scCoLLM: An LLM-Based Model Using Context-Space Adaptation for Single-Cell Analysis under Domain Shift

Anonymous authors
Paper under double-blind review

Abstract

Deep learning has been adopted throughout single-cell analysis with remarkable success. However, once the target data to be analyzed depart from the distribution a single-cell model was trained on, current single-cell models exhibit marked performance degradation. Such *domain shift* is the rule rather than the exception in practice: newly profiled data routinely come from unfamiliar cell types, tissues, and donors. We attribute much of this degradation to a missing conditioning signal: existing methods predict each query cell from its own expression profile alone, ignoring the *target-domain-specific biological context* carried by the target dataset itself. Closing this gap by training or fine-tuning on the target domain is often infeasible, since every such remedy requires an additional training run for each new target, its supervised variants further require ground truth annotations that a newly profiled dataset does not carry, and the strongest models may not release their weights. We therefore recast target-domain adaptation from *parameter space* to *context space*. We introduce **s**ingle-**c**ell **Co**ntext **LLM** (**scCoLLM**), a training-free framework that performs **Context-Space Adaptation** on top of general-purpose LLMs such as GPT. Given a query cell, scCoLLM retrieves a query-conditioned biological context set from the *unlabeled* target data, selected to balance query-level local relevance against target-domain global diversity, yielding a set of cell–profile pairs through zero-shot reasoning. The resulting cell–profile pairs form the **Bio-Context**. Conditioned on this Bio-Context, scCoLLM analyzes the query cell while aware of the biological context of the target domain, and thereby adapts to that domain without gradient updates, ground-truth labels, or access to model parameters. Under three domain shift tasks—cross-cell type, cross-tissue, and cross-donor—scCoLLM attains state-of-the-art performance, improving on the strongest of 12 baselines by up to 9%. Further experiments confirmed the soundness of the Bio-Context design and its decisive role in improvement, as well as the robustness and transferability of scCoLLM.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has become a foundational technology in modern biology, revealing cellular heterogeneity that is obscured in bulk assays (Tang et al., 2009; Macosko et al., 2015). Across most downstream **single-cell analyses lies a common prediction problem: given the query cell, infer its identity from the measured gene expression profile.** A series of AI-based single-cell models, including dataset-specific models (Lopez et al., 2018; Stuart et al., 2019; Korsunsky et al., 2019), single-cell foundation models (Theodoris et al., 2023; Cui et al., 2024; Rosen et al., 2026; Yuan et al., 2024) and LLM-based methods (Hou & Ji, 2024; Fang et al., 2026) have been proposed, making automated single-cell analysis increasingly accurate and practical.

This progress, however, is largely measured under an assumption that real applications violate. Existing methods are predominantly evaluated on held-out cells drawn from the same distribution as the data on which they were developed (**in-domain setting**), whereas datasets generated in real workflows typically arrive under **domain shift**, comprising unfamiliar cells that differs significantly from the training corpus. Such shifts change how a cell expresses its identity even when the identity itself is unchanged. Figure 1 illustrates this for T cells from the exocrine pancreas and abdominal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

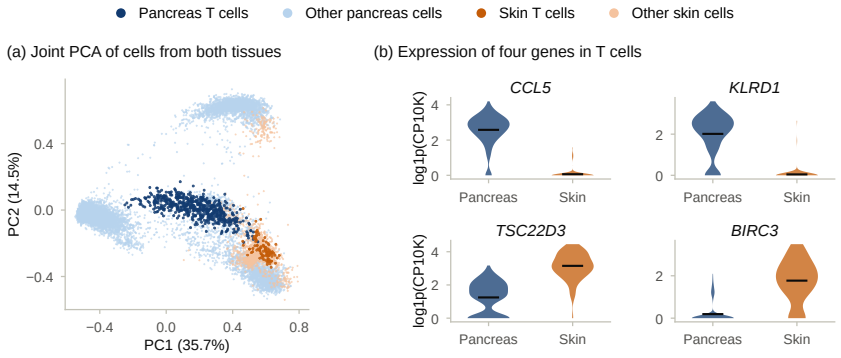


Figure 1: T cells differ markedly between tissues and donors despite sharing the same cell type. (a) Joint PCA of all cells from the two domains. (b) Expression of four genes in these T cells, measured as log1p(CP10K) of the genes retained in the model input; black bars mark the means.

skin, sampled from different tissues and donors. Although these cells carry the same cell type, T cells from the two tissues and donors occupy distinct regions of the joint PCA space (Figure 1a) and differ sharply at the level of individual genes (Figure 1b). A model that has learned the expression signature of T cells in one therefore encounters a substantially different signature for the same cell type in another, which makes cell analysis under domain shift a realistic and difficult problem.

Performance degradation under such domain shifts in single-cell analysis is a repeatedly documented finding rather than a conjecture: supervised classifiers lose accuracy in inter-dataset evaluation (Abdelaal et al., 2019); scFMs prove unreliable without finetuning and are at times outperformed by simpler baselines (Kedzierska et al., 2025; Boiarsky et al., 2024; Yuan et al., 2024). Because discovery-oriented studies are precisely those that probe unfamiliar biological territory, domain shift is the dominant operating regime rather than a corner case, and **strengthening adaptation under domain shift is therefore a problem of both methodological and biological importance.**

We hypothesize that the reason for this degradation is a missing conditioning signal. A shifted target dataset is not merely a collection of unfamiliar cells; it carries target-specific structure—the composition of cell populations present in the target, the distribution of expression states within them, and the local similarity and contrast structure surrounding a given query. Existing single-cell methods largely ignore this signal: they predict each query cell from its own profile alone. The obvious remedy—retraining or fine-tuning on the target domain—is rarely viable in practice. Every such remedy requires an additional training run for each new target dataset, which is expensive and must be repeated per target; its supervised variants additionally require ground truth annotations, which for a newly profiled dataset are typically unavailable. **This raises the question we study: How can a single-cell model adapt to a target domain without updating or accessing model parameters and without target-domain ground truth annotations? Our answer is Context-Space Adaptation**, relocating target-domain adaptation from parameter space to context space.

General-purpose LLMs exhibit strong in-context learning (ICL) ability, providing a training-free adaptation mechanism during inference (Brown et al., 2020; Wei et al., 2022). We build on this property and propose **s**ingle-**c**ell **Co**ntext **LLM** (**scCoLLM**), an LLM-based framework for single-cell analysis under domain shift that realizes *context-space adaptation*. Figure 2 illustrates the overall framework: for a given query cell, scCoLLM retrieves a query-conditioned biological context under two complementary criteria: *query-level local relevance* and *target-domain global diversity*. scCoLLM then generates cell profiles based on biological context using zero-shot reasoning; these are free-text descriptions of the cells' identity attributes. The resulting cell–profile pairs form what we call the **Bio-Context**, an explicit representation of the domain-specific biological context of the target domain. At inference, the Bio-Context is inserted into scCoLLM's context prompt, so that the model analyzes the query cell while aware of the biological context of the target domain.

We evaluate scCoLLM on a benchmark involving three types of domain shifts: cross-cell-type, cross-tissue, and cross-donor, and observe state-of-the-art performance in all three settings. Beyond accuracy, **scCoLLM offers three practical advantages**: 1) it is **training-free**, which removes the

cost of per-target adaptation; 2) it **requires no target-domain labels**, which substantially broadens its applicability; 3) and it **requires no access to model parameters**, which allows it to be carried over to future and stronger LLMs. Further analyses support the mechanism we claim: 1) ablations confirm that the two retrieval criteria are complementary, ensuring that the retrieved context possesses both relevance and diversity; 2) we verify that the constructed Bio-Context is what drives the performance gain, rather than the mere presence of additional in-context examples; 3) scCoLLM is robust both to errors in the Bio-Context and to context size; 4) consistent behavior across LLM backbones of different families and scales demonstrates the portability of scCoLLM, allowing the method to benefit from continued progress in general-purpose LLMs.

Our contributions are summarized as follows.

- **Framework.** We study single-cell analysis under domain shift and formulate *Context-Space Adaptation*, a framework that realizes target-domain adaptation without additional training, access to model parameters, or target-domain ground truth annotations.
- **Method.** We propose scCoLLM, which constructs Bio-Context by jointly accounting for query-level local relevance and target-domain global diversity, thereby enabling scCoLLM to analyze each query using explicit biological context from the target domain.
- **Empirical evidence.** scCoLLM achieves state-of-the-art performance under three types of domain shift. Further experiments confirmed the soundness of the Bio-Context design and its decisive role in improvement, as well as the robustness and transferability of scCoLLM.

2 Related Work

Artificial intelligence for single-cell analysis. Existing approaches fall into three families. *Single-cell foundation models* (scFMs) pretrain large neural networks on millions of cells and transfer to downstream tasks (Theodoris et al., 2023; Cui et al., 2024; Fischer et al., 2024; Rosen et al., 2026; Yuan et al., 2024; Dong et al., 2026). *Dataset-specific methods* derive representations from the dataset at hand without external pretraining (Stuart et al., 2019; Korsunsky et al., 2019; Lopez et al., 2018). *LLM-based methods* instead draw on the general biological knowledge encoded in large language models and reason about cells in natural language (Hou & Ji, 2024; Fang et al., 2026).

Adaptation under domain shift. However, most existing single-cell models exhibit significant performance degradation when faced with domain-shift scenarios. This is a broadly documented concern rather than an isolated observation (Abdelaal et al., 2019; Boiarsky et al., 2024; Kedzierska et al., 2025; Nadig et al., 2025). Several recent methods move beyond query-centric prediction by exploiting context, but at a considerable cost, such as large-scale pre-training (Dong et al., 2026), supervised training (Fang et al., 2026), or the requirement that all candidate labels be known (Fang et al., 2026). scCoLLM removes these requirements and realizes context-space adaptation.

3 Method

This section presents scCoLLM. Section 3.1 formalizes single-cell analysis under domain shift and the setting of context-space adaptation. Section 3.2 describes how scCoLLM constructs a query-conditioned Bio-Context from unlabeled target-domain data, and Section 3.3 describes how the Bio-Context conditions scCoLLM to produce cell features. Figure 2 provides an overview of scCoLLM.

3.1 Problem Formulation

Single-cell models. Let $\mathcal{G} = \{g_1, \ldots, g_G\}$ denote a vocabulary of G genes. A cell c profiled by scRNA-seq is described by its expression profile $\boldsymbol{x}_c \in \mathbb{R}^G_{\geq 0}$, whose j-th entry $x_{c,j}$ is the measured expression of gene g_j. Single-cell models can be formulated using a cell encoder:

$$f : \mathbb{R}^G_{\geq 0} \to \mathbb{R}^d, \qquad \boldsymbol{z}_c = f(\boldsymbol{x}_c), \tag{1}$$

which maps an expression profile to a d-dimensional cell feature $\boldsymbol{z}_c$.

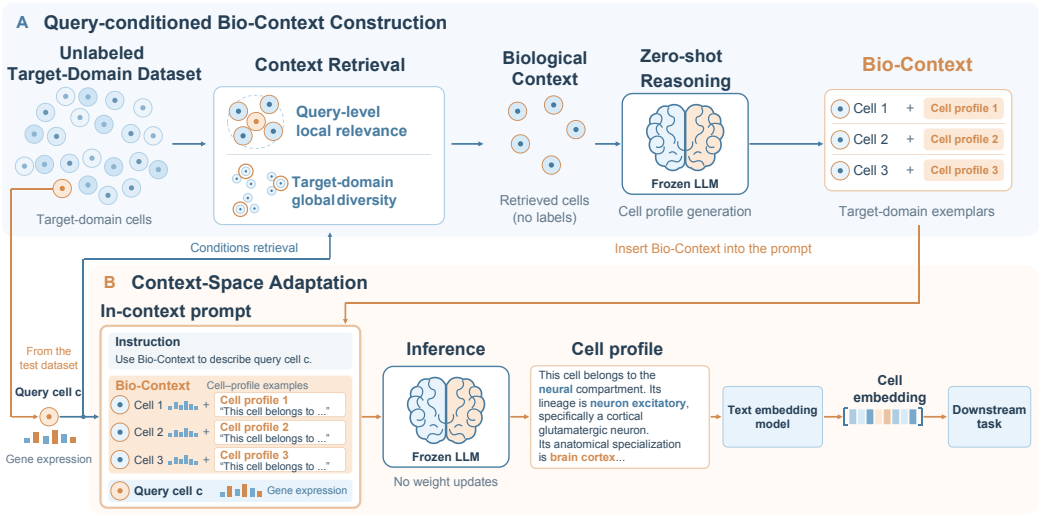


Figure 2: Overview of scCoLLM. (A) Query-conditioned Bio-Context construction. scCoLLM retrieves a query-conditioned biological context by jointly considering query-level local relevance and target-domain global diversity. scCoLLM then generates Bio-Context using zero-shot reasoning. (B) Context-space adaptation. The Bio-Context and the query cell are inserted into an in-context prompt, from which the frozen LLM infers the cell profile of the query cell.

Domain-shift setting. Let P_{S} denote the distribution of the trained data from which cell encoder f acquires its knowledge, and let P_{T} denote the distribution of a newly profiled target domain. At test time, only an unlabeled dataset from target domain is observed, $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{N}, \boldsymbol{x}_i \sim P_{\mathrm{T}}$, whose annotations $y_i \in \mathcal{Y}_{\mathrm{T}}$ are withheld; the target label space $\mathcal{Y}_{\mathrm{T}}$ may contain identities that never occur under P_{S}. We consider the **domain-shift setting** $P_{\mathrm{T}} \neq P_{\mathrm{S}}$, in which there is a significant difference between the two distributions. With a slight abuse of notation, we identify each cell by $c \in \mathcal{D}$.

From parameter space to context space. The conventional remedy for domain shift is **parameter-space adaptation**, which updates the parameters θ of an encoder f_θ on target data $\mathcal{D}^{\mathrm{ad}}$:

$$\theta_{\mathrm{T}} = \arg\min_{\theta'} \frac{1}{|\mathcal{D}^{\mathrm{ad}}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}^{\mathrm{ad}}} \ell\big(h(f_{\theta'}(\boldsymbol{x})),\, y\big), \qquad \boldsymbol{z}_c = f_{\theta_{\mathrm{T}}}(\boldsymbol{x}_c). \tag{2}$$

Here $h : \mathbb{R}^d \to \mathcal{Y}$ is a predictor over label space $\mathcal{Y}$, ℓ is a loss function, y is either a ground-truth annotation in supervised fine-tuning, or a pseudo-label in self-supervised training on the target data. Every instance of Eq. (2) therefore requires an additional training procedure for each new target domain, with gradient-based optimization and white-box access to model parameters θ, and its supervised variants additionally require massive ground truth annotations.

Parameter-space adaptation is costly. We therefore study **context-space adaptation**, in which the parameters remain fixed and adaptation is carried by a context constructed from the unlabeled $\mathcal{D}$:

$$\boldsymbol{z}_c = f_\theta\big(\boldsymbol{x}_c;\, \mathcal{C}(c,\mathcal{D})\big), \tag{3}$$

where $\mathcal{C}(c,\mathcal{D})$ is the constructed context. This transductive setting is reasonable and reflects the practice of single-cell analysis, where an entire target dataset is typically available. In-context learning makes Eq. (3) realizable with LLMs: conditioning on a prompt alters the predictive distribution of the model without any parameter update (Brown et al., 2020; Xie et al., 2021). scCoLLM instantiates f_θ by composing a frozen LLM $\mathcal{M}$ with a frozen text embedding model $\mathcal{E}$, both accessed only through their inference interfaces, and instantiates $\mathcal{C}(c,\mathcal{D})$ with the Bio-Context of Section 3.2.

Textual interface. Because $\mathcal{M}$ operates on token sequences over a vocabulary $\mathcal{V}$, each cell c is serialized into a cell sentence $\tau(\boldsymbol{x}_c)$ (Levine et al., 2024) using a deterministic map $\tau : \mathbb{R}^G_{\geq 0} \to \mathcal{V}^*$. Appendix C.2 specifies the gene-ranking rule and serialization details.

3.2 BIO-CONTEXT CONSTRUCTION

The first stage of scCoLLM constructs a Bio-Context tailored to each query cell (Figure 2A). It selects reference cells called biological context set from the unlabeled target dataset and pairs their expression profiles with descriptions generated by the frozen LLM. Selection balances two complementary requirements: *query-level local relevance*, which provides close biological references, and *target-domain global diversity*, which supplies context from the broader target population.

Reference pool. We sample a context set $\mathcal{D}_{\text{ctx}}$ uniformly at random, from the entire target dataset $\mathcal{D}$, using 2% of its cells. This set is sampled once and reused across queries. For each query c, we exclude the query itself from the candidate pool, thereby achieving query inference:

$$\mathcal{D}_{\text{ctx}} \subset \mathcal{D}, \qquad |\mathcal{D}_{\text{ctx}}| = \lceil 0.02N \rceil, \qquad \mathcal{Z}_c = \mathcal{D}_{\text{ctx}} \setminus \{c\}, \qquad n_c = |\mathcal{Z}_c|. \tag{4}$$

Query-conditioned biological context selection. We instantiate the biological context as a set of reference cells S_c. Drawing inspiration from in-context example selection methods (Liu et al., 2022; Rubin et al., 2022; Ye et al., 2023) and model the utility of the biological context using determinantal point process that combines local relevance with global diversity.

Let $\phi : \mathbb{R}^G_{\geq 0} \to \mathbb{S}^{d_u - 1}$ be a fixed, label-free feature map and $\boldsymbol{u}_i = \phi(\boldsymbol{x}_i)$ the unit-normalized feature of cell i (Appendix C specifies it), $\boldsymbol{u}_i^\top \boldsymbol{u}_c$ is the local relevance of a candidate $i \in \mathcal{Z}_c$ to the query c. Let $\boldsymbol{L} \in \mathbb{R}^{n_c \times n_c}$ be the Gram matrix with $L_{ij} = \boldsymbol{u}_i^\top \boldsymbol{u}_j$. It is positive semidefinite with unit diagonal. Following Kulesza & Taskar (2012), we incorporate query-conditioned relevance by

$$\boldsymbol{L}^{(c)} = \operatorname{Diag}\left(\exp\left(\boldsymbol{w}^{(c)}\right)\right) \boldsymbol{L} \operatorname{Diag}\left(\exp\left(\boldsymbol{w}^{(c)}\right)\right), \qquad w_i^{(c)} = \frac{\boldsymbol{u}_i^\top \boldsymbol{u}_c}{2\lambda}, \tag{5}$$

where $\lambda > 0$ controls the balance between the two criteria. For a subset $S \subseteq \mathcal{Z}_c$, this kernel gives

$$\log\det\left(\boldsymbol{L}_S^{(c)}\right) = \underbrace{\frac{1}{\lambda} \sum_{i \in S} \boldsymbol{u}_i^\top \boldsymbol{u}_c}_{\text{local relevance}} + \underbrace{\log\det\left(\boldsymbol{L}_S\right)}_{\text{global diversity}}, \tag{6}$$

where $\boldsymbol{L}_S = [L_{ij}]_{i,j \in S}$. The determinant is the squared volume spanned by the selected features: it penalizes redundant references and rewards complementary ones. Small λ emphasizes local relevance, whereas larger λ gives diversity greater weight; intermediate values combine close references with a broader view of the target population. Section 4.5 examines this balance empirically.

We seek the size-K biological context set with maximum determinantal point process probability, equivalently the maximum a posteriori subset:

$$S_c^\star = \underset{S \subseteq \mathcal{Z}_c,\ |S| = K}{\arg\max} \log\det\left(\boldsymbol{L}_S^{(c)}\right). \tag{7}$$

We take $K \leq \operatorname{rank}(\boldsymbol{L}) \leq \min(n_c, d_u)$ so that a size-K subset with positive determinant exists. Problem (7) is NP-hard (Ko et al., 1995), and we use greedy approximate maximum a posteriori inference Chen et al. (2018): starting from $S^{(0)} = \varnothing$, it adds the candidate j_t from $\mathcal{Z}_c$ by:

$$j_t = \underset{i \in \mathcal{Z}_c \setminus S^{(t-1)}}{\arg\max} \left[\log\det\left(\boldsymbol{L}^{(c)}_{S^{(t-1)} \cup \{i\}}\right) - \log\det\left(\boldsymbol{L}^{(c)}_{S^{(t-1)}}\right)\right], \tag{8}$$

with $S^{(t)} = S^{(t-1)} \cup \{j_t\}$, and $S_c = S^{(K)}$ is the biological context obtained after K selections.

From biological context to Bio-Context. We make that biological context accessible to the LLM by pairing each reference $s \in S_c = (\boldsymbol{x}_{j_1}, ..., \boldsymbol{x}_{j_K})$ with a structured cell profile $\hat{\mathcal{T}}_s$ using zero-shot reasoning by LLM. Each profile contains five text attributes: `broad_compartment`, `lineage`, `sublineage`, `maturity_or_developmental_stage` and `anatomical_specialization`. These describe cell identity at successive levels of granularity together with developmental and anatomical context (Appendix B). The Bio-Context is the resulting sequence of gene expression–cell profile pairs, ordered by greedy selection:

$$\mathcal{B}_c = \left(\left(\tau(\boldsymbol{x}_{j_1}), \hat{\mathcal{T}}_{j_1}\right), \ldots, \left(\tau(\boldsymbol{x}_{j_K}), \hat{\mathcal{T}}_{j_K}\right)\right). \tag{9}$$

3.3 Inference with Bio-Context

The second stage uses the Bio-Context to interpret the query cell in its target-domain background (Figure 2B) and condition scCoLLM to produce cell features for downstream analysis.

Bio-Context-conditioned cell profiling and embedding. We place the Bio-Context $\mathcal{B}_c$ and the serialized query $\tau(\boldsymbol{x}_c)$ in a shared prompt to the frozen LLM. The resulting cell profile $\mathcal{T}_c = \mathcal{M}\big(\mathcal{B}_c, \tau(\boldsymbol{x}_c)\big)$ describes the query cell c using both its gene-expression pattern and the target-domain context represented in the Bio-Context $\mathcal{B}_c$. A frozen text embedding model $\mathcal{E} : \mathcal{V}^* \to \mathbb{R}^d$ encodes the cell profile, followed by ℓ_2 normalization: $\boldsymbol{z}_c = \frac{\mathcal{E}(\mathcal{T}_c)}{\|\mathcal{E}(\mathcal{T}_c)\|_2}$. The shared profile format organizes every cell along the same biological attributes, providing a common semantic basis for the embedding. Writing $\bar{\mathcal{E}}$ for the normalized embedding map, the cell feature from scCoLLM is:

$$\boldsymbol{z}_c = f_{\text{scCoLLM}}\big(\boldsymbol{x}_c; \mathcal{D}\big) = f_{\text{scCoLLM}}\big(\boldsymbol{x}_c; \mathcal{C}(c, \mathcal{D}_{\text{ctx}})\big) = \bar{\mathcal{E}}\Big(\mathcal{M}\big(\mathcal{B}_c, \tau(\boldsymbol{x}_c)\big)\Big). \tag{10}$$

The resulting cell features are passed to downstream tasks. Target-domain adaptation through the query-dependent Bio-Context realizes the context-space adaptation with unchanged parameters.

4 Experiments

4.1 Experimental Setup

We evaluate on 3 domain-shift datasets from benchmark (Yuan et al., 2024): $\mathcal{D}_{\text{ct}}$ (cross-cell-type), $\mathcal{D}_{\text{ts}}$ (cross-tissue), and $\mathcal{D}_{\text{dn}}$ (cross-donor). On all three, scCoLLM transfers a frozen general-purpose LLM pre-trained on natural language to single-cell expression analysis without single-cell-specific training, facing a substantial domain shift in both input domain and task. **The experiment includes 12 baselines in 3 major categories:** 6 single-cell foundation models (Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), scTab (Fischer et al., 2024), UCE (Rosen et al., 2026), Stack (Dong et al., 2026), and scCello (Yuan et al., 2024)), 4 dataset-specific methods (Raw Data, Seurat (Stuart et al., 2019), Harmony (Korsunsky et al., 2019), and scVI (Lopez et al., 2018)); and 2 LLM-based methods (GPTCelltype (Hou & Ji, 2024) and Cell-o1 (Fang et al., 2026)). We also compared scCoLLM-V, a simplified version which employs k-NN majority voting (bypassing LLM inference) using samples from Bio-Context while keeping others unchanged. **Evaluation metrics including** normalized mutual information (NMI), adjusted Rand index (ARI), average silhouette width (ASW), and their mean, AvgBio. More details are given in Appendices E and F.

4.2 Single-cell Analysis under Domain Shift: Clustering Results

We first evaluate whether scCoLLM produces cell features that preserve cell-type structure under cross-cell-type, cross-tissue, and cross-donor shifts. Table 1 shows clustering performance.

Overall performance. scCoLLM achieves the highest AvgBio under all three shifts. It reaches 0.802 on $\mathcal{D}_{\text{ct}}$, 0.712 on $\mathcal{D}_{\text{ts}}$, and 0.704 on $\mathcal{D}_{\text{dn}}$, exceeding the strongest comparable baseline by 0.030, 0.007, and 0.061, respectively. The gains across distinct shifts indicate that target-domain Bio-Context supports transferable cell representations without dataset-specific training or ground truth labels. Furthermore, scCoLLM-V significantly underperforms scCoLLM, indicating that only a simple kNN is insufficient and LLMs reasoning plays a crucial role.

Largest gain under donor shift is consistent with the motivation. On $\mathcal{D}_{\text{dn}}$, scCoLLM improves AvgBio over scCello from 0.643 to 0.704 and ARI from 0.486 to 0.692, showing that the improvement includes closer agreement with annotated cell identities. This result is consistent with the motivation for context-space adaptation: biological knowledge acquired during pretraining can be interpreted using reference cells that reflect the expression background of the target donor. The ablations in Section 4.4 further examine the contribution of this target-domain evidence.

4.3 Cell-Type Annotation

We next test whether the frozen cell features also support supervised annotation on $\mathcal{D}_{\text{ct}}$, using labeled target-domain cells to train a classifier. Table 2 shows the results and details are in Appendix F.

Table 1: Clustering performance under $\mathcal{D}_{\mathrm{ct}}$, $\mathcal{D}_{\mathrm{ts}}$, and $\mathcal{D}_{\mathrm{dn}}$ domain shifts. In each column, the best and second-best comparable results are bold and underlined, respectively. For scCoLLM, standard deviations appear below the means. "/" denotes unavailable results (Appendix E.2). LLM-based baselines use predicted cell types as clusters for NMI/ARI. Their ASW is computed in expression PCA space, so ASW and AvgBio (†) are not directly comparable (Appendices E.2 and E.3).

	$\mathcal{D}_{\mathrm{ct}}$ (cross-cell-type)				$\mathcal{D}_{\mathrm{ts}}$ (cross-tissue)				$\mathcal{D}_{\mathrm{dn}}$ (cross-donor)			
Model	NMI	ARI	ASW	**AvgBio**	NMI	ARI	ASW	**AvgBio**	NMI	ARI	ASW	**AvgBio**
Geneformer	0.846	0.697	0.525	0.689	0.787	0.499	0.505	0.597	0.686	0.327	0.433	0.482
scGPT	0.866	0.705	0.551	0.707	0.794	0.556	0.531	0.627	0.677	0.298	0.456	0.477
scTab	0.886	**0.807**	0.584	0.759	0.815	0.616	0.541	0.657	/	/	/	/
UCE	0.902	<u>0.802</u>	0.612	<u>0.772</u>	0.836	0.610	0.562	0.670	0.737	0.284	0.496	0.506
scCello	0.887	0.781	<u>0.640</u>	0.769	<u>0.839</u>	<u>0.675</u>	<u>0.601</u>	<u>0.705</u>	<u>0.794</u>	<u>0.486</u>	**0.649**	<u>0.643</u>
Raw Data	0.864	0.718	0.529	0.703	0.800	0.585	0.508	0.631	0.665	0.251	0.462	0.460
Seurat	0.893	0.773	0.590	0.752	0.813	0.560	0.535	0.636	0.711	0.335	0.420	0.489
Harmony	0.553	0.241	0.432	0.432	0.684	0.400	0.460	0.515	0.690	0.324	0.408	0.474
scVI	<u>0.905</u>	0.797	0.577	0.760	0.816	0.550	0.537	0.634	0.722	0.311	0.471	0.502
GPTCelltype	0.855	0.710	0.618†	0.728†	0.780	0.572	0.546†	0.633†	0.670	0.302	0.374†	0.449†
Cell-o1	0.667	0.349	0.485†	0.500†	0.679	0.251	0.444†	0.458†	0.652	0.125	0.417†	0.398†
scCoLLM-V	0.873	0.725	0.616	0.738	0.823	0.604	0.428	0.618	0.725	0.433	0.265	0.474
scCoLLM	**0.905** ±0.002	0.790 ±0.003	**0.710** ±0.002	**0.802** ±0.003	**0.841** ±0.003	**0.681** ±0.008	**0.614** ±0.003	**0.712** ±0.003	**0.816** ±0.002	**0.692** ±0.004	<u>0.603</u> ±0.005	**0.704** ±0.001

Table 2: Cell-type annotation on $\mathcal{D}_{\mathrm{ct}}$ with an MLP classifier trained on frozen features. Scores are means ± standard deviations over three runs with different random seeds.

	Single-Cell Foundation Models		Dataset-specific Models	Ours
Metric	UCE	scCello	Seurat	**scCoLLM**
Acc	$\mathbf{0.951}_{\pm 0.003}$	$0.936_{\pm 0.002}$	$0.882_{\pm 0.002}$	$\underline{0.945}_{\pm 0.003}$
Macro-F1	$\underline{0.819}_{\pm 0.005}$	$0.792_{\pm 0.016}$	$0.693_{\pm 0.007}$	$\mathbf{0.832}_{\pm 0.012}$
Weighted-F1	$\mathbf{0.958}_{\pm 0.002}$	$0.939_{\pm 0.002}$	$0.896_{\pm 0.001}$	$\underline{0.951}_{\pm 0.003}$
BAcc	$\underline{0.910}_{\pm 0.008}$	$0.892_{\pm 0.016}$	$0.855_{\pm 0.009}$	$\mathbf{0.925}_{\pm 0.017}$

scCoLLM achieves strong annotation performance across cell types. scCoLLM achieves the highest macro-F1 (0.832) and balanced accuracy (0.925), exceeding UCE by 0.013 and 0.015, respectively. Accuracy and weighted-F1 remain within 0.007 of UCE, which leads these frequency-weighted metrics. Thus, the cell representation from scCoLLM supports accurate downstream annotation while retaining competitive performance under both ways of weighting cell types.

Complementarity with the label-free evaluation. scCoLLM leads UCE in clustering and remains competitive after supervised classifier training. The comparison of them under these two settings illustrates that the Bio-Context provides information about the target domain that would otherwise have to be supplied by supervision. The two evaluations establish the usefulness of features from scCoLLM both when target labels are unavailable for adaptation and available for annotation.

4.4 Importance of Bio-Context

We examine whether the gains arise from target-domain Bio-Context rather than from more examples alone. On $\mathcal{D}_{\mathrm{ct}}$, we compare Bio-Context with query-only, random-context, cross-domain-context, relevance-only-context variants, together with an oracle using annotated cells. Table 3 evaluates both clustering and the accuracy of generated cell profiles. More details are in Appendix G.

Bio-Context improves both representations and profiles. Adding Bio-Context raises AvgBio over query-only variant from 0.702 to 0.802 and improves every clustering metric. Profile-attribute accuracy increases from 0.719 to 0.788 and macro-F1 from 0.596 to 0.654. Since these 2 variants

Table 3: Effect of the context supplied to scCoLLM on $\mathcal{D}_{\text{ct}}$. Relevance-only context uses kNN.

	Label-free conditioning					Oracle
Metric	(i) Bio-Context (**scCoLLM**)	(ii) Query-only	(iii) Random context	(iv) Cross-domain context	(v) Relevance-only context	(vi) Ground-truth labels
NMI	**0.905**	0.829	0.567	0.747	0.879	0.943
ARI	**0.790**	0.685	0.323	0.480	0.705	0.896
ASW	**0.710**	0.592	0.596	0.491	0.704	0.805
AvgBio	**0.802**	0.702	0.462	0.573	0.763	0.881
Accuracy	**0.788**	0.719	0.488	0.261	0.773	0.972
Macro-F1	**0.654**	0.596	0.291	0.239	0.618	0.897

share the query, backbone, and remaining pipeline, the comparison isolates the contribution of Bio-Context. The improvement in the profile attributes also links the feature gains to more accurate biological descriptions, rather than to clustering alone. Appendix L shows a correction case study.

Query relevance and target-domain context make Bio-Context a domain-specific biological context. Random target-domain context reduces AvgBio to 0.462, while query-conditioned context from another domain yields 0.573; both are below the query-only. The relevance-only variant retrieves cells improves upon query-only variant but still lags behind scCoLLM. Thus, neither access to target-domain without query conditioning, nor query-conditioned from another domain, nor considering query relevance alone reproduces the benefit of Bio-Context. Only their combination improves every metric. The oracle variant reaches high AvgBio further confirming that Bio-Context contain useful information about the query.

4.5 Balancing Local Relevance and Global Diversity

We test whether balancing local relevance and global diversity yields informative Bio-Contexts. Figure 3 shows how λ trades off these two properties. See Appendix H for more details.

Balancing relevance and coverage of Bio-Context when $\lambda = 0.1$. Pure-relevance retrieval ($\lambda \to 0$ limit) selects references with high query similarity (0.920) and same-type rate (93.5%), but the references span only 1.20 cell types on average. At the default $\lambda = 0.1$, this rises to 2.03 types, and answer coverage increases from 81.5% to 85.0%. Increasing λ beyond 0.1 yields only marginal additional coverage while further reducing relevance. These results support $\lambda = 0.1$ as a practical balance: diversity broadens the evidence supplied by Bio-Context, while relevance keeps that evidence tied to the query. The two retrieval criteria consequently serve complementary roles.

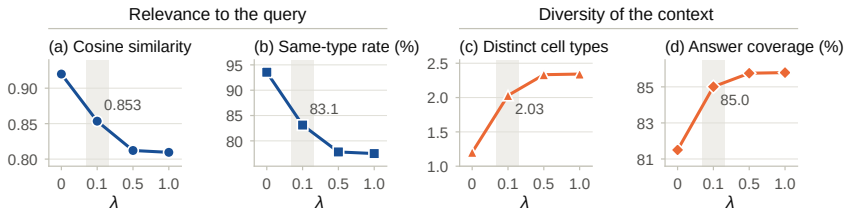


Figure 3: Relevance metrics (cosine similarity and same-type rate) and diversity metrics (distinct cell types and answer coverage) of the Bio-Context as functions of the trade-off λ on $\mathcal{D}_{\text{ct}}$.

4.6 Robustness to Bio-Context Noise and Variations in Context Size

We examine robustness to errors in generated cell profiles of Bio-Context and to the size K of Bio-Context. The noise and context-size protocols are detailed in Appendices I and J. **scCoLLM is robust to noisy Bio-Context.** Figure 4 shows that corrupting 10% or 30% of cell profiles reduces AvgBio by less than 1.5%, while profile-attribute accuracy and macro-F1 change slightly. Even when half of the cell profiles are corrupted, AvgBio decreases by less than 6%, and attribute accuracy

and macro-F1 decline by only 0.010 and 0.024. These results support robustness to errors in the Bio-Context. **scCoLLM is robust to Bio-Context size.** Performance in clustering and profile-attribute accuracy remains stable from $K = 5$ to $K = 64$ (Figure 4).

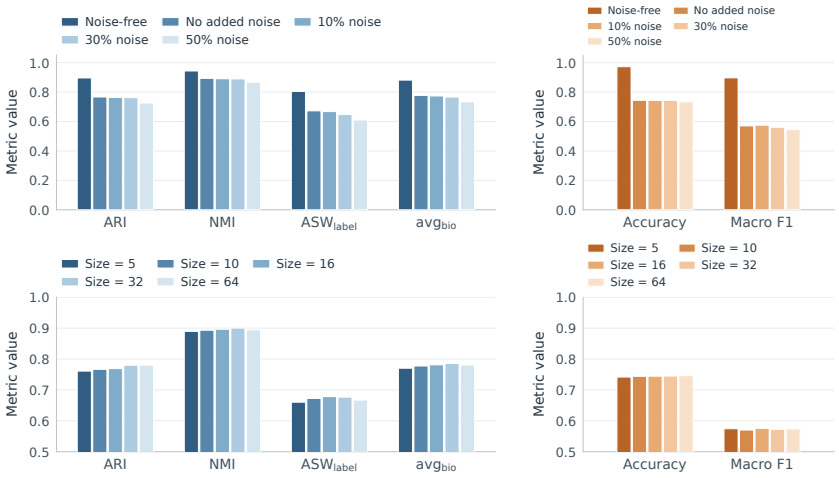


Figure 4: Robustness of scCoLLM to Bio-Context noise (above) and size (below).

4.7 Portability Across LLM Backbones

Table 4 shows scCoLLM with different LLM backbones. More details are in Appendix K.

Portability of scCoLLM across backbones. GLM-5.3 closely matches GPT-5.6 in clustering and profile attributes. GPT-oss-120B reaches 0.770, comparable to the strongest baselines UCE and scCello, while GPT-oss-20B reaches 0.753, close to Seurat. scCoLLM with GPT-oss-20B also outperforms the much larger query-only GPT-5.6, highlighting the pivotal role played by the Bio-Context in scCoLLM, a capability unattainable through the use of LLM alone. Thus, competitive performance is retained across LLM backbones, without single-cell-specific model training. Moving from the smaller 20B to the larger 120B model raises performance showing that LLM backbone improvements can translate into better cell profiles and cell embeddings.

Table 4: scCoLLM with different LLM backbones on the cross-cell-type dataset $\mathcal{D}_{\text{ct}}$.

		Proprietary	Open-weight		
Evaluation	Metric	GPT-5.6	GLM-5.3	GPT-oss-120B	GPT-oss-20B
Clustering	NMI	$\underline{0.904}_{\pm 0.002}$	$\mathbf{0.906}_{\pm 0.005}$	0.895	0.861
	ARI	$\mathbf{0.790}_{\pm 0.003}$	$\underline{0.784}_{\pm 0.013}$	0.761	0.696
	ASW	$\mathbf{0.710}_{\pm 0.002}$	$\underline{0.703}_{\pm 0.007}$	0.654	$\underline{0.703}$
	AvgBio	$\mathbf{0.802}_{\pm 0.003}$	$\underline{0.798}_{\pm 0.007}$	0.770	0.753
Profile Attributes	Accuracy	**0.788**	**0.788**	$\underline{0.766}$	0.744
	Macro-F1	$\underline{0.654}$	**0.656**	0.600	0.570

5 Conclusion

We introduced *Context-Space Adaptation* for single-cell analysis under domain shift and instantiated it as scCoLLM. By selecting biological context from unlabeled target domain that balance local relevance with global diversity, scCoLLM constructs a Bio-Context that conditions LLM to generate cell profiles for downstream task. Adaptation requires neither parameter updates nor new ground truth labels. Experiments demonstrate state-of-the-art performance across domain shifts. Further experiments confirmed the soundness of the Bio-Context design and its decisive role in enhancing model performance, as well as the robustness and transferability of scCoLLM.

REPRODUCIBILITY STATEMENT

The code that implements scCoLLM is provided in the supplementary material. Appendix C specifies every component of the pipeline: the frozen backbones and the settings under which they are queried, the text embedding model and the feature dimension it fixes, the serialization length, the retrieval hyperparameters and the size of the context set, and the prompt template reproduced verbatim. Reproduction requires access to a language model, either through paid API calls to the proprietary backbone used in the main experiments (GPT-5.6) or through a local deployment of the open-weight backbones (the GPT-oss family and GLM-5.3), which incurs no LLM API cost; Appendix D reports the measured cost per cell for both routes. The three evaluation datasets are those released with scCello (Yuan et al., 2024), whose data preprocessing we follow; Appendix E.1 summarizes their construction and statistics, and Appendix E.3 states the evaluation protocol.

AI USE STATEMENT

Generative AI tools assisted with drafting and revising this manuscript, and preparing LaTeX formatting. The authors are responsible for reviewing these outputs and verifying the final text, references, and empirical claims.

REFERENCES

Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel JT Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome biology*, 20(1):194, 2019.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Rebecca Boiarsky, Nalini M Singh, Alejandro Buendia, Ava P Amini, Gad Getz, and David Sontag. Deeper evaluation of a single-cell foundation model. *Nature Machine Intelligence*, 6(12):1443–1446, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in neural information processing systems*, 31, 2018.

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(8):1470–1480, 2024.

Mingze Dong, Abhinav Adduri, Dhruv Gautam, Christopher Carpenter, Rohan Shah, Chiara Ricci-Tam, Yuval Kluger, Dave P Burke, and Yusuf H Roohani. Stack: In-context learning of single-cell biology. *bioRxiv*, 2026.

Yin Fang, Qiao Jin, Guangzhi Xiong, Bowen Jin, Xianrui Zhong, Siru Ouyang, Yifan Yang, Aidong Zhang, Jiawei Han, and Zhiyong Lu. Cell-o1: training llms to solve single-cell reasoning puzzles with reinforcement learning. *Bioinformatics*, 42(5):btag208, 2026.

Felix Fischer, David S Fischer, Roman Mukhin, Andrey Isaev, Evan Biederstedt, Alexandra-Chloé Villani, and Fabian J Theis. sctab: Scaling cross-tissue single-cell annotation models. *Nature Communications*, 15(1):6611, 2024.

Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology*, 42(2): 293–304, 2024.

Wenpin Hou and Zhicheng Ji. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature methods*, 21(8):1462–1465, 2024.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Kasia Z Kedzierska, Lorin Crawford, Ava P Amini, and Alex X Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025.

Chun-Wa Ko, Jon Lee, and Maurice Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. Cell2sentence: teaching large language models the language of biology. *BioRxiv*, pp. 2023–09, 2024.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pp. 100–114, 2022.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Ajay Nadig, Akshaya Thoutam, Madeline Hughes, Anay Gupta, Andrew W Navia, Nicolo Fusi, Srivatsan Raghavan, Peter S Winter, Ava P Amini, and Lorin Crawford. Consequences of training data composition for deep learning models in single-cell biology. *bioRxiv*, 2025.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, 2026. Accessed: 2026-09-17.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, et al. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic acids research*, 53(D1):D886–D900, 2025.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embedding provides a foundation model for cell biology. *Nature*, pp. 1–9, 2026.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 2655–2671, 2022.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pp. 39818–39833. PMLR, 2023.

Xinyu Yuan, Zhihao Zhan, Zuobai Zhang, Manqi Zhou, Jianan Zhao, Boyu Han, Yue Li, and Jian Tang. Cell ontology guided transcriptome foundation model. *Advances in Neural Information Processing Systems*, 37:6323–6366, 2024.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

A A Second Instance of the Domain Gap

Figure 1 illustrates the gap that motivates this work with T cells from two domains. Because a single example could reflect a property of that particular cell type or tissue pair, we report a second case here. Figure 5 contrasts endothelial cells of the exocrine pancreas with endothelial cells of the myometrium, again taken from the cross-tissue dataset $\mathcal{D}_{\text{ts}}$ and again matched in source dataset, assay, and health status while differing in donor.

The two groups carry the same cell-type annotation, yet they occupy distinguishable regions of the joint principal-component space, with a silhouette score of 0.443 when grouped by tissue (Figure 5a), and individual genes separate them in both directions: *FABP5* and *TXNIP* are expressed in the pancreatic endothelial cells and almost absent from the myometrial ones, whereas *SERPINE1* and *SOD2* show the opposite pattern (Figure 5b). The split is less pronounced than for T cells, whose silhouette is 0.667, which is itself informative: the size of the gap varies with the cell type and the tissue pair, so a model cannot be designed on the assumption that it is small. In both cases the identity to be recovered is unchanged across the two tissues while the expression evidence for it is not. Target-domain context is intended to help the model interpret these domain-specific expression patterns.

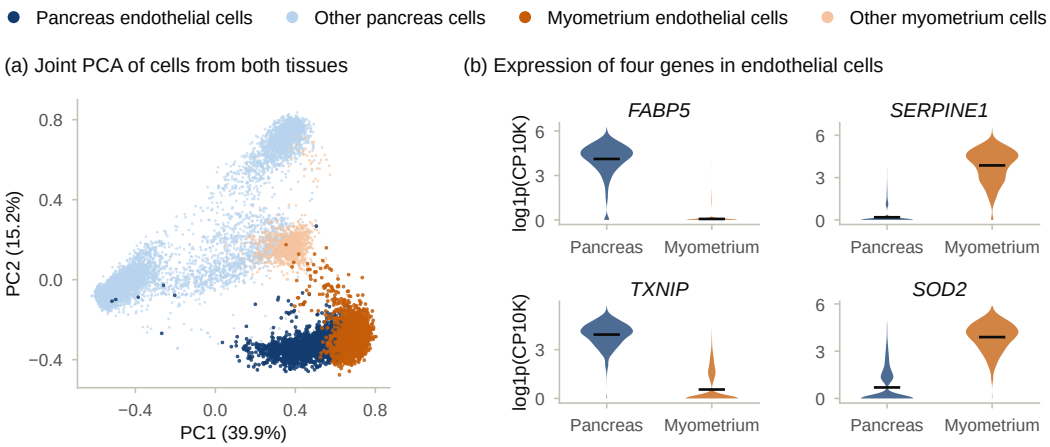


Figure 5: Endothelial cells differ between the exocrine pancreas and the myometrium despite sharing the same cell-type annotation, a second instance of the shift shown for T cells in Figure 1. Both tissues are taken from the cross-tissue dataset $\mathcal{D}_{\text{ts}}$ and come from the same source dataset, assay, and healthy condition, but from different donors. (a) Joint PCA of all 14,565 cells, comprising 1,927 pancreatic endothelial cells, 1,972 myometrial endothelial cells, and 9,613 and 1,053 other cells from the two tissues, respectively. The endothelial cells of the two tissues occupy distinguishable regions, with a silhouette score of 0.443 when grouped by tissue. (b) Expression of four genes in these endothelial cells, measured as log1p(CP10K) of the genes retained in the model input, with genes absent from the input set to zero; black bars mark the means.

B The Cell Profile

Section 3.2 introduces the cell profile, the object that the frozen LLM produces for every reference cell and, conditioned on the Bio-Context, for the query. This appendix states its five fields and explains the choice. The same five fields are used for every dataset, every context size, and every backbone reported in Section 4; none of them is adapted to a target domain.

During the evaluation phase, to calculate prediction accuracy (accuracy and Macro-F1) for the fields within the cell profiles, we manually curated the ground-truth cell profiles corresponding to each cell type in the dataset. Note that this portion of the data is not used for inference or the generation of cell embeddings; it serves solely to evaluate model performance.

Table 5: The five fields of a cell profile. The model fills them by reasoning from the genes the cell expresses. The fields are free text and the listed values are illustrative only.

Field	What it states	Illustrative values
`broad_compartment`	The compartment of the tissue to which the cell belongs	epithelial, immune, stromal, endothelial
`lineage`	The developmental lineage within that compartment	myeloid, lymphoid, fibroblast, vascular endothelium
`sublineage`	The specific subpopulation within that lineage	CD8-positive T cell, classical monocyte, capillary endothelial cell
`maturity_or_developmental_stage`	How far the cell has progressed through development, differentiation, or activation	progenitor, immature, mature, naive, activated
`anatomical_specialization`	The organ- or site-specific specialization that the expression indicates	pancreatic, myometrial, mucosal, tissue-resident

Graded granularity. The first three fields are nested, each constraining the next, so a profile stays informative even when its finest field is wrong. This matters under distribution shift, where the fine level is the one at which zero-shot reasoning fails first: a reference whose `sublineage` is mistaken still contributes a correct compartment and lineage to the context. The corruption study of Section 4.6 probes the extreme form of this situation, in which all five fields of a reference are corrupted, and shows that the method degrades gracefully.

Two axes that a single identity term cannot carry. Cells that agree on all three hierarchical fields still differ in how far they have progressed through development or activation and in the organ in which they reside. The second of these is exactly what a tissue shift changes: Figure 1 and Figure 5 show two cell types whose annotation is constant across tissues while their expression is not. `maturity_or_developmental_stage` and `anatomical_specialization` give the model a place to record that variation, instead of forcing it to absorb the variation into a coarser or a finer identity term, or to omit it.

Commensurability of references and query. The fields are the same for every reference and for the query, so the text embedding compares descriptions written along the same axes. Distance in the resulting feature space then reflects disagreement about identity rather than differences in what the model happened to mention, which is the property the downstream clustering and classification rely on.

C Implementation Details

This appendix states the components and settings that define scCoLLM in the reported experiments. None of them is tuned per dataset: a single configuration is used for $\mathcal{D}_{\mathrm{ct}}$, $\mathcal{D}_{\mathrm{ts}}$, and $\mathcal{D}_{\mathrm{dn}}$ alike.

C.1 Frozen Models

The default backbone $\mathcal{M}$ is GPT-5.6 (OpenAI, 2026), queried through its API with reasoning enabled. The noise-robustness and context-size studies in Section 4.6 use GPT-oss-20B (Agarwal et al., 2025). The backbone study of Section 4.7 additionally evaluates GLM-5.3 (Zeng et al., 2026), GPT-oss-120B, and GPT-oss-20B. The text embedding model $\mathcal{E}$ is `openai/text-embedding-3-small` in all experiments; its output dimension fixes the cell feature at $d = 1536$. Both models are used strictly through their inference interfaces, and no weight of either is read or updated.

C.2 GENE-EXPRESSION SERIALIZATION

Following the rank-value encoding of Theodoris et al. (2023) and the rank-based cell sentences of Levine et al. (2024), τ lists the symbols of the m genes that are most *enriched* in the cell, in decreasing order of enrichment,

$$\tau(\boldsymbol{x}_c) = \big(g_{\pi_c(1)}, \ldots, g_{\pi_c(m)}\big), \qquad \tilde{x}_{c,j} = \frac{x_{c,j}}{\mathrm{med}_j}, \\ \tilde{x}_{c,\pi_c(1)} \geq \tilde{x}_{c,\pi_c(2)} \geq \cdots \geq \tilde{x}_{c,\pi_c(G)}, \tag{11}$$

where med_j is the median expression of gene g_j across a large reference corpus and π_c is a permutation of $\{1, \ldots, G\}$ with ties broken by a fixed gene order. Normalizing by med_j makes the ranking reflect how unusual a gene is for this cell rather than how abundant it is in general.

The serializer τ of Eq. (11) emits at most $m = 2048$ gene symbols per cell, so a cell with more detected genes is truncated after its 2048 most enriched ones. The symbols are presented in tiers of decreasing rank, which signals to the model how much evidence each part of the list carries.

C.3 RETRIEVAL

The feature map ϕ returns dimension $d_u = 256$ and is obtained by label-free preprocessing of the expression data. The trade-off of Eq. (6) is $\lambda = 0.1$ and the Bio-Context size is $K = 10$; both are fixed for every dataset and backbone, and Sections 4.5 and 4.6 report the behavior around these values. We draw $\mathcal{D}_{\mathrm{ctx}}$ uniformly at random without replacement from the entire target dataset, with $|\mathcal{D}_{\mathrm{ctx}}| = \lceil 0.02|\mathcal{D}| \rceil$. For each query c, retrieval uses $\mathcal{Z}_c = \mathcal{D}_{\mathrm{ctx}} \setminus \{c\}$ (Eq. (4)), ensuring that the query is excluded from its own Bio-Context.

C.4 PROMPT TEMPLATE

Figure 6 shows the prompt template. Placeholders in square brackets are filled per cell; in the context-conditioned inference of Section 3.3, the serialized Bio-Context is inserted ahead of the query block, and the reference cells are described with the same template and the same five fields.

The simplified variant, scCoLLM-V, utilizes the Bio-Context constructed by scCoLLM but does not generate prompts by inserting this context into the LLM for inference. Instead, it determines the attributes of the cell profile through k-nearest neighbor (kNN) majority voting based on the Bio-Context, while the remaining steps mirror those of scCoLLM. scCoLLM-V aims to investigate the role played by LLM inference within the overall model and to determine whether a simple kNN approach is sufficient to address the task.

```
You are a computational biologist reasoning single cells from
their transcriptome.

You will be given the top-ranked genes of ONE cell. The ranking
is not raw expression: genes are ordered by their expression in
this cell divided by that gene's median expression across a
large reference corpus, so a high rank means the gene is
unusually high for this cell rather than merely abundant.
Reason about which cell identity this profile is characteristic
of, then describe that cell with a structured semantic cell
profile: broad_compartment, lineage, sublineage,
maturity_or_developmental_stage, anatomical_specialization

Cell [cell_id]

Top [num_genes] genes, ranked by how much each gene is enriched
in this cell relative to that gene's typical level across a
large reference corpus (rank 1 = most enriched; this suppresses
housekeeping genes that are high in every cell). They are
grouped into tiers by rank; earlier tiers carry more evidence
about cell identity:

[genes list]
```

Figure 6: The prompt template used by scCoLLM. The same template defines the five fields of the cell profile for Bio-Context cells and for queries, so that the two are described along identical axes.

D Inference Cost

Because scCoLLM performs no training, its entire computational cost is inference: one description call per cell of the context set, which is cached and reused across queries, and one profiling call per query cell. Every call uses the same template and differs only in the serialized cells it carries, so the cost is both easy to measure and easy to predict in advance.

For API cost: cell costs \$0.000687 with GPT-5.6 and \$0.000495 with GLM-5.3, that is, \$0.69 and \$0.49 per thousand cells. These are upper bounds on what is actually paid: in the prompts, the system prompt is identical across all cells, and the instruction-related prefix within the user prompt is also consistent; therefore, prompt caching can be leveraged to reduce costs.

The two open-weight backbones remove the LLM API cost altogether. GPT-oss-120B and GPT-oss-20B are released with their weights and are straightforward to serve locally, and Section 4.7 shows that scCoLLM built on them remains strong: GPT-oss-120B attains an AvgBio of 0.770 on $\mathcal{D}_{\mathrm{ct}}$, on par with the best pretrained baselines of Table 1 (UCE 0.772, scCello 0.769) and within 0.032 of the proprietary GPT-5.6. scCoLLM can therefore be run at no LLM API cost while retaining performance close to that of the strongest backbone.

E Details of the Experimental Setup

This appendix expands the summary in Section 4.1. Appendix E.1 describes the construction of the three benchmark datasets, Appendix E.2 describes the baselines and their outputs, and Appendix E.3 formalizes the clustering protocol and metrics. The cell-type annotation experiment is detailed separately in Appendix F.

E.1 BENCHMARK DATASETS

Source and curation. The three datasets belong to the benchmark released with scCello (Yuan et al., 2024), which was assembled from the human scRNA-seq collections of the CELLxGENE census (release 2023-07-25) (Program et al., 2025). Three filters were applied during its curation. Cells flagged as non-primary were discarded. Only cells profiled with 10x Genomics protocols were retained, which limits variability introduced by the sequencing technology. Cells from cancer samples were removed, as their expression programs deviate strongly from those of healthy tissue.

Construction of the distribution shifts. Each shift is created by withholding categories along one metadata attribute. For the cell-type, tissue, and donor attributes, a random subset containing 15%, 15%, and 10% of the categories is drawn, respectively, and categories that cover more than 0.1% of the curated cells are excluded from the draw. The cells annotated with a withheld category form the evaluation dataset for that attribute, data statistics are shown in Table 6. The shifts are in this sense *benchmark-defined*: they are fixed by the construction of the benchmark and are identical for every method compared in this work. We denote the resulting cross-cell-type, cross-tissue, and cross-donor datasets by $\mathcal{D}_{\mathrm{ct}}$, $\mathcal{D}_{\mathrm{ts}}$, and $\mathcal{D}_{\mathrm{dn}}$.

Table 6: Statistics of the three benchmark datasets. Bold entries mark the withheld attribute.

Dataset	Shift	#Cells	#Cell types	#Tissues	#Donors	#Conditions	#Batches
$\mathcal{D}_{\mathrm{ct}}$	Cell type	486,810	**87**	125	122	35	90
$\mathcal{D}_{\mathrm{ts}}$	Tissue	341,681	205	**32**	2,052	7	25
$\mathcal{D}_{\mathrm{dn}}$	Donor	2,521,868	404	101	**525**	33	123

Target-domain shift setting for scCoLLM. scCoLLM is based on a general-purpose LLM like GPT. Although the specific composition of its training dataset is not explicitly disclosed, such models are typically trained on vast amounts of natural language text; while this corpus may contain information regarding cells and biology, the data format differs significantly from the scRNA-seq data found in benchmark datasets. Furthermore, scCoLLM has not undergone training or fine-tuning on scRNA-seq datasets or cell-specific tasks. Consequently, scCoLLM transfers a frozen general-purpose LLM—pre-trained on natural language—to single-cell expression analysis without any single-cell-specific training, thereby confronting a substantial domain shift in both the input domain and the task.

Context sampling and query exclusion. For each target dataset $\mathcal{D}$, scCoLLM samples the context set $\mathcal{D}_{\mathrm{ctx}}$ once, uniformly at random without replacement, from the entire dataset, with $|\mathcal{D}_{\mathrm{ctx}}| = \lceil 0.02|\mathcal{D}| \rceil$ (Eq. (4)). Sampling uses no annotation and is fixed before inference. For every evaluated query c, the reference subset S_c is selected from $\mathcal{Z}_c = \mathcal{D}_{\mathrm{ctx}} \setminus \{c\}$, so $c \notin S_c$ even when c belongs to the sampled context set. Evaluation follows the subsets described above and is not restricted to the remaining of cells. With the reference pool and retrieval map fixed, each query is processed separately using only distinct reference cells; the same pool can also serve subsequently arriving queries without rebuilding the reference pool. This supports adaptation at the query-inference stage.

E.2 BASELINES

Feature-based methods yield one vector per cell for the clustering protocol of Appendix E.3; GPT-Celltype and Cell-o1 instead yield cell-type predictions, which are evaluated as described below. Pretrained single-cell models are used as frozen feature extractors, without fine-tuning on the target data. Stack is evaluated separately in Appendix E.4 because its fixed gene vocabulary covers only part of the benchmark genes. GPTCelltype and Cell-o1 are evaluated in this work.

Dataset-specific methods. **Raw Data** takes the expression values of highly variable genes directly as the cell feature; this simple reference has been shown to rival pretrained models in zero-shot evaluations (Kedzierska et al., 2025). **Seurat** (Stuart et al., 2019; Hao et al., 2024) identifies anchors, i.e., pairs of mutually nearest cells across batches, and uses them to align the batches in a shared

low-dimensional space. **Harmony** (Korsunsky et al., 2019) starts from a principal-component embedding and alternates between soft clustering and a cluster-wise linear correction that removes batch-dependent offsets. **scVI** (Lopez et al., 2018) fits a variational autoencoder with a count likelihood and batch covariates, and the inferred latent variables are used as cell features. Seurat, Harmony, and scVI fit their representations directly on the unlabeled target data and therefore serve as unsupervised target-domain baselines. None of these methods uses cell-type annotations.

Single-cell foundation models. **Geneformer** (Theodoris et al., 2023) represents a cell by its genes ranked according to normalized expression and is pretrained with masked gene prediction on about 30 million cells; its cell feature is the mean of the output gene embeddings. **scGPT** (Cui et al., 2024) is a generative transformer pretrained on more than 33 million cells, whose `[CLS]` embedding serves as the cell feature. **scTab** (Fischer et al., 2024) is a feature-attention model trained with cell-type supervision on 22.2 million cells; its learned representation is used as the cell feature. **UCE** (Rosen et al., 2026) builds gene tokens from protein-language-model embeddings of the encoded proteins and is pretrained on 36 million cells; we use its `[CLS]` embedding. **scCello** (Yuan et al., 2024) complements masked gene prediction with objectives derived from cell-type labels and the cell ontology during pretraining on 22 million cells, and likewise provides a `[CLS]` embedding. On $\mathcal{D}_{\mathrm{dn}}$, which comprises about 2.5 million cells, scTab could not be applied because processing the dataset exceeded the available memory (Yuan et al., 2024); its results for this dataset are therefore marked with / in Table 1.

LLM-based methods. **GPTCelltype** (Hou & Ji, 2024) performs cluster-level annotation: marker genes are derived for each cell cluster by differential expression and passed to an LLM, which names the cell type of that cluster, and every cell inherits the annotation of its cluster. We query it with GPT-5.6, matching the default backbone of scCoLLM to control for the choice of LLM. **Cell-o1** (Fang et al., 2026) is an LLM post-trained with supervised fine-tuning on distilled reasoning traces followed by reinforcement learning; given a batch of cells described by their top expressed genes, donor-level context, and a list of candidate cell types, it reasons jointly over the batch and assigns a cell type to each cell. We use the released model without further training. For these two methods, we use the predicted cell types directly as cluster assignments when computing NMI and ARI against the reference annotations. ASW is computed on the principal components of the expression profiles, with the predicted cell types defining the groups (Appendix E.3). Their ASW and AvgBio thus use an external expression space and are not directly comparable with those of feature-based methods. We mark these entries with † in Table 1 and exclude them from the best and second-best ranking.

E.3 Clustering Protocol and Metrics

Notation. Consider an evaluation of N cells with features $\{\boldsymbol{z}_i\}_{i=1}^{N}$ produced by a method and reference annotations $\{y_i\}_{i=1}^{N}$, $y_i \in \mathcal{Y}$. The annotations induce a partition $\mathcal{A} = \{A_1, \ldots, A_{|\mathcal{A}|}\}$ of the cells, where A_u collects the cells of the u-th cell type, and a clustering algorithm yields a partition $\Pi = \{\Pi_1, \ldots, \Pi_{|\Pi|}\}$. We write $n_{uv} = |A_u \cap \Pi_v|$, $a_u = |A_u|$, and $b_v = |\Pi_v|$ for the entries and margins of their contingency table.

Clustering protocol. Cell features are clustered with the graph-based protocol of that benchmark, applied consistently to all feature-based methods. For GPTCelltype and Cell-o1, Π instead groups cells by their predicted cell types, as described in Appendix E.2. All clustering metrics are computed with scib (Luecken et al., 2022).

Normalized mutual information (NMI). The empirical entropies of the two partitions and their mutual information are

$$H(\mathcal{A}) = -\sum_{u=1}^{|\mathcal{A}|} \frac{a_u}{N} \log \frac{a_u}{N}, \qquad H(\Pi) = -\sum_{v=1}^{|\Pi|} \frac{b_v}{N} \log \frac{b_v}{N}, \tag{12}$$

$$I(\mathcal{A}; \Pi) = \sum_{u=1}^{|\mathcal{A}|} \sum_{v=1}^{|\Pi|} \frac{n_{uv}}{N} \log \frac{N\, n_{uv}}{a_u\, b_v}, \tag{13}$$

with the convention $0 \log 0 = 0$. NMI divides the mutual information by the arithmetic mean of the two entropies,

$$\mathrm{NMI}(\mathcal{A}, \Pi) = \frac{2\, I(\mathcal{A}; \Pi)}{H(\mathcal{A}) + H(\Pi)} \in [0, 1], \tag{14}$$

which equals 1 when the partitions coincide and 0 when they are independent.

Adjusted Rand index (ARI). The Rand index counts the pairs of cells on which two partitions agree, i.e., pairs grouped together in both or separated in both. ARI subtracts the agreement expected when the partitions are drawn at random with their cluster sizes held fixed and rescales the result (Hubert & Arabie, 1985):

$$\mathrm{ARI}(\mathcal{A}, \Pi) = \frac{\sum_{u,v} \binom{n_{uv}}{2} - E}{\frac{1}{2}\big[\sum_u \binom{a_u}{2} + \sum_v \binom{b_v}{2}\big] - E}, \qquad E = \frac{\sum_u \binom{a_u}{2} \sum_v \binom{b_v}{2}}{\binom{N}{2}}. \tag{15}$$

ARI equals 1 for identical partitions, has expectation 0 under random labeling, and becomes negative when agreement falls below chance.

Average silhouette width (ASW). ASW measures how well the annotated cell types are separated in the feature space and does not depend on the clustering. For cell i, let α_i denote its mean distance to the other cells sharing its annotation y_i, and β_i the smallest mean distance from cell i to the cells of any other single annotation. The silhouette of cell i (Rousseeuw, 1987) and the rescaled average reported as ASW are

$$s_i = \frac{\beta_i - \alpha_i}{\max(\alpha_i, \beta_i)} \in [-1, 1], \qquad \mathrm{ASW} = \frac{1}{2}\Big(1 + \frac{1}{N}\sum_{i=1}^{N} s_i\Big) \in [0, 1], \tag{16}$$

so that higher values indicate compact and well-separated cell types. For GPTCelltype and Cell-o1, $\boldsymbol{z}_i$ is replaced by the principal-component representation of the expression profile and y_i by the predicted cell type. Their ASW therefore measures separation of the predicted groups in expression space, subject to the comparison convention stated in Appendix E.2.

AvgBio. The summary clustering score is the unweighted mean of the three metrics,

$$\mathrm{AvgBio} = \tfrac{1}{3}\big(\mathrm{NMI} + \mathrm{ARI} + \mathrm{ASW}\big). \tag{17}$$

E.4 Evaluation of Stack

Stack (Dong et al., 2026) is pretrained on 149 million cells and embeds each cell together with a set of context cells. We applied its released pretrained model to $\mathcal{D}_{\mathrm{ct}}$ without further training and formed the contexts in two ways: from cells of the same donor, as in the official tutorial, and from blocks of cells of the same source dataset grouped by their annotated cell type. The second variant yields the most homogeneous contexts, but it relies on the reference annotations and is therefore the setting most favorable to Stack. Table 7 reports the clustering metrics of Appendix E.3 together with the accuracy and macro-F1 of a k-nearest-neighbor (kNN) classifier on the Stack features. With either context, Stack attains a high kNN accuracy of about 0.97 but an ARI below 0.46 and an AvgBio below 0.60, and the label-informed contexts improve its scores only marginally.

Gene vocabulary mismatch. The genes of $\mathcal{D}_{\mathrm{ct}}$ are indexed by the vocabulary of Geneformer, which contains 25,427 gene tokens. Stack, in contrast, accepts a fixed set of 15,012 genes, which determines the input dimension of its first layer, a linear map from 15,012 to 1,600 dimensions stored in the pretrained weights. Of the gene tokens that actually occur in $\mathcal{D}_{\mathrm{ct}}$, only 71.4% belong to this set; the remaining 28.6% were not among the highly variable genes selected for Stack and must be discarded. The resulting loss of input information arises from the incompatible vocabularies rather than from the modeling approach of Stack, so its scores are not directly comparable with those of the other methods. We therefore report Stack here instead of in Table 1.

F Details of the Cell-Type Annotation Experiment

This appendix describes the supervised annotation study of Section 4.3, which asks how accurately a classifier trained on labeled target cells of $\mathcal{D}_{\mathrm{ct}}$ from the frozen features of different methods.

Table 7: Stack on the cross-cell-type dataset $\mathcal{D}_{\mathrm{ct}}$ with two ways of forming the context. The cell-type blocks use the reference annotations.

	Clustering			
Context	NMI	ARI	ASW	**AvgBio**
Same donor (official tutorial)	0.802	0.438	0.515	0.585
Cell-type blocks (label-informed)	0.807	0.454	0.515	0.592

F.1 Compared Features

We compare four feature extractors that cover the method families considered in this work. **scCoLLM** provides the text-embedding features. **UCE** (Rosen et al., 2026), **scCello** (Yuan et al., 2024) and **Seurat** (Stuart et al., 2019; Hao et al., 2024) are included. All features are extracted with frozen models. Each feature vector is normalized to unit ℓ_2 norm,

$$\tilde{\boldsymbol{z}}_i = \frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|_2}, \tag{18}$$

F.2 Data Sampling and Splits

We use all cells of $\mathcal{D}_{\mathrm{ct}}$ and divide them into disjoint training, validation, and test sets in the ratio $7 : 1.5 : 1.5$. The division is constructed so that the training set contains cells of all 87 cell types, which guarantees that every cell type encountered at test time has labeled training examples. The validation set is used only for model selection, and all reported scores are computed on the test set. The procedure is repeated with three random seeds, and we report the mean and standard deviation of each metric.

F.3 Classifier and Training

On top of the frozen normalized features, we train an MLP $h_\psi : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ with parameters ψ, where d is the feature dimensionality of the respective method and $\mathcal{Y}$ is the set of 87 cell types. The classifier defines

$$p_\psi(y \mid \tilde{\boldsymbol{z}}) = \mathrm{softmax}\big(h_\psi(\tilde{\boldsymbol{z}})\big)_y, \tag{19}$$

and ψ is obtained by minimizing the cross-entropy loss over the training set; the feature extractors are not updated. The classifier is a three-layer perceptron with hidden widths 256 and 128, i.e., $d \to 256 \to 128 \to 87$, and both the architecture and the optimization settings are identical for the four feature types, so that the reported differences are attributable to the features alone. The model with the best validation performance is used for testing.

F.4 Evaluation Metrics

Let $\mathcal{D}_{\mathrm{te}}$ denote the test set of n cells with annotations y_i and predictions $\hat{y}_i = \arg\max_{y \in \mathcal{Y}} p_\psi(y \mid \tilde{\boldsymbol{z}}_i)$, let $\mathcal{Y}_{\mathrm{te}} \subseteq \mathcal{Y}$ be the set of cell types present among the test annotations, and let n_y be the number of test cells of type y. Accuracy is the fraction of correctly annotated cells,

$$\mathrm{Acc} = \frac{1}{n} \sum_{i \in \mathcal{D}_{\mathrm{te}}} \mathbf{1}\big[\hat{y}_i = y_i\big]. \tag{20}$$

For a cell type y, the precision and recall are

$$\mathrm{P}_y = \frac{|\{i \in \mathcal{D}_{\mathrm{te}} : \hat{y}_i = y_i = y\}|}{|\{i \in \mathcal{D}_{\mathrm{te}} : \hat{y}_i = y\}|}, \qquad \mathrm{R}_y = \frac{|\{i \in \mathcal{D}_{\mathrm{te}} : \hat{y}_i = y_i = y\}|}{n_y}, \tag{21}$$

and $\mathrm{F1}_y = 2\,\mathrm{P}_y\mathrm{R}_y/(\mathrm{P}_y{+}\mathrm{R}_y)$, where ratios with a zero denominator are set to 0. The class-averaged metrics are

$$\text{Macro-F1} = \frac{1}{|\mathcal{Y}_{\mathrm{te}}|} \sum_{y \in \mathcal{Y}_{\mathrm{te}}} \mathrm{F1}_y, \qquad \mathrm{BAcc} = \frac{1}{|\mathcal{Y}_{\mathrm{te}}|} \sum_{y \in \mathcal{Y}_{\mathrm{te}}} \mathrm{R}_y, \tag{22}$$

$$\text{Weighted-F1} = \sum_{y \in \mathcal{Y}_{\text{te}}} \frac{n_y}{n} \, \text{F1}_y. \tag{23}$$

Following scikit-learn (Pedregosa et al., 2011), which is used for all computations, a cell type that appears only among the predictions additionally enters macro-F1 with $\text{F1}_y = 0$. Macro-F1 and balanced accuracy weight all cell types equally and therefore emphasize rare types, whereas weighted-F1 weights each cell type by its frequency in the test set.

G Details of the Bio-Context Analysis

Conditioning variants. The study in Section 4.4 evaluates five conditioning variants on $\mathcal{D}_{\text{ct}}$. They differ in the context supplied to the frozen LLM, while the query serialization, query-profiling procedure, and text embedding model are kept fixed:

(i) **Bio-Context (scCoLLM):** reference cells are selected from the unlabeled target dataset by the query-conditioned procedure of Section 3.2 and paired with their generated profiles.

(ii) **Query-only:** the query is profiled without reference examples, i.e., $\mathcal{B}_c = \varnothing$.

(iii) **Random context:** K reference cells are drawn uniformly at random from $\mathcal{D}_{\text{ct}}$ and paired with zero-shot generated profiles.

(iv) **Cross-domain context:** the same Bio-Context construction procedure selects references from the cross-tissue dataset $\mathcal{D}_{\text{ts}}$ instead of the target dataset $\mathcal{D}_{\text{ct}}$.

(v) **Relevance-only context:** By using k-NN retrieval to identify the k (k=10) cells most similar to the query cell, this approach focuses solely on relevance while disregarding global diversity. This variant investigates whether simple k-NN retrieval—based exclusively on query relevance—is sufficient to explain the performance gains observed with scCoLLM.

(vi) **Oracle:** the retrieved cells of variant (i) are retained, but their generated profiles are replaced with ground-truth cell-type annotations. This provides a comparison with annotated target-domain references.

Clustering uses the metrics of Appendix E.3. The random and cross-domain clustering scores are means over ten runs; the query-only condition was evaluated once. Table 3 reports the scores and their differences from scCoLLM, with the oracle shown separately from the label-free variants.

Profile-attribute evaluation. In addition to evaluating cell embeddings, we compare each generated query profile with the reference values of its five identity attributes: broad compartment, lineage, sublineage, maturity or developmental stage, and anatomical specialization (Appendix B). Accuracy and macro-F1 are computed for each attribute and then averaged over the five attributes. These metrics are also used in the robustness and backbone studies. Reference annotations are used for evaluation; only the explicitly labeled oracle supplies them to the LLM.

H Details of the Retrieval Analysis

This appendix defines the statistics reported in Section 4.5 and lists their exact values. We evaluate $\lambda \in \{0, 0.1, 0.5, 1.0\}$, with 0.1 as the default. All four settings use the same set $\mathcal{Q}_{\text{ret}}$. Let $S_c^\lambda \subseteq \mathcal{Z}_c$ be the $K = 10$ references retrieved for $c \in \mathcal{Q}_{\text{ret}}$ by greedy MAP inference on Eq. (7), where $\lambda = 0$ denotes the pure-relevance limit. Let y_i be the annotated cell type of cell i. For each identity attribute $f \in \{1, \ldots, 5\}$, i.e., broad compartment, lineage, sublineage, maturity or developmental stage, and anatomical specialization (Section 3.2), let $v_f(c)$ be the ground-truth value of the query and $\hat{v}_f(s)$ the value stated in the generated profile $\hat{\mathcal{T}}_s$ of reference s. Annotations are used only to evaluate retrieval. Local relevance is measured by mean query–reference cosine similarity and the fraction of references sharing the query's cell type:

$$\text{Sim}(\lambda) = \frac{1}{|\mathcal{Q}_{\text{ret}}|} \sum_{c \in \mathcal{Q}_{\text{ret}}} \frac{1}{K} \sum_{s \in S_c^\lambda} \boldsymbol{u}_s^\top \boldsymbol{u}_c, \tag{24}$$

$$\text{Same}(\lambda) = \frac{1}{|\mathcal{Q}_{\text{ret}}|} \sum_{c \in \mathcal{Q}_{\text{ret}}} \frac{1}{K} \sum_{s \in S_c^\lambda} \mathbf{1}\big[y_s = y_c\big], \tag{25}$$

Context richness is measured by the number of distinct reference cell types and answer coverage. The latter is the fraction of queries for which at least one reference profile supplies a matching attribute value, averaged over the five attributes. Every context contains at least one cell type, so a fraction p_1 of single-type contexts implies $\mathrm{Types}(\lambda) \geq p_1 + 2(1 - p_1)$, i.e., $p_1 \geq 2 - \mathrm{Types}(\lambda)$; for $\lambda = 0$, this gives $p_1 \geq 0.798$. Table 8 lists the values plotted in Figure 3.

Table 8: Relevance and richness of the retrieved Bio-Context on $\mathcal{D}_{\mathrm{ct}}$ for different values of λ. The default setting is marked.

	Relevance		Richness	
λ	Cosine similarity	Same-type rate (%)	Distinct cell types	Answer coverage (%)
0	0.920	93.54	1.202	81.50
0.1 (default)	0.853	83.12	2.030	85.01
0.5	0.812	77.81	2.333	85.76
1.0	0.810	77.50	2.342	85.79

I Details of the Noise Robustness Analysis

We conduct the noise-robustness study on $\mathcal{D}_{\mathrm{ct}}$. Let $\mathcal{U} \subseteq \mathcal{D}_{\mathrm{ct}}$ be the set of cells that appear as a reference in the Bio-Context of at least one query. For each corruption rate $\rho \in \{10\%, 30\%, 50\%\}$, we draw a subset $\mathcal{U}_\rho \subseteq \mathcal{U}$ containing a fraction ρ of these cells uniformly at random. For every $s \in \mathcal{U}_\rho$, the generated profile $\hat{\mathcal{T}}_s$ is replaced by a profile whose value for each of the five identity attributes is drawn at random from the values of that attribute. The same replacement profile is reused in every Bio-Context $\mathcal{B}_c$ with $s \in S_c$, so corruption is consistent across queries that share a reference cell.

We include two controls: scCoLLM with its original generated reference profiles and no added noise ($\rho = 0$, labeled *No added noise* in Figure 4), and an oracle that supplies ground-truth reference annotations in the Bio-Context (labeled *Noise-free*). All five conditions use the same retrieval procedure, prompt template, and inference procedure, with GPT-oss-20B as the frozen LLM backbone; only the content of the reference profiles changes. Each condition is evaluated with the clustering protocol of Appendix E.3 and the profile-attribute metrics of Appendix G. Figure 4 reports all five conditions, and Section 4.6 describes the effect of corruption relative to the unperturbed condition.

J Details of the Context-Size Analysis

The context-size study in Section 4.6 varies the number of reference cells over $K \in \{5, 10, 16, 32, 64\}$ on $\mathcal{D}_{\mathrm{ct}}$, using GPT-oss-20B as the frozen backbone. All other components are held fixed within the sweep. We evaluate the resulting cell embeddings with the clustering metrics of Appendix E.3 and the generated profiles with the attribute metrics of Appendix G. Figure 4 reports all five settings; its labels $\mathrm{ASW}_{\mathrm{label}}$ and $\mathrm{avg}_{\mathrm{bio}}$ denote ASW and AvgBio, respectively. The default context size is $K = 10$.

K Details of the Backbone Comparison

Section 4.7 evaluates scCoLLM on $\mathcal{D}_{\mathrm{ct}}$ with four frozen LLMs: the default proprietary GPT-5.6 (OpenAI, 2026), the open-weight GLM-5.3 from Z.ai (Zeng et al., 2026), and OpenAI's GPT-oss-120B and GPT-oss-20B (Agarwal et al., 2025). The comparison exchanges the LLM used to generate reference and query profiles while retaining the retrieval procedure, prompt format, and text embedding model. Each backbone is accessed through its inference interface without single-cell-specific training or parameter updates.

Table 4 reports clustering performance and profile-attribute accuracy and macro-F1, using the metrics in Appendices E.3 and G. Clustering scores without standard deviations are means over ten runs.

For the within-family scale comparison, GPT-oss-20B and GPT-oss-120B have approximately 21B and 117B total parameters, respectively. Inference costs are reported separately in Appendix D.

L Case Study: A Correction Made by the Bio-Context

Section 4.4 establishes in aggregate that the Bio-Context improves the accuracy of the cell identities inferred by the LLM. This appendix follows a single cell of $\mathcal{D}_{\mathrm{ct}}$ through the two conditions of that comparison, the query-only variant (ii) and scCoLLM (i), which differ in nothing but the context inserted into the prompt.

The query. Cell 390757 is annotated as an early double-negative thymocyte. Its serialization opens with the tier of its eight most enriched genes, *CD99*, *IGLL1*, *MACROH2A1*, *NUCB2*, *DUT*, *MZB1*, *PTMA*, and *HINT1*.

Query-only inference. Given this list alone, the frozen LLM places the cell in the immune compartment with lineage `lymphoid_B` and sublineage `pre_B`. The reading is defensible on the evidence available: *IGLL1*, ranked second, encodes a component of the pre-B-cell receptor, and *MZB1*, ranked sixth, is associated with immunoglobulin production in B cells, so two of the strongest signals in the list point to the B lineage and a ranked gene list offers nothing that contradicts them.

Inference with the Bio-Context. The Bio-Context retrieved for this query supplies a second and independent kind of evidence: reference cells of the Bio-Context that lie close to the query in the retrieval space ϕ, each accompanied by the profile the same frozen LLM generated for it. These profiles are not unanimous, but they are skewed to state the lineage `lymphoid_T`, and the sublineage `thymocyte_early_DN`. Conditioned on this context, scCoLLM answers immune, `lymphoid_T`, `thymocyte_early_DN`, which agrees with the reference annotation on all three fields (Table 9).

Table 9: A single cell of $\mathcal{D}_{\mathrm{ct}}$ (index 390757) under the two conditions of Section 4.4.

Field	Query-only	**scCoLLM**	Reference
`broad_compartment`	immune	immune	immune
`lineage`	lymphoid B	**lymphoid T**	lymphoid T
`sublineage`	pre B	**thymocyte early DN**	thymocyte early DN

What the example shows. Because the two conditions share the query, the backbone, the template, and the embedding model, the correction is attributable to the context alone. What the context adds is not a longer or better gene list but evidence of a different kind. In the query-only prompt the model must resolve one transcriptome against priors formed during pretraining, and markers that are ambiguous in isolation are resolved in favor of the identity those priors associate with them most strongly. The Bio-Context instead places beside the query a set of cells from the same target domain, together with how the same model reads them, and the ambiguity is settled by what the neighborhood of the query actually looks like in this dataset rather than by what a marker most often means in general. The two fields that change here, lineage and sublineage, are also the fields at which cells of different identity are separated in the resulting feature space, which is how a correction of this kind reaches the clustering scores of Section 4.2.