heterogeneous source tissues—a hypothesis worth pursuing in dedicated benchmarks.

4.4 The Cost–Benefit Calculus of scLLMs

From a resource perspective, the cost differential between scLLMs and classical methods is striking. Our scGPT embedding extraction required approximately four hours of compute time on a high-end H100 GPU at a cost of approximately USD 15. In contrast, XGBoost and Logistic Regression, which substantially outperformed scGPT, can be run on standard laptop hardware in minutes. For the majority of laboratories and clinical institutions without access to high-performance GPU clusters, the practical barriers to scLLM deployment are substantial.

This compute cost would be justified if scLLMs provided superior accuracy. Our results demonstrate that they do not, at least in the frozen-representation linear-probing paradigm with a logistic-regression head that we evaluated here. Until scLLMs demonstrate consistent, calibrated superiority over classical methods in fair benchmark conditions, their broad adoption for routine cell-type annotation is difficult to justify on either performance or resource grounds.

4.5 Limitations

Several limitations of this study should be acknowledged.

First, we evaluated only one scLLM (scGPT). The frozen-representation pipeline was tested with three downstream learners (Section 3.5), and an additional end-to-end fine-tuning experiment was performed on TNBC (Section 3.6); however, alternative scLLMs (Geneformer, scFoundation, UCE) may show different accuracy and calibration profiles, and the end-to-end fine-tuning experiment was performed on a single dataset only. We do not generalise our results beyond the scGPT family of models.

Second, the calibration analysis pertains to the end-to-end classification pipeline—scGPT embeddings followed by a downstream classifier—rather than to scGPT in isolation. The predicted probabilities used to compute the ECE originate from the downstream classifier, not from scGPT itself. Our defensible claim is therefore that the scGPT-embedding+LogReg pipeline is markedly less well calibrated than classical HVG-based pipelines on these benchmarks, not that scGPT is intrinsically uncalibrated as a representation model.

Third, our results are based on a single stratified 80/20 train/test split per dataset, with HVG selection performed on the full dataset prior to the split. The split is a *random* stratified split that preserves class proportions but does *not* group by donor or batch; cells from the same donor may therefore appear in both the training and test folds, so the reported accuracies reflect within-cohort generalisation and may be optimistic relative to a donor- or batch-held-out protocol. Because all model families are evaluated on the identical split, this affects the absolute numbers rather than the cross-model ranking, but a donor-grouped split is an important refinement for

future work. Test-set uncertainty is quantified by 1,000 bootstrap resamples (percentile-method 95% CIs on Macro F1, reported throughout Tables 3 and 4), which captures sampling variability for a fixed training set but not variability across alternative training partitions. Critically, because the same HVG set is provided to all four model families on each dataset, none of the cross-model comparisons in this paper are biased by HVG leakage—only the absolute performance numbers may be modestly optimistic. The same dataset-wide HVG-selection convention has been adopted by prior benchmarks against which we compare [10, 22, 32], so our numbers are directly comparable to that literature. We did not perform k-fold cross-validation: each of the six atlases requires a multi-hour scGPT embedding pass on a high-end GPU, and repeating the entire pipeline (including HVG selection nested within each fold) for all six datasets and all six fairness-control models, plus one end-to-end fine-tuning run, was outside our compute budget. A stricter nested k-fold CV protocol reporting mean $\pm$ std across folds is a worthwhile refinement for future work.

Fourth, our classical ML methods used default hyperparameters without optimisation, which may slightly underestimate their potential performance but reflects realistic out-of-the-box usage.

Fifth, the GRN analysis is an embedding-derived gene co-expression similarity analysis rather than an attention-based or causal regulatory inference, and it is qualitative: the recovered modules are validated against the published immunology literature for each module (Section 3.4; 29–31), but we did not quantify enrichment against a null or baseline graph (random gene sets, a PCA-derived similarity graph, or a raw-expression correlation graph), and experimental in-vitro validation of any putatively novel predictions remains beyond the scope of this benchmarking study. A quantitative, null-anchored network benchmark is a natural extension.

Sixth, our benchmark compares scGPT against generic machine-learning baselines (Logistic Regression, Random Forest, XGBoost) but does not include established reference-based or marker-based single-cell annotation tools such as SingleR [6], Azimuth [9], scType [35], scPred [8], CellTypist [12], or scmap [7]. Comprehensive head-to-head comparisons among such tools have been reported elsewhere [10, 11]; our contribution is specifically the head-to-head comparison of an scLLM against the supervised-learning baseline that scLLMs are most often *implicitly* positioned to replace. A direct benchmark of scGPT against the reference-based annotators above (using the same six atlases, the same evaluation protocol, and the same calibration metrics) is a natural and important follow-up, and we expect it to further sharpen the cost–benefit picture for clinical deployment.

Seventh, our calibration audit reports ECE and the bin-level reliability diagrams but not the scalar Brier score, negative log-likelihood, or signed and class-wise ECE decompositions, because per-cell predicted-probability arrays were not retained from the original runs. Since ECE is an unsigned summary and those arrays would be required to compute signed or strictly proper scoring rules, we restrict our claims to the *magnitude* of miscalibration captured by ECE and the reliability diagrams, and do not interpret its direction.

5 Conclusions

We present, to our knowledge, a focused benchmarking of statistical calibration in pipelines built on scGPT at scale. Across six large, diverse single-cell atlases totalling 1.3 million cells,

classical machine learning methods—XGBoost, Logistic Regression, and Random Forest trained on highly variable genes—substantially and consistently outperform a frozen-representation linear-probing pipeline built on scGPT in both predictive accuracy (Macro F1) and calibration (ECE). Critically, the scGPT-embedding+LogReg pipeline is systematically miscalibrated across all evaluated datasets, with ECE far exceeding that of every classical baseline, so that its confidence scores cannot be used as reliability estimates. Despite this, scGPT embeddings retain biologically meaningful structure, as evidenced by the recapitulation of known co-expression relationships in zero-shot, embedding-derived gene-network inference.

Our findings argue for three immediate changes in how the field evaluates and deploys scLLMs: (1) calibration metrics should be mandatory in all scLLM benchmarking publications, with a clear distinction between calibration of the foundation model and calibration of the downstream pipeline; (2) classical ML baselines should be rigorously benchmarked against scLLM pipelines—using matched downstream learners and matched input gene sets—before claiming scLLM superiority; and (3) the practical and financial costs of scLLM deployment must be weighed against demonstrated performance advantages. Until scLLMs can demonstrate calibrated, reproducible superiority over classical methods under matched and externally validated benchmark conditions, their routine adoption for cell-type annotation in large-scale atlases is premature.

List of Abbreviations

CZ CELLxGENE: Chan Zuckerberg CELL by GENE Discover portal; ECE: Expected Calibration Error; GRN: Gene Regulatory Network; HVG: Highly Variable Gene; LogReg: Logistic Regression; MHC-II: Major Histocompatibility Complex class II; PBMC: Peripheral Blood Mononuclear Cell; scLLM: Single-Cell Large Language Model; scRNA-seq: Single-Cell RNA Sequencing; TCR: T-Cell Receptor; TME: Tumour Microenvironment; TNBC: Triple-Negative Breast Cancer.

Declarations

Ethics approval and consent to participate. Not applicable. This study used only publicly available, fully de-identified scRNA-seq datasets distributed through the Chan Zuckerberg CELLxGENE Discover portal under the original data generators' ethics approvals.

Consent for publication. Not applicable.

Availability of data and materials. All six datasets used in this study are publicly available through the CZ CELLxGENE Discover portal (https://cellxgene.cziscience.com/). The dataset landing pages, together with the interactive CELLxGENE Explorer pages for the exact datasets analysed, are:

- TNBC Breast Cancer — "Single-cell atlas of untreated human triple-negative breast cancer". Landing page: https://cellxgene.cziscience.com/collections/ceef2841-5333-46ac-92ef-ccbe0
  dataset: https://cellxgene.cziscience.com/e/6f9de485-58cd-4342-bfc4-b3d3dd223aa8.cxg/.

- Indonesia PBMC — "Ancestral and environmental diversity shape the immune landscape in Indonesia". Landing page: https://cellxgene.cziscience.com/collections/d1e0e64d-6d2a-4a3e-b7f4-43ed909a9d9c; dataset: https://cellxgene.cziscience.com/e/ca7d95ac-53aa-4a94-8c76-300c608ce6a0.cxg/.
- Brain Atlas — "ALS motor cortex and spinal cord single-nucleus multiome dataset". Landing page: https://cellxgene.cziscience.com/collections/0986e4cd-7a58-405d-9b91-4b199bb412 dataset: https://cellxgene.cziscience.com/e/e2a00644-0a48-4815-ae87-d562045114f5.cxg/.
- Multi-Tissue TME — "A multi-tissue single-cell tumour microenvironment atlas". Landing page: https://cellxgene.cziscience.com/collections/3f7c572c-cd73-4b51-a313-207c7f20f188; dataset: https://cellxgene.cziscience.com/e/b617ee1b-f8c8-4de9-b82b-e803ab93550d.cxg/.
- Human Pancreas — "A transcriptional cross-species map of pancreatic islet cells". Landing page: https://cellxgene.cziscience.com/collections/0a77d4c0-d5d0-40f0-aa1a-5e1429bcbd7e; dataset: https://cellxgene.cziscience.com/e/3294d050-6eeb-4a00-b24c-71aacc9b777f.cxg/.
- Pig Pancreas — "A transcriptional cross-species map of pancreatic islet cells". Landing page: https://cellxgene.cziscience.com/collections/0a77d4c0-d5d0-40f0-aa1a-5e1429bcbd7e; dataset: https://cellxgene.cziscience.com/e/db4a9ed2-e994-40c1-b7ec-4091fdf7b6c1.cxg/.

The scGPT pre-trained whole-human model weights are distributed by the original authors at https://github.com/bowang-lab/scGPT. All analysis code, the comprehensive benchmarking pipeline (including end-to-end fine-tuning, bootstrap 95% confidence intervals, reliability diagrams, and per-class confusion matrices), and the bootstrap CSV outputs are openly available at https://github.com/alirezakhosravii/scGPTvsClassical under the MIT licence.

Competing interests. The authors declare that they have no competing interests.

Funding. This research was supported by the OPUS grant of the National Science Centre, Poland (grant number 2020/37/B/NZ7/02069) and by the National Recovery Plan grant of the Medical Research Agency, Poland (grant number KPOD.07.07-IW.07-0007/24). The cloud GPU computational costs were personally funded by A.K. The funding bodies played no role in the design of the study; in the collection, analysis, or interpretation of data; or in writing the manuscript.

Authors' contributions. **A.K.** conceptualised the study, designed the methodology, implemented the software, performed the formal analysis and investigation, curated the data, produced the visualisations, drafted the original manuscript, and contributed to manuscript review and editing. **Ar.K.** contributed to software implementation, formal analysis, and manuscript review and editing. **K.Ł.** and **M.S.** contributed to manuscript review and editing. **K.P.** contributed

to manuscript review and editing and to supervision. **A.S.** contributed to conceptualisation, manuscript review and editing, supervision, and project administration. **A.J.Ż.** contributed to manuscript review and editing, supervision, and funding acquisition. All authors read and approved the final manuscript.

Acknowledgements. The authors thank the staff at the Centre of Biostatistics and Bioinformatics, Medical University of Gdańsk, and in particular Małgorzata Madej for her support. The authors also thank the Chan Zuckerberg Initiative for providing open access to the CELLxGENE Discover portal and the curated single-cell atlas datasets used in this study, and acknowledge the scGPT development team at the Wang Lab (University of Toronto) for open-sourcing the pre-trained model weights and the scGPT Python package. Computational resources were provided by RunPod cloud GPU services.

References

[1] Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, et al. The Human Cell Atlas. *eLife*, 6:e27041, 2017. doi: 10.7554/eLife.27041.

[2] The Tabula Sapiens Consortium. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022. doi: 10.1126/science.abl4896.

[3] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015. doi: 10.1016/j.cell.2015.05.002. URL https://www.cell.com/fulltext/S0092-8674(15)00549-8.

[4] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017. doi: 10.1038/ncomms14049. URL https://www.nature.com/articles/ncomms14049.

[5] Amos Tanay and Aviv Regev. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541:331–338, 2017. doi: 10.1038/nature21350. URL https://www.nature.com/articles/nature21350.

[6] Dvir Aran, Agnieszka P. Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P. Naikawadi, Paul J. Wolters, Adam R. Abate, Atul J. Butte, and Mallar Bhattacharya. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20:163–172, 2019. doi: 10.1038/s41590-018-0276-y. URL https://www.nature.com/articles/s41590-018-0276-y.

[7] Vladimir Yu. Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods*, 15:359–362, 2018. doi: 10.1038/nmeth.4644. URL https://www.nature.com/articles/nmeth.4644.

[8] Jose Alquicira-Hernandez, Anuja Sathe, Hanlee P. Ji, Quan Nguyen, and Joseph E. Powell. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology*, 20:264, 2019. doi: 10.1186/s13059-019-1862-5. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1862-5.

[9] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, et al.

Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021. doi: 10.1016/j.cell.2021.04.048. URL https://www.cell.com/cell/fulltext/S0092-8674(21)00583-3.

[10] Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J. T. Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, 20:194, 2019. doi: 10.1186/s13059-019-1795-z. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1795-z.

[11] Giovanni Pasquini, Jesus Eduardo Rojo Arias, Patrick Schäfer, and Volker Busskamp. Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal*, 19:961–969, 2021. doi: 10.1016/j.csbj.2021.01.015. URL https://www.csbj.org/article/S2001-0370(21)00020-X/fulltext.

[12] C. Domínguez Conde, C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, S. K. Howlett, O. Suchanek, K. Polanski, H. W. King, L. Mamanova, N. Huang, P. A. Szabo, L. Richardson, L. Bolt, E. S. Fasouli, K. T. Mahbubani, M. Prete, L. Tuck, N. Richoz, Z. K. Tuong, L. Campos, H. S. Mousa, E. J. Needham, S. Pritchard, T. Li, R. Elmentaite, J. Park, E. Rahmani, D. Chen, D. K. Menon, O. A. Bayraktar, L. K. James, K. B. Meyer, N. Yosef, M. R. Clatworthy, P. A. Sims, D. L. Farber, K. Saeb-Parsy, M. Zambon, M. Noseda, M. Haniffa, W. Sungnak, D. R. Gaya, M. Scoôse, S. Nejentsev, A. L. Haber, A.-C. Villani, A. Lun, A. Horowitz, M. Elowitz, and S. A. Teichmann. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022. doi: 10.1126/science.abl5197. URL https://www.science.org/doi/10.1126/science.abl5197.

[13] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018. doi: 10.1038/s41592-018-0229-2. URL https://www.nature.com/articles/s41592-018-0229-2.

[14] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://www.nature.com/articles/s41592-024-02201-0.

[15] Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Ziad R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023. doi: 10.1038/s41586-023-06139-9. URL https://www.nature.com/articles/s41586-023-06139-9.

[16] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4:852–866, 2022. doi: 10.1038/s42256-022-00534-z. URL https://www.nature.com/articles/s42256-022-00534-z.

[17] Minsheng Hao, Jing Gong, Xin Zeng, Chengyu Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7. URL https://www.nature.com/articles/s41592-024-02305-7.

[18] Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2023. doi: 10.1101/2023.11.28.568918. URL https://www.biorxiv.org/content/10.1101/2023.11.28.568918v2.

[19] OpenAI. GPT-4 technical report. arXiv preprint arXiv:2303.08774, OpenAI, 2023.

[20] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a/guo17a.pdf.

[21] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.

[22] Kayla Z. Kedzierska, Lauren Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26:103, 2025. doi: 10.1186/s13059-025-03574-x. URL https://link.springer.com/article/10.1186/s13059-025-03574-x.

[23] CZI Single-Cell Biology Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, et al. CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv*, 2023. doi: 10.1101/2023.10.30.563174. URL https://www.biorxiv.org/content/10.1101/2023.10.30.563174v1.

[24] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018. doi: 10.1186/s13059-017-1382-0. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1382-0.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

[26] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785. URL https://arxiv.org/abs/1603.02754.

[27] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324. URL https://link.springer.com/article/10.1023/A:1010933404324.

[28] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE*, 5(9): e12776, 2010. doi: 10.1371/journal.pone.0012776. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0012776.

[29] Walter Reith and Bernard Mach. The bare lymphocyte syndrome and the regulation of MHC expression. *Annual Review of Immunology*, 19:331–373, 2001. doi: 10.1146/annurev.immunol.19.1.331.

[30] Ilia Voskoboinik, James C. Whisstock, and Joseph A. Trapani. Perforin and granzymes: function, dysfunction and human pathology. *Nature Reviews Immunology*, 15(6):388–400, 2015. doi: 10.1038/nri3839.

[31] Jennifer E. Smith-Garvin, Gary A. Koretzky, and Martha S. Jordan. T-cell activation. *Annual Review of Immunology*, 27:591–619, 2009. doi: 10.1146/annurev.immunol.021908.132706.

[32] Rebecca Boiarsky, Nalini Singh, Alejandro Buendia, Gad Getz, and David Sontag. A deep dive into single-cell RNA sequencing foundation models. *bioRxiv*, page 2023.10.19.563100, 2023. doi: 10.1101/2023.10.19.563100.

[33] Graham Heimberg, Rajat Bhatnagar, Hana El-Samad, and Matt Thomson. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Systems*, 2(4):239–250, 2016. doi: 10.1016/j.cels.2016.04.001.

[34] Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40 (1):121–130, 2022. doi: 10.1038/s41587-021-01001-7.

[35] Aleksandr Ianevski, Anil K. Giri, and Tero Aittokallio. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nature Communications*, 13:1246, 2022. doi: 10.1038/s41467-022-28803-w.

Additional File 1 — Supplementary Figures

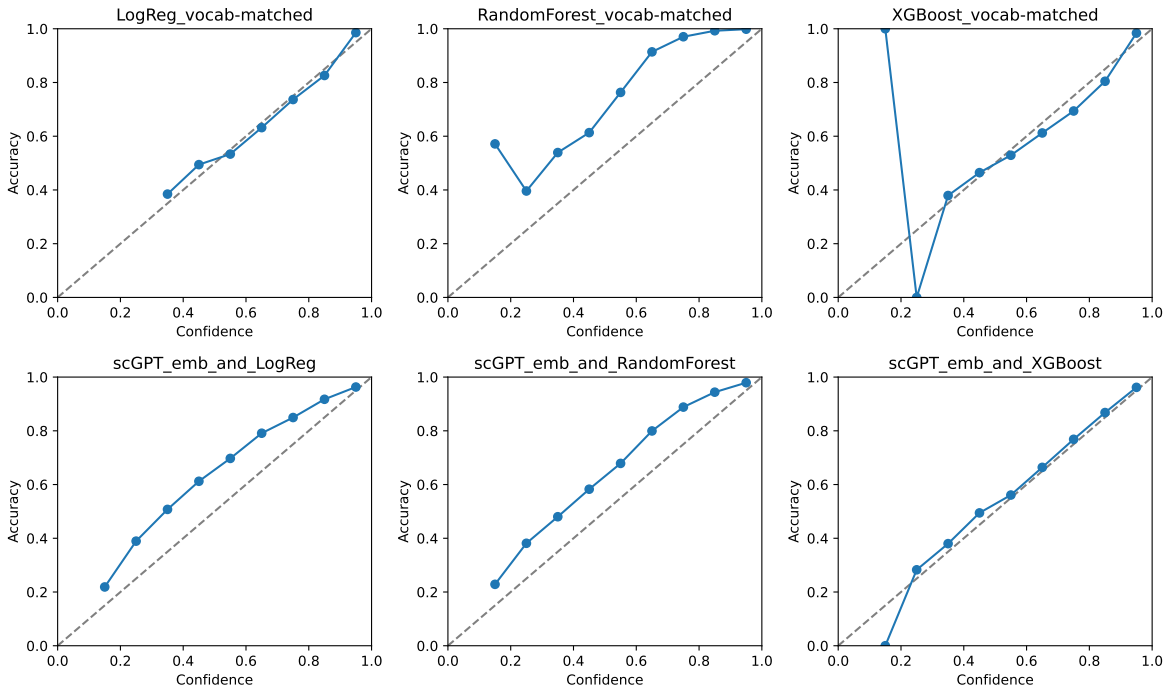


Figure 4: **Supplementary Figure S1.** Reliability diagrams for the six fairness-control models on the Indonesia PBMC dataset. Layout matches Figure 2.

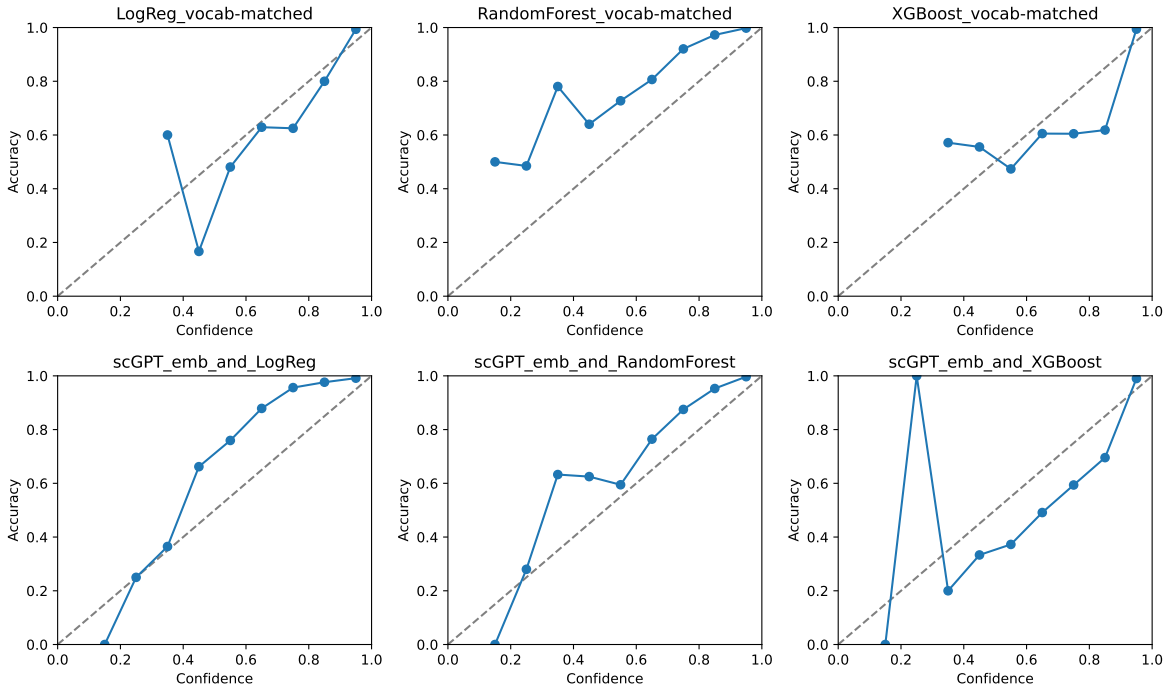


Figure 5: **Supplementary Figure S2.** Reliability diagrams for the six fairness-control models on the Brain Atlas dataset.

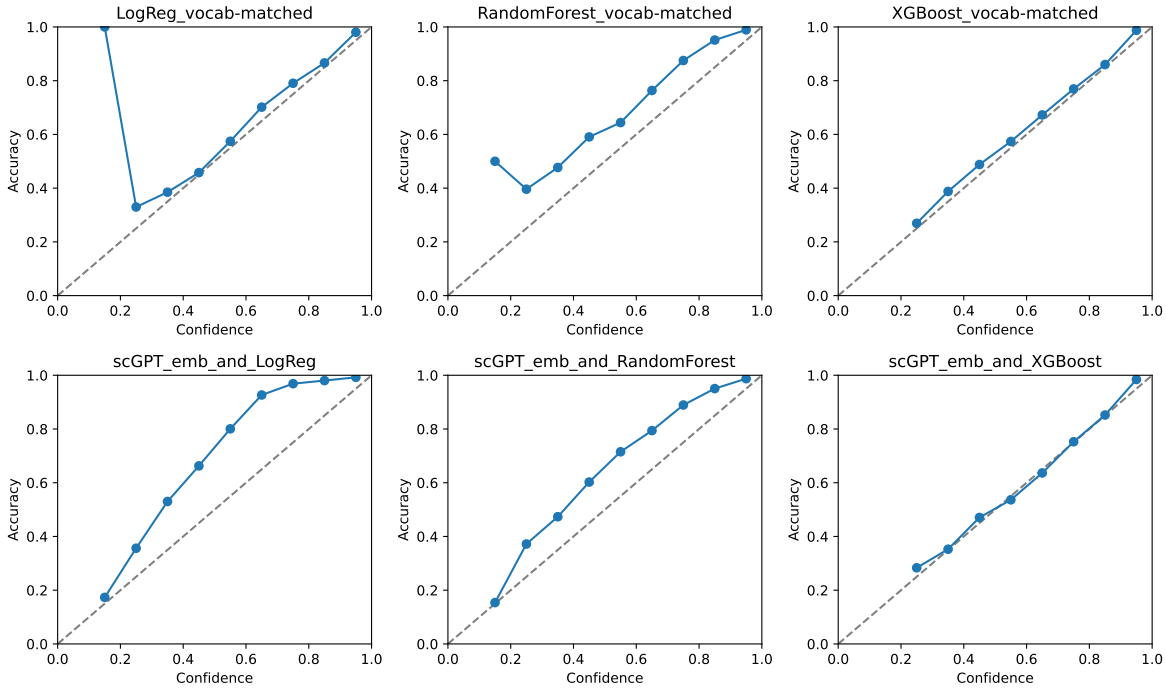


Figure 6: **Supplementary Figure S3.** Reliability diagrams for the six fairness-control models on the Multi-Tissue TME dataset. Note that on this dataset the scGPT + XGBoost panel and the LogReg (vocab-matched) panel are visually almost identical, in line with the F1 tie reported in Table 3.

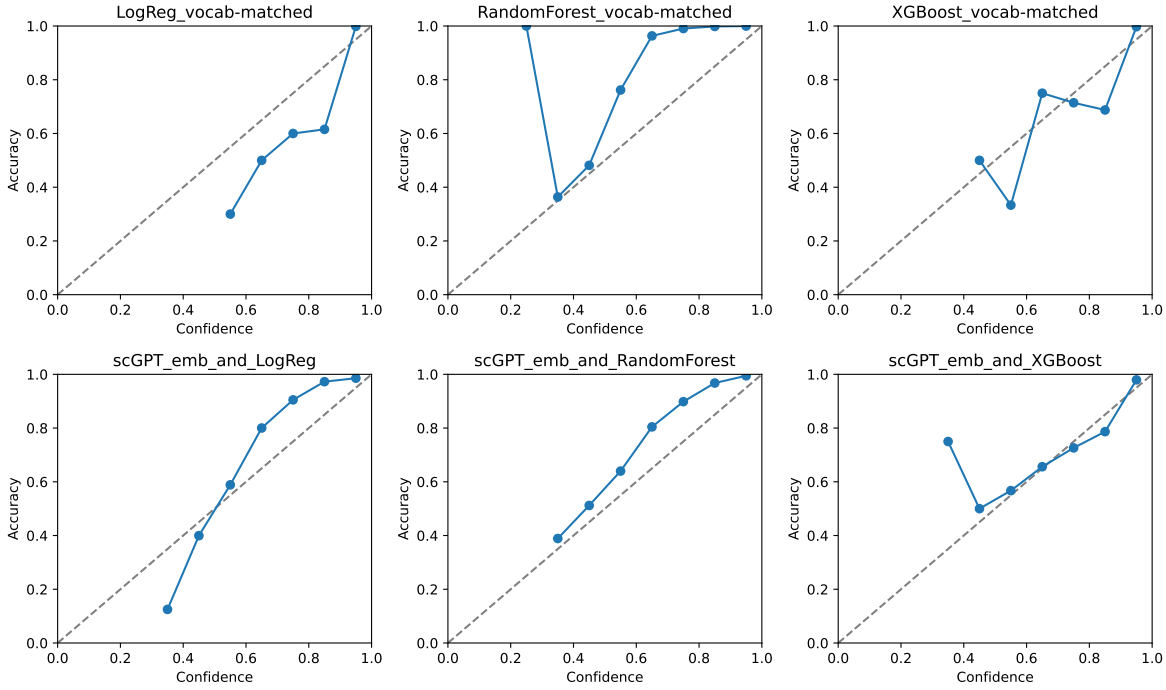


Figure 7: **Supplementary Figure S4.** Reliability diagrams for the six fairness-control models on the Human Pancreas dataset.

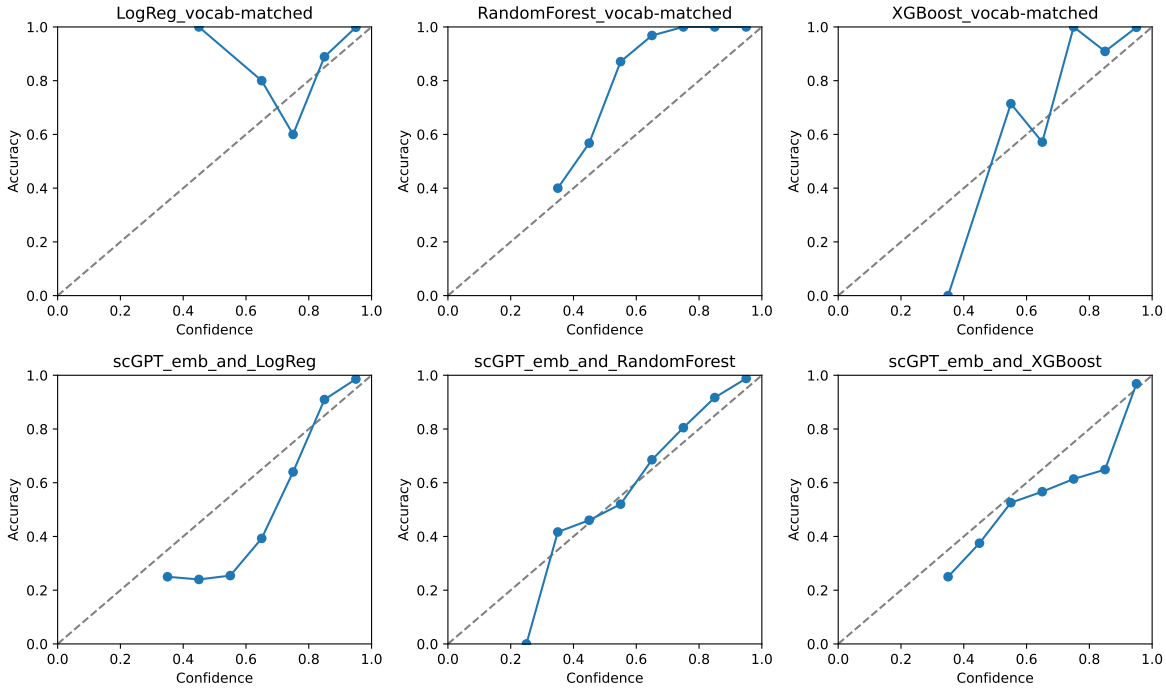


Figure 8: **Supplementary Figure S5.** Reliability diagrams for the six fairness-control models on the Pig Pancreas dataset.

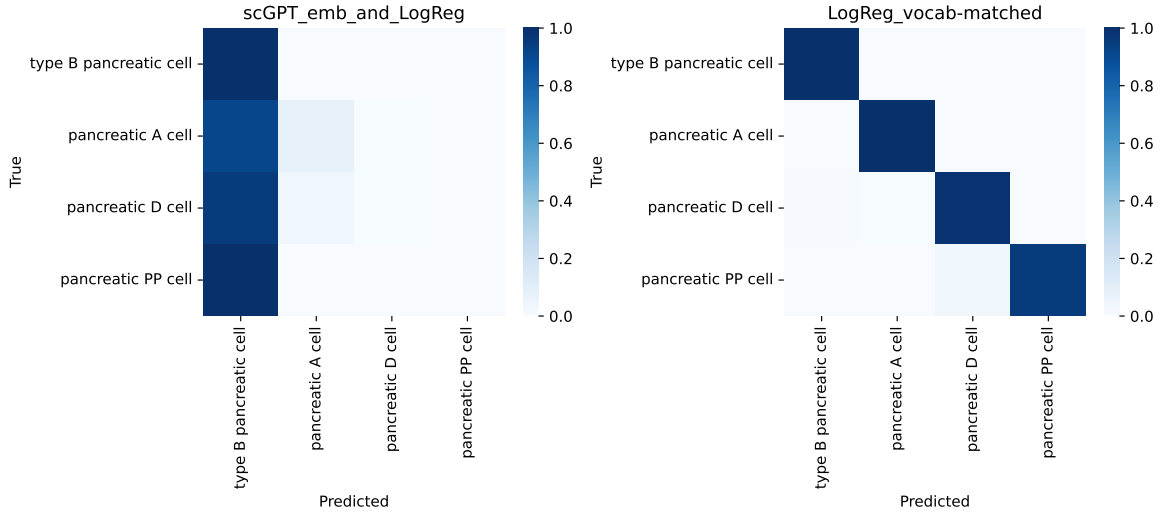


Figure 9: **Supplementary Figure S6.** Per-class row-normalised confusion matrices on the Pig Pancreas dataset. Left: scGPT embeddings + LogReg. Right: LogReg on the vocabulary-matched HVG subset. The scGPT pipeline systematically confuses the four endocrine cell types (type B, A, D, PP); the classical baseline produces a near-diagonal confusion matrix.