Biological Reasoning with Reinforcement Learning through Natural Language Enables Generalizable Zero-Shot Cell Type Annotations

Xi Wang [*] [1] Runzi Tan [*] [1] Bo Wang [2] Simona Cristea [1] [3]

Abstract

Single-cell RNA-sequencing (scRNAseq) has reshaped biomedical research, enabling the high-resolution characterization of cellular populations. Yet cell type annotation, a process typically performed by domain experts interpreting gene expression patterns by manual curation or with specialized algorithms, remains labor-intensive and limited by prior knowledge. In addition, while reasoning large language models (LLMs) have demonstrated remarkable performance on mathematics, coding and general-reasoning benchmarks, their potential in scRNAseq analyses remains underexplored. Here, we investigate the advantages and limitations of employing DeepSeek-R1-0528, a recently developed open-source 671B-parameter reasoning LLM, for zero-shot scRNAseq cell type annotation. We find that DeepSeek-R1 prompted with a ranked list of 10 differentially expressed marker genes per cluster of single cells outperforms both its reasoning-enhanced, non-reasoning equivalent (DeepSeek-V3-0324) and GPT-4o in cluster-level annotations. At the level of single cells, DeepSeek-R1 prompted with the top 500 expressed genes in a cell outperforms its non-reasoning counterpart DeepSeek-V3, illustrating test-time scaling for bioinformatics tasks through natural language. Running DeepSeek-R1 in zero-shot classifier mode, with a prompt that presents a broad catalogue of cell type labels to choose from, improves its performance and generalizability across different datasets. On data curated by the expert model scTab (termed in-domain data), the DeepSeek-R1 classifiers perform better than the expert model scGPT and on par with the specialized cell genomics LLM C2S-Scale-1B, but lag behind scTab. On out-of-distribution data unseen by the two expert models, DeepSeek-R1 and its classifier versions generalize better and outperform the other models in the majority of the evaluated datasets. Notably, DeepSeek-R1 supports its cell type calls with interpretable textual biological rationales underlying its reasoning, providing a learning opportunity for researchers. Nevertheless, peak annotation performance remains modest, highlighting the intrinsic complexity of scRNAseq cell type annotation.

1. Introduction

Single-cell RNA-sequencing (scRNAseq) has transformed modern biology, enabling the study of gene expression in individual cells and revealing previously unrecognized cellular states. From immunology to developmental biology and precision oncology, scRNAseq approaches are now integral to addressing diverse research questions (Papalexi & Satija, 2018; Park et al., 2022; Tang et al., 2009; Patel et al., 2014; Tirosh et al., 2016; Hwang et al., 2018). However, despite immense algorithmic progress within the past 10 years, with hundreds of advanced methods developed by the research community, cell type annotation remains a bottleneck step in scRNAseq bioinformatics pipelines. In practice, biomedical researchers often resort to manually annotating scRNAseq data by interpreting representative marker genes using their domain expertise. But, human experts can also be subject to bias in their annotations due to their specific expertise, experience level, or different perceptions of the granularity required for the task (Hou & Ji, 2024; Ergen et al., 2024; Andreatta et al., 2024).

Alternatively, state-of-the-art (SOTA) supervised methods or foundation models can accelerate cell type annotation. Yet, an inherent limitation of such approaches is that models can only recognize the cell identities present during train-

[*] Equal contribution [1] Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA [2] Department of Computer Science, University of Toronto, Canada; Vector Institute for Artificial Intelligence, Toronto, Canada; Peter Munk Cardiac Center, University Health Network, Toronto, Canada; Department of Laboratory Medicine and Pathobiology, University of Toronto, Canada [3] Department of Biostatistics, Harvard T.H. Chan School of Public Health. Correspondence to: Simona Cristea <scristea@ds.dfci.harvard.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

ing. As a result, novel or *out-of-domain* (OOD) cells are either forced into the closest known class or reported as *unknown*. Therefore, the true extent to which such algorithms generalize to novel datasets and offer practical utility for researchers as an end-to-end solution for cell type annotation remains unclear (Heumos et al., 2023; Luecken & Theis, 2019; Abdelaal et al., 2019; Stuart & Satija, 2019; Fischer et al., 2024; Cui et al., 2024).

LLMs have captured extensive attention for their capacity to reason through complex tasks using chain-of-thought (CoT), performing particularly well in mathematics, coding, or clinical decision-making (DeepSeek-AI et al., 2025; Zhang et al., 2022; Wei et al., 2023; Brown et al., 2020; Vaswani et al., 2017; Radford et al., 2019; OpenAI et al., 2024; Sandmann et al., 2025; Tordjman et al., 2025; Skarlinski et al., 2024). DeepSeek-R1-0528 was recently introduced as an open-source general-purpose 671B-parameter reasoning model, specifically trained to strengthen its reasoning capabilities (DeepSeek-AI et al., 2025). Reasoning models like DeepSeek-R1-0528 (referred to as DeepSeek-R1 or simply R1 from here onwards) can parse new problems at inference time, a concept known as test-time scaling, with little or no additional training (Wei et al., 2023). This approach mirrors how human experts retrieve and synthesize their domain knowledge, suggesting that reasoning LLMs might contribute particularly well to tasks that rely on the dynamic interpretation and aggregation of complex data, such as scRNAseq cell type annotation.

As scRNAseq data presents itself in a quantitative format of gene expression profiles, rather than a textual format, the potential of applying out-of-the-box existing generic LLMs to scRNAseq analysis tasks such as cell type annotation remains underexplored (Hou & Ji, 2024; Lu et al., 2024; Liu et al., 2024; Szałata et al., 2024; Chen & Zou, 2024; Choi et al., 2024; Levine et al., 2024). Attempts at using LLMs for cell type annotation include zero-shot cluster-level labeling with GPT-4 by Hou & Ji (2024), generating embedding representations via natural language prompts with CELLama (Choi et al., 2024) or GenePT (Chen & Zou, 2024), as well as fine-tuning LLMs with scRNAseq data with Cell2Sentence and C2S-Scale (Levine et al., 2024; Rizvi et al., 2025). Specifically, the Hou & Ji (2024) paper showed how an early version of GPT-4 prompted with the top 10 differentially expressed marker genes per cluster can achieve near-expert accuracy and outperform existing SOTA expert algorithms such as singleR (Aran et al., 2019). However, none of the existing works investigate the CoT reasoning capabilities of very recent generic reasoning LLMs such as DeepSeek-R1 for zero-shot scRNAseq cell type annotation, at both the cluster and single-cell levels.

We hypothesized that DeepSeek-R1 is able to deploy its test-time logical reasoning abilities to interpret its pre-training biological knowledge and ultimately reliably annotate scRNAseq data, in a process conceptually similar to manual annotation by domain experts. Here, we tested this hypothesis by zero-shot prompting DeepSeek-R1 with representative marker genes for clusters or single cells, as either a standalone model or as a classifier contextualized with cell type labels to choose from. We focused specifically on assessing zero-shot cell type annotation capabilities, *i.e.* without relying on supervised fine-tuning, as supervised fine-tuning would require access to both bioinformatics expertise and labeled data, introducing practical real-world bottlenecks.

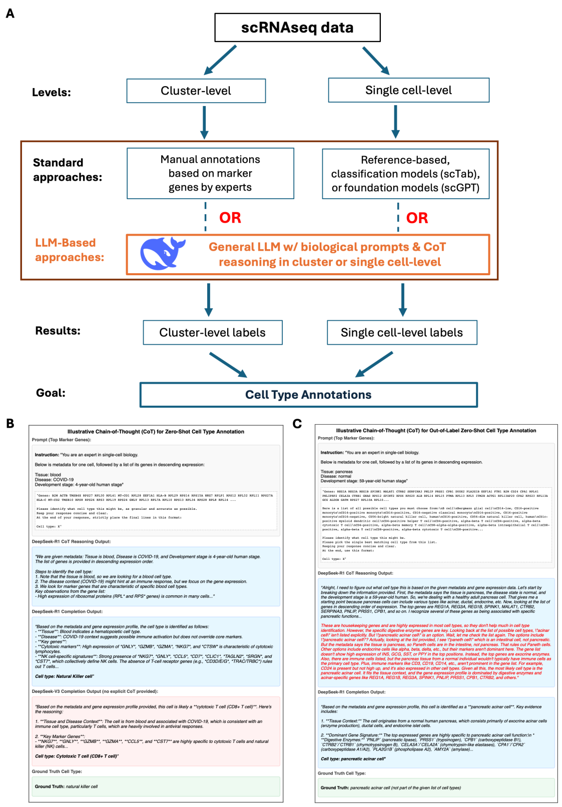
We compared DeepSeek-R1's performance against non-reasoning LLMs, the expert models scTab (Fischer et al., 2024) and scGPT (Cui et al., 2024), as well as the specialized cell genomics pretrained LLM C2S-Scale-1B (Rizvi et al., 2025). We documented the advantages and limitations of annotating scRNAseq data with a general reasoning LLM across various tissues and datasets, and identified generalization challenges faced by the specialized models on unseen data (di Montesano et al., 2025). Due to its reasoning nature, DeepSeek-R1 justified its annotations with biologically interpretable CoT, preserving the interpretability of manual marker-based annotation workflows and providing researchers with the opportunity to understand which genes were relevant for annotation and how they linked to the predicted label. With the amount of scRNAseq data increasing exponentially, reasoning LLMs such as DeepSeek-R1 can be leveraged to re-frame the cell type annotation problem altogether and identify the middle ground between generalizability and contextual expertise needed for annotating cells from novel experimental and biological setups.

2. Results

2.1. Reasoning with large language models for interpretable cell type annotations

Broadly, scRNAseq cell type annotation analyses adopt one of two main strategies: 1) a cluster-level approach, in which a cell type label is given to an entire cluster of cells by manual or reference-driven annotations, and 2) a single-cell level approach, in which each cell's expression profile is mapped directly to a label, often using large classification models, reference-based tools or, more recently, foundation models. Despite the open-source availability of increasingly sophisticated cell type annotation algorithms, researchers still rely on manual inspection of canonical marker genes, regarded in practice by most molecular biologists and medical professionals as the "gold standard" to ensure biological interpretability. However, manual annotation remains a labor- and time-intensive process.

Here, we propose an alternative cell type annotation approach that leverages the general-purpose reasoning LLM


A
scRNAseq data
Levels:
Cluster-level
Single cell-level
Standard approaches:
Manual annotations based on marker genes by experts
Reference-based, classification models (scTab), or foundation models (scGPT)
OR
OR
LLM-Based approaches:
General LLM w/ biological prompts & CoT reasoning in cluster or single cell-level
Results:
Cluster-level labels
Single cell-level labels
Goal:
Cell Type Annotations
B
C

Illustrative Chain-of-Thought (CoT) for Zero-Shot Cell Type Annotation

Prompt (Top Marker Genes):

Instruction: "You are an expert in single-cell biology.

Below is metadata for one cell, followed by a list of its genes in descending expression:

Tissue: blood
Disease: COVID-19
Development stage: 4-year-old human stage"

```
"Genes: B2M ACTB TMSB4X RPS27 RPL30 RPL41 MT-CO1 RPL28 EEF1A1 HLA-B RPL29 RPS14 RPS15A NKG7 RPLP1 RPS12 RPL32 RPL11 RPS27A
HLA-C MT-CO2 TMSB10 RPS8 RPS26 RPS3 RPL19 RPS24 GNLY RPL13 RPL7A RPL10 RPS13 RPL34 RPS28 RPL8 RPL14 ...

Please identify what cell type this might be, as granular and accurate as possible.
Keep your response concise and clear.
At the end of your response, strictly place the final lines in this format:

Cell type: X"
```

DeepSeek-R1 CoT Reasoning Output:

*"We are given metadata: Tissue is blood, Disease is COVID-19, and Development stage is 4-year-old human stage.
The list of genes is provided in descending expression order.*

*Steps to identify the cell type:
1. Note that the tissue is blood, so we are looking for a blood cell type.
2. The disease context (COVID-19) might hint at an immune response, but we focus on the gene expression.
3. We look for marker genes that are characteristic of specific blood cell types.
Key observations from the gene list:
- High expression of ribosomal proteins (RPL* and RPS* genes) is common in many cells..."*

DeepSeek-R1 Completion Output:

*"Based on the metadata and gene expression profile, the cell type is identified as follows:
- **Tissue**: Blood indicates a hematopoietic cell type.
- **Disease**: COVID-19 context suggests possible immune activation but does not override core markers.
- **Key genes**:
- **Cytotoxic markers**: High expression of *GNLY*, *GZMB*, *GZMA*, *NKG7*, and *CTSW* is characteristic of cytotoxic lymphocytes.
- **NK cell-specific signatures**: Strong presence of *NKG7*, *GNLY*, *CCL5*, *CD7*, *CLIC1*, *TAGLN2*, *SRGN*, and *CST7*, which collectively define NK cells. The absence of T-cell receptor genes (e.g., *CD3D/E/G*, *TRAC/TRBC*) rules out T cells...*

Cell type: Natural Killer cell"

DeepSeek-V3 Completion Output (no explicit CoT provided):

*"Based on the metadata and gene expression profile provided, this cell is likely a **cytotoxic T cell (CD8+ T cell)**. Here's the reasoning:*

*1. **Tissue and Disease Context**: The cell is from blood and associated with COVID-19, which is consistent with an immune cell type, particularly T cells, which are heavily involved in antiviral responses.*

*2. **Key Marker Genes**:
- **NKG7**, **GNLY**, **GZMB**, **GZMA**, **CCL5**, and **CST7** are highly specific to cytotoxic T cells and natural killer (NK) cells...*

Cell type: Cytotoxic T cell (CD8+ T cell)"

Ground Truth Cell Type:

Ground Truth: *natural killer cell*

Illustrative Chain-of-Thought (CoT) for Out-of-Label Zero-Shot Cell Type Annotation

Prompt (Top Marker Genes):

Instruction: "You are an expert in single-cell biology.

Below is metadata for one cell, followed by a list of its genes in descending expression.

Tissue: pancreas
Disease: normal
Development stage: 59-year-old human stage"

```
"Genes: REG1A REG3A REG1B SPINK1 MALAT1 CTRB2 SERPINA3 PNLIP PRSS1 CPB1 DUOX2 PLA2G1B EEF1A1 FTH1 B2M CD24 CPA1 RPL41
PNLIPRP2 CELA3A CTRB1 GNAS RPS12 SPINT2 RPS6 RPS20 ALB RPL14 RPL15 PTMA RPL13 RPL5 ITM2B ACTG1 RPL13AP20 CPA2 RPS23 RPL13A
GCG ALDOB GATM RPS27 RPL10A RPL10...

Here is a list of all possible cell types you must choose from:\nB cell\nBergmann glial cell\nCD14-low, CD16-positive
monocyte\nCD14-positive monocyte\nCD14-positive, CD16-negative classical monocyte\nCD14-positive, CD16-positive
monocyte\nCD16-negative, CD56-bright natural killer cell, human\nCD16-positive, CD56-dim natural killer cell, human\nCD1c-
positive myeloid dendritic cell\nCD4-positive helper T cell\nCD4-positive, alpha-beta T cell\nCD4-positive, alpha-beta
cytotoxic T cell\nCD4-positive, alpha-beta memory T cell\nCD8-alpha-alpha-positive, alpha-beta intraepithelial T cell\nCD8-
positive, alpha-beta T cell\nCD8-positive, alpha-beta cytotoxic T cell\nCD8-positive...

Please identify what cell type this might be.
Please pick the single best matching cell type from this list.
Keeping your response concise and clear.
At the end, use this format:

Cell type: X"
```

DeepSeek-R1 CoT Reasoning Output:

"Alright, I need to figure out what cell type this is based on the given metadata and gene expression data. Let's start by breaking down the information provided. First, the metadata says the tissue is pancreas, the disease state is normal, and the development stage is a 59-year-old human. So, we're dealing with a healthy adult pancreas cell. That gives me a starting point because pancreas cells can include various types like acinar, ductal, endocrine, etc. Now, looking at the list of genes in descending order of expression. The top genes are REG1A, REG3A, REG1B, SPINK1, MALAT1, CTRB2, SERPINA3, PNLIP, PRSS1, CPB1, and so on. I recognize several of these genes as being associated with specific pancreatic functions...

These are housekeeping genes and are highly expressed in most cell types, so they don't help much in cell type identification. However, the specific digestive enzyme genes are key. Looking back at the list of possible cell types, \"acinar cell\" isn't listed explicitly. But \"pancreatic acinar cell\" is an option. Wait, let me check the list again. The options include \"pancreatic acinar cell\"? Actually, looking at the list provided, I see \"paneth cell\" which is an intestinal cell, not pancreatic. But the metadata says the tissue is pancreas, so Paneth cells are in the intestine, not pancreas. That rules out Paneth cells. Other options include endocrine cells like alpha, beta, delta, etc., but their markers aren't dominant here. The gene list doesn't show high expression of INS, GCG, SST, or PPY in the top positions. Instead, the top genes are exocrine enzymes. Also, there are immune cells listed, but the pancreas tissue from a normal individual wouldn't typically have immune cells as the primary cell type. Plus, immune markers like CD3, CD19, CD14, etc., aren't prominent in the gene list. For example, CD24 is present but not high up, and it's also expressed in other cell types. Given all this, the most likely cell type is the pancreatic acinar cell. It fits the tissue context, and the gene expression profile is dominated by digestive enzymes and acinar-specific genes like REG1A, REG1B, REG3A, SPINK1, PNLIP, PRSS1, CPB1, CTRB2, and others."

DeepSeek-R1 Completion Output:

*"Based on the metadata and gene expression profile, this cell is identified as a **pancreatic acinar cell**. Key evidence includes:*

*1. **Tissue Context:** The cell originates from a normal human pancreas, which consists primarily of exocrine acinar cells (enzyme production), ductal cells, and endocrine islet cells.*

*2. **Dominant Gene Signature:** The top expressed genes are highly specific to pancreatic acinar cell function:\n * **Digestive Enzymes:** `PNLIP` (pancreatic lipase), `PRSS1` (trypsinogen), `CPB1` (carboxypeptidase B1), `CTRB2`/`CTRB1` (chymotrypsinogen B), `CELA3A`/`CELA2A` (chymotrypsin-like elastases), `CPA1`/`CPA2` (carboxypeptidase A1/A2), `PLA2G1B` (phospholipase A2), `AMY2A` (amylase)...*

Cell type: pancreatic acinar cell"

Ground Truth Cell Type:

Ground Truth: *pancreatic acinar cell (not part of the given list of cell types)*

Figure 1. **Overview of the proposed reasoning LLM-based cell type annotation framework (A) and illustrative CoT biological reasonings (B, C).** **A**: Standard pipelines rely on a combination of manual marker gene inspection, reference-based methods or deep learning classifiers. Our approach employs a general-purpose reasoning LLM (DeepSeek-R1) to process ranked gene signatures for both clusters (top 10 differentially expressed genes) and single-cells (top 500 highest expressed genes). **B**: DeepSeek-R1 (correct prediction) vs. DeepSeek-V3 (incorrect). **C**: DeepSeek-R1 proposes the correct label, despite it being outside the labels that the model is instructed to choose from.

DeepSeek-R1 (DeepSeek-AI et al., 2025) (Fig. 1A). We adopted a prompt engineering technique similar to that of Hou & Ji (2024) and Lu et al. (2024) that exploits the LLM's zero-shot cell type annotation capabilities at inference (Hou & Ji, 2024; Levine et al., 2024). We prompted the LLM with either a cluster's top 10 differentially expressed marker genes, or a single cell's top 500 expressed genes, along with relevant metadata (*e.g.*, species or tissue), and asked it to identify the granular cell type label. DeepSeek-R1 responded with CoT reasoning, culminating in an annotated cell type label (Fig. 1B, C). Furthermore, we investigated the impact of prompt engineering on cell type annotation by comparing "long" prompts against "short" prompts asking for more concise responses.

2.2. LLM reasoning enhances cluster-level cell type annotations

Our cluster-level analysis used the 1, 130 single-cell clusters from the work by Hou & Ji (2024) for benchmarking. This aggregated dataset originates from different studies encompassing multiple tissues and cell types from human and mouse, including lung, skin, blood, prostate, fetal development, and many others (Fig. 2A). For performance evaluation and benchmarking, we adapted scTab's evaluation framework (Fischer et al., 2024) with an additional label-matching step, as done by Hou & Ji (2024). Specifically, for each cluster, given a label predicted by a model and a ground truth label provided by the original dataset, we first matched both labels to their corresponding Cell Ontology (CL) database terms (Côté et al., 2006), to remove potential ambiguities in LLM outputs. Then, using Ubergraph (Balhoff et al., 2022), we compared the ground truth label first with the LLM output matched against CL, then further with all its child descendants in the ontology tree. If a match was recorded between the ground truth label and either the LLM label or the label of any of its descendants, the prediction was recorded as TRUE. Otherwise, the match was unsuccessful, and the prediction was recorded as FALSE. Lastly, we aggregated these TRUE/FALSE predictions across all the 1,130 tested clusters (Methods A.3).

As the Hou & Ji (2024) study showed that a now-outdated version of GPT-4 was superior to SOTA expert models such as SingleR (Aran et al., 2019), scType (Ianevski et al., 2022), and CellMarker 2.0 (Hu et al., 2023) for cluster-level cell type annotation, in our cluster-level analysis we benchmarked DeepSeek-R1 only against the contemporary GPT-4o (version 2025-0326) and DeepSeek-V3-0324. DeepSeek-V3-0324 is an updated version of DeepSeek-V3, the instruction-tuned and Reinforcement Learning with Human Feedback (RLHF)-fine-tuned version of the DeepSeek-V3-base model (unavailable at the time of testing). Notably, DeepSeek-V3-0324 incorporates reasoning enhancements from the first version of DeepSeek-R1 (DeepSeek-R1-0120) via improved post-training and RL insights. However, while both DeepSeek-R1 and DeepSeek-V3-0324 can explain their answers step-by-step, the R1 models were built from the ground up as reasoning models, while V3-0324 learned to reason by distilling DeepSeek-R1-0120's CoT (DeepSeek-AI et al., 2025). Therefore, the R1 models are placed in the Arena's reasoning tier, while V3-0324 is considered reasoning-enhanced but classified as non-reasoning. In contrast, GPT-4o has not been explicitly trained with techniques like long-form CoT distillation or targeted RLHF that reward internal reasoning steps and is considered a full non-reasoning model.

Our evaluations revealed that both DeepSeek models outperformed GPT-4o in annotation accuracy, with the highest

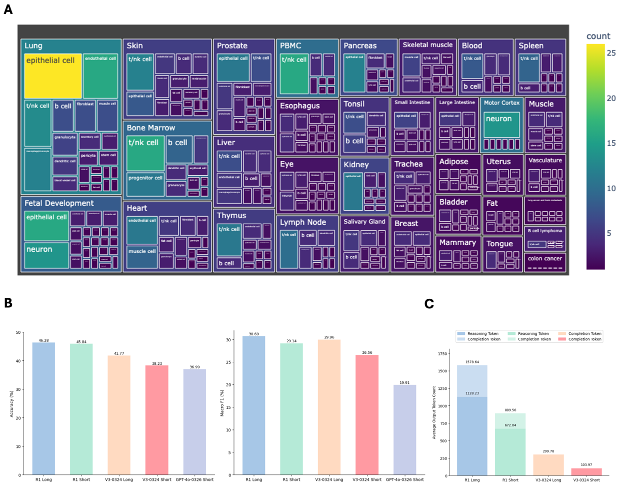


Figure 2. **Benchmarking performance of DeepSeek-R1 on cluster level cell type annotation against DeepSeek V3 and GPT-4o.** **A**: Treemap of the cluster-level dataset composition across cell types and tissues; the color scale and the area of each box indicate the number of clusters per each unique tissue and cell type combination. **B**: Bar plots comparing the accuracy and Macro-F1 score of the tested models on cluster-level cell type annotation across the 1,130 clusters. GPT-4o-0326 was tested using the same short prompts as R1 and V3. **C**: Bar plots showing the average token count for cluster-level cell type annotation, averaged across clusters. For reasoning models, the darker color represents the reasoning token count (unique to reasoning models, representing the computational steps during the internal thought process), and the lighter bar shows the completion token count (representing all the final generated answer text); for non-reasoning models, the bar shows the completion token count.

cluster-level accuracies of 46.28% and 45.84% obtained by DeepSeek-R1 with the long reasoning and short prompt respectively, and the lowest accuracy of 36.99% obtained by GPT-4o (Fig. 2B). DeepSeek-R1's accuracy also consistently surpassed DeepSeek-V3-0324 under the same prompt (46.28% vs 41.77% for long prompts and 45.84% vs 38.23% for short ones), with Macro-F1 evaluations following a similar trend (Fig. 2B; Supplementary Table S1).

We further examined how different prompting strategies affected test-time computation by investigating the average token count per query (Fig. 2C). For DeepSeek-R1, the output includes internal *reasoning tokens* (representing the model's step-by-step thought process before generating the final answer) along with the final output tokens intended for the user. The *completion token* count for R1 is the sum of both reasoning and final output tokens. In contrast, V3-0324 does not have explicit internal reasoning tokens; rather its output consists solely of *completion tokens*. We found that, on average, R1 used far more tokens than V3-0324: 5.27 times more for the long prompt, and 8.56 times more for the short prompt. The absolute 0.44% increase in R1's accuracy when using a long versus a short prompt came at the cost of average reasoning and completion token increases of 1.68 and 1.77-fold respectively, while for V3-0324, an absolute 3.54% accuracy gain came at the cost of 2.88-fold more completion tokens.

Altogether, these results show that: 1) enhanced LLM logical reasoning during inference, particularly longer test-time, translates into improved biological interpretation of cluster-level differential gene signatures compared to lack of reasoning, and 2) the performance benefit of full reasoning (R1) comes at a high token cost compared to learned reasoning capabilities (V3-0324), and similarly for long versus short prompts, in line with recent reports of sub-optimal token consumption of full reasoning models (Nayab et al., 2025; Lee et al., 2025; Fu et al., 2024; Han et al., 2025; Ballon et al., 2025). Lastly, the performance metrics reported here were lower than those of Hou & Ji (2024) for similar models ran on the same cluster-level data because of the different scoring strategy employed by Hou & Ji (2024), in which they also considered partial matches between ground-truth and predicted labels.

2.3. DeepSeek-R1 and its variants overperform expert models on out-of-domain single-cell data and scTab excels on known data

Compared to cluster-level annotation, the annotation of individual cells is more challenging due to higher granularity and lower signal-to-noise ratio. Here, using the same evaluation methodology as for the cluster-level analysis, we benchmarked DeepSeek-R1 first against the non-reasoning model DeepSeek-V3, and further against three specialized models: the two expert models scTab (Fischer et al., 2024), a multi-class classifier, and scGPT (Cui et al., 2024), a scRNAseq foundation model, and the single-cell-genomics-pre-trained LLM C2S-Scale-1B (Rizvi et al., 2025). For benchmarking, we chose DeepSeek-V3 as it is the backbone for both DeepSeek-V3-0324 and DeepSeek-R1 and it does not incorporate any reasoning objective, scTab and scGPT as recent SOTA models for scRNAseq single-cell annotation, and C2S-Scale-1B as a representative tool for LLM scRNAseq cell-type annotation using similar gene-list prompts. C2S-Scale-1B is based on the Pythia-1B architecture and fine-tuned with the Cell2Sentence (Levine et al., 2024) framework on a wide array of scRNAseq datasets.

Our benchmarking utilized four datasets derived from CellXGene: 1) An *in-domain validation dataset*: $10,000$ cells randomly subsampled from the 3.5 million cells used as curated validation data by scTab (Fischer et al., 2024); 2) similar to 1), an *in-domain test dataset* consisting of $10,000$ cells randomly subsampled from scTab's curated test dataset of 3.4 million cells; 3) An *out-of-domain (OOD) random dataset*: $10,000$ cells randomly subsampled from the 7.9

million cells added to CellXGene after May 15th 2023, representing data not seen by neither scTab nor scGPT during training or testing (Program et al., 2024) (Fig. 3A; 4) similar to 3), an *OOD balanced tissue dataset*, with a different subset of randomly sampled $10,000$ cells from the same set of 7.9 million cells, this time covering eight different tissues, with $1,250$ sampled cells per tissue (Fig. 4A; Supplementary Table S2; Methods A.1.2). As the fine-tuning cut-off date for C2S-Scale-1B was later than the release dates of the *OOD* datasets, the *OOD* cells were likely explicitly seen by the model before inference, particularly given C2S-Scale-1B's fine-tuning strategy with cell sentences consisting of lists of marker genes and corresponding ground-truth labels sourced from CellXGene (Program et al., 2024) and the Human Cell Atlas (Regev et al., 2017). In contrast, even though DeepSeek-R1 could have in principle also encountered material related to the *OOD* datasets during training, the DeepSeek pre-training corpus is almost exclusively open-access text and excludes scRNAseq matrices or preprocessed marker-gene tables. Therefore, any potential exposure would have likely been limited to statements linking selected canonical marker genes to cell types, rather than to the verbatim comprehensive gene lists used in our prompts.

Before benchmarking, we first identified the optimal number N of highly expressed genes per cell to include in DeepSeek-R1's prompt, using $1,000$ randomly subsampled cells from the *OOD* random dataset for multiple values of N, ranging from 5 (corresponding to the top 5 most highly expressed genes per cell) to the full set of all genes expressed in a cell (Fig. 3B). We found that providing the top 500 highly expressed genes in the prompt yielded the highest accuracy and the second-highest Macro-F1 score, and we therefore set $N = 500$ as prompt input for subsequent analyses. For the *in-domain* and random *OOD* datasets, we computed confidence intervals (CI) for performance metrics, as well as pairwise comparisons of model performance (Methods A.3.4).

2.3.1. In-domain benchmarking

We first evaluated the performance of the reasoning model DeepSeek-R1 against its non-reasoning counterpart DeepSeek-V3 on the *in-domain validation dataset* using both long and short prompts (Methods A.2). R1 outperformed V3 in both accuracy and Macro-F1 score, while long prompts performed similarly to short prompts, with a significant boost in long prompting for R1 (Fig. 3C, Supplementary Fig. S1). We then benchmarked DeepSeek-R1 with long prompts against scTab, scGPT and C2S-Scale-1B. Recognizing the classifier nature of the expert models scTab and scGPT, we also evaluated R1 in classifier mode by constraining the model to choose a single label from a pre-defined set of labels given in the prompt (Methods A.2). scTab imposed stricter data curation restrictions than scGPT, such as requiring any cell type to have at least $5,000$ unique cells present in at least 30 donors. In consequence, scTab's label set consisted of 164 cell types versus 593 for scGPT.

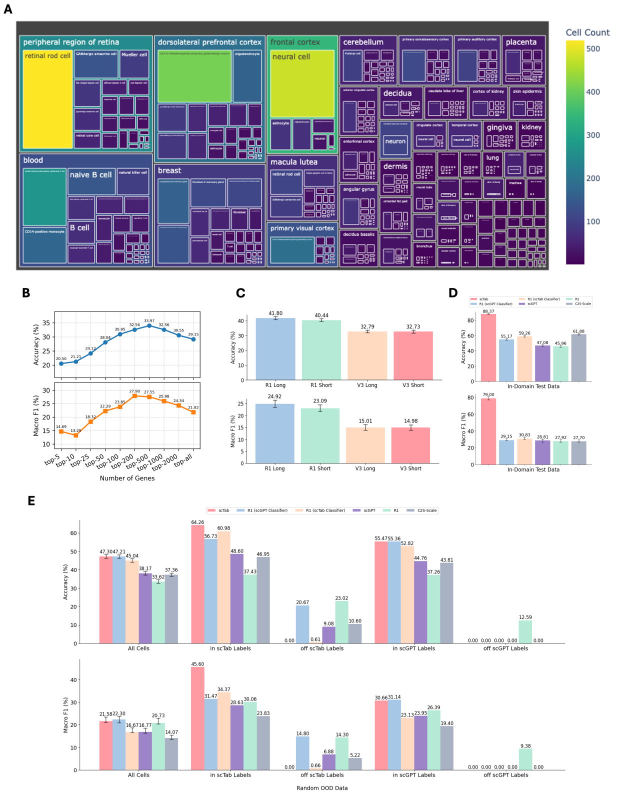


Figure 3. **Benchmarking performance of DeepSeek-R1 on single-cell level cell type annotation against DeepSeek-V3 and specialized cell type annotation models on *in-domain* and *OOD* random datasets.** **A**: Treemap plot showing the tissue and cell type composition of the $10,000$ cells in the *OOD random dataset*, added to the CellXGene database after May 15th, 2023, the cutoff for both scGPT and scTab development. The color scale and the area of each box indicate the number of cells per each unique tissue and cell type combination. **B**: DeepSeek-R1's accuracy and Macro-F1 score as a function of N, the number of top expressed genes included in the prompt, based on a subset of $1,000$ randomly sampled cells from the *OOD* random dataset. **C**: Accuracy and Macro-F1 for DeepSeek-R1, the two DeepSeek-R1 classifier versions with the cell type label set of either scTab or scGPT, the two expert models scTab and scGPT and C2S-Scale-1B on the *in-domain* test dataset consisting of $10,000$ randomly sampled cells. **D**: Accuracy and Macro-F1 score of DeepSeek-R1, the two classifier versions of DeepSeek-R1 using the cell type label set of either scTab or scGPT, the two expert models scTab and scGPT and C2S-Scale-1B on the *in-domain* test dataset consisting of $10,000$ randomly sampled cells. **E**: Performance of the same six models as in (D) on the *OOD* random dataset on all cells, as well as split by whether the ground-truth cell type labels were within scTab's or scGPT's labels.

For *in-domain* testing, we used scTab's test data. As expected, scTab outperformed all models on this curated dataset, followed by the C2S-Scale and DeepSeek-R1 classifiers with scTab and scGPT labels, with the classifiers performing better than C2S-Scale on Macro-F1 scores (Fig. 3D). scGPT and the unconstrained DeepSeek-R1 scored similar accuracy (Supplementary Fig. S1), highlighting the capacity of general reasoning LLMs to annotate single cells at similar levels to SOTA expert models, even in controlled settings where all ground truth cell type labels are within the dictionaries of the expert models.

2.3.2. Out-of-domain (OOD) random sampling benchmarking

When tested on $10,000$ cells randomly downloaded from CellXGene after 15th 2023 and unseen by either of the two expert models, scTab and R1 with scGPT classifier scored statistically indistinguishable highest accuracy (Fig. 3E, upper left panel, labeled *All Cells*; Supplementary Fig. S1). scTab's random *OOD* accuracy dropped by 46% relative to *in-domain*, and C2S-Scale's accuracy dropped by 40%. In contrast, R1's performance was more robust *in-domain* and *OOD*, with relative drops of 14% for R1 with scGPT classifier, 24% for R1 with scTab classifier, and 27% for unconstrained R1. Similarly, scGPT's accuracy only dropped 19% *OOD* relative to *in-domain*. The Macro-F1 score was highest for R1 with scGPT classifier labels, followed by scTab and unconstrained R1 (Fig. 3E, lower panel labeled *All Cells*). The difference in ranking between accuracy and Macro-F1 score reflects different performances on more frequent and less frequent cell types, with models scoring higher on Macro-F1 generally having a more balanced performance across cell types of various frequencies.

The change in the models' performance on *OOD* relative to *in-domain* data reflects variations in batch effects, data distribution and curation, annotation granularities and unseen cell types. Since the *OOD* cells were novel to scTab and scGPT, we investigated how much of the models' decrease in performance was due to a lack of label overlap (*in-label* and *off-label* cells). As expected, the accuracy of scTab rebounded to 64.26% when only assessing on its labels and its Macro-F1 score to 45.60%, closely followed by R1 in classifier mode with scTab labels (60.98%), which nevertheless scored lower Macro-F1 (34.37%, Fig. 3E, panels labeled *in scTab Labels*). When evaluating only on cell types among scGPT's labels, both scTab and R1 classifier with scGPT labels performed best (55.47% and 55.36% accuracy, Fig. 3E, upper panel labeled *in scGPT Labels*). For the subset of cells with labels not among scTab's labels, R1 with scGPT classifier and unconstrained R1 had the highest accuracy and Macro-F1 scores, substantially outperforming scGPT, despite the expert model and the R1 classifier having access to the exact same set of labels (Fig. 3E, panels labeled *off scTab Labels*). The non-zero performance of R1 with scTab classifier was a consequence of the rare situations of R1 proposing a cell type label outside its instruction list from the prompt (3.79% of cases for R1 with scTab labels, and 0.63% for R1 with scGPT labels), and that prediction being correct (Fig. 1C). On cells off scGPT labels, unconstrained R1 was the only model with non-zero performance (Fig. 3E, panels labeled *off scGPT Labels*), underscoring DeepSeek-R1's adaptability for labeling less commonly encountered cell types.

In summary, on the *OOD* random data, R1 in classifier mode with scGPT labels emerged as the most reliable overall cell type annotation strategy, consistently scoring high in most settings. The model also showed comparable accuracy and Macro-F1 scores when tested on the two sets of curated cells (164 cell type labels for scTab and 593 for scGPT) and performed on par with unconstrained R1 on cells off scTab labels. We note that the peak performance of all tested models was modest on the *OOD* random data, reflective of the realistic situation of employing LLMs or expert models for zero-shot annotating scRNAseq data. Even though all cells in this dataset had been QCed and assigned ground truth labels by their respective studies, this data has not undergone additional cell-type-targeted curation, and all cell types were considered before sampling, regardless of their relative frequencies and granularities, in contrast to scTab's *in-domain* data.

2.3.3. OOD balanced tissue benchmarking

To more accurately mimic real-world scenarios that researchers might encounter as daily bioinformatics tasks, we created an *OOD balanced tissue* dataset, in which we sampled $1,250$ cells from eight selected tissues (blood, peripheral region of retina, kidney, breast, lung, trachea, pancreas and cerebellum), amounting to $10,000$ cells (Fig. 4A; Methods A.1.2; Supplementary Table S2). On this aggregated data, the R1 classifiers were the most accurate models across all meaningful settings (Fig. 4B, upper panels). R1 with scGPT labels performed best when tested across all cells, on cells off scTab labels and on cells within scGPT labels, while R1 with scTab labels scored highest when restricting the evaluation only to cells within scTab labels. The unconstrained R1 model emerged as top performing by Macro-F1 scoring in all scenarios except when restricting to scTab labels, further demonstrating how DeepSeek-R1 can generalize and reliably zero-shot annotate a wide variety of cell types of various frequencies and with widely different phenotypes.

To better understand why the *OOD* balanced-tissue benchmark was more challenging for the specialized models, especially scTab, we investigated the label overlap between the labels in these datasets and scTab's label set (Supple-

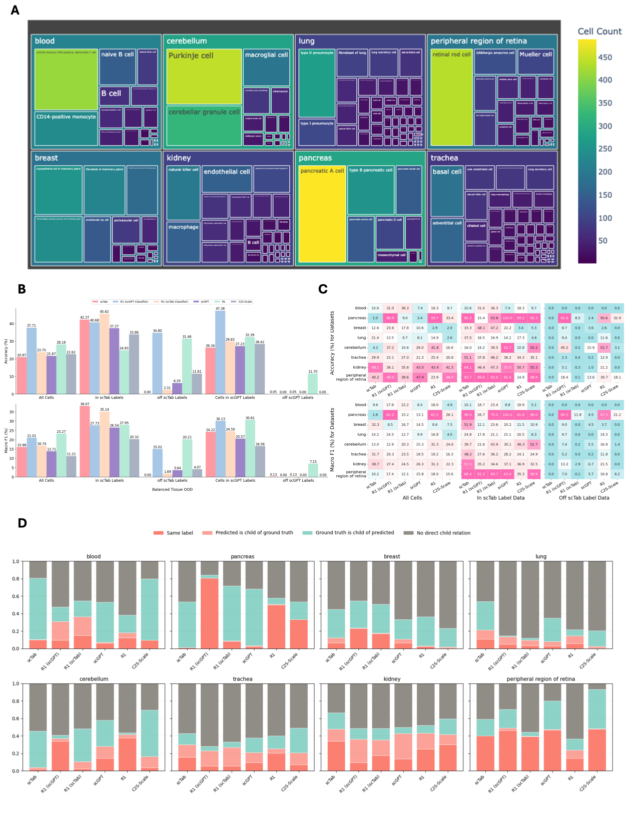


Figure 4. **Benchmarking performance of DeepSeek-R1 on single-cell level cell type annotation against specialized cell type annotation models on the *OOD* balanced tissue dataset. A**: Treemap plot showing the tissue and cell type composition of the 10, 000 cells belonging to eight selected tissues (1, 250 cells per tissue) in the *OOD* balanced tissue dataset used for benchmarking, added to the CellXGene database after May 15th, 2023, the cutoff date for both scGPT and scTab development. The color scale and the area of each box indicate the number of cells per each unique tissue and cell type combination. **B**: Performance of the same six models as in Fig. 3D and E, with the same figure layout. **C**: Accuracy and Macro-F1 score of the same six models as in (B), split by the eight selected tissues, as well as by whether the ground-truth cell type labels were part of scTab's labels. **D**: Stacked barplots showing the distribution of cell-level annotations across all six models and all eight tissues, split by whether the predicted and the ground truth label either matched perfectly or predicted was a child of ground truth (recorded as correct predictions), or ground truth was a child of predicted (incorrect prediction), or there was no direct child-parent relation between predicted and ground truth (also incorrect). For the lung tissue dataset, 382 of the 1, 250 cells had no match of the ground truth label in Cell Ontology and were therefore excluded from accuracy (B and C) and frequency (D) evaluations.

mentary Fig. S2). While scTab covered some lineages well, with 0.08% of blood cells, 24.96% of kidney cells, 41.44% of trachea cells, and 42.86% of lung cells falling outside its vocabulary, it missed a larger fraction of the remaining four tissues, with 57.04% of retina, 62.24% of breast, 74.08% of cerebellum and virtually the entire pancreas dataset (98.48%) falling outside of scTab's 164 labels. In total, 50.44% of cells with a ground truth Cell Ontology match (Supplementary Fig. S3) were off label for scTab, despite all eight tissues included in scTab's classification. This discrepancy might have happened either because the newly added datasets included novel cell types that scTab's datasets did not include, or because these cell types were not large or frequent enough to pass scTab's strict curation criteria.

Further detailed breakdown of tissue-specific performance showed interesting patterns for the different models (Fig. 4C). For blood cells, the two DeepSeek-R1 classifiers yielded the best performance in both accuracy and Macro-F1 score, while scTab and scGPT performed much more poorly (Fig. 4C, panels labeled *All Cells*, first row), despite almost complete cell type label overlap (99.92%) with scTab and blood being the most frequent tissue in scTab's training set. A closer inspection of the cell-level predictions (Fig. 4D) showed that the specialized models, especially scTab and C2S-Scale-1B, often missed the right granularity to match the ground-truth data (*e.g.* predicting *leukocyte* for a ground-truth label of *naïve B cell*). This suggests generalization challenges for the specialized models on novel datasets, despite having access to the right labels and having encountered a large number of similar cells during training. In contrast, when given specific indications to choose from either scGPT's or scTab's set of labels, DeepSeek-R1 in classifier mode correctly labeled a much larger fraction of cells, with even greater granularity than the ground truth data, providing additional biological information.

For pancreas, scTab's training corpus only listed six nonspecific and infrequent cell types: B cell, T cell, endothelial cell, mast cell, mature NK T cell, and plasma cell (Supplementary Table S3). In contrast, the pancreas *OOD* dataset analyzed in this study (Muraro et al., 2016) consisted of ten highly specific pancreatic cell types: pancreatic A cell, pancreatic D cell, pancreatic ductal cell, pancreatic PP cell, type B pancreatic cell, endothelial cell, mesenchymal cell, pancreatic acinar cell, pancreatic endocrine cell, and pancreatic epsilon cell. Endothelial cells, representing less than 5% of all pancreatic cells, were the only common cell type between the two datasets, while acinar cells represent 80−85% of the entire pancreatic tissue mass. Due to very low cell type label overlap (1.52%), scTab and the DeepSeek-R1 scTab classifier performed poorly on this dataset (Fig. 4C). In contrast, while scGPT's performance was also poor (3.4%), contextualizing DeepSeek-R1 with scGPT's labels increased its accuracy to 80.9% (compared to 50.7% for unconstrained R1), further suggesting generalization struggles of expert models. Closer inspection showed that the predictions of the two expert models scTab and scGPT were too general, borrowing similar cell types from other tissues (Fig. 4D).

On the breast tissue, DeepSeek-R1 in scGPT classifier mode showed the highest accuracy, while scTab had the highest Macro-F1 score (Fig. 4C). The performance of unconstrained R1 was particularly poor in accuracy (2.9%), due to granularity issues such as predicting *myoepithelial cell* instead of a ground truth of *basal myoepithelial cell* or mistaking the two lineages basal and luminal. The lung data showed highest accuracy for scTab and highest Macro-F1 score for unconstrained R1, likely a consequence of high cell type label overlap with scTab (Supplementary Fig. S2), as well as many lung cells used in scTab's training. For cerebellum, the unconstrained R1 scored highest in both accuracy and Macro-F1 score, with the overwhelming majority of their correct predictions having the exact same label as the ground truth data, and very few incorrect predictions of lower granularity (Fig. 4D). For trachea and kidney, scTab yielded the highest accuracy and Macro-F1 score (Fig. 4C), likely also a consequence of high cell type label overlap (Supplementary Fig. S2). Lastly, results on the peripheral region of retina tissue showed highest accuracy of R1 in scGPT classifier mode (Fig. 4C) and high percentages of perfect matches for all models except unconstrained R1 (Fig. 4D), even though unconstrained R1 was the model with the highest Macro-F1 score.

Taken together, these analyses on the *OOD balanced tissue* dataset revealed that balancing the tissue distribution and assessing cell type annotation performance separately for each tissue is essential for understanding model performance. Overall, the DeepSeek-R1 model demonstrated superior adaptability, especially in classifier mode when contextualized with appropriate label constraints from the large scGPT label vocabulary.

3. Conclusion and Discussion

Formulating the cell type annotation problem as a benchmarking task is inherently difficult. A necessary ingredient for successful cell typing is accurately labeling biological knowledge as either known or novel. In practice, some cell types are frequent, generic and display a distinctive marker profile, making them relatively straightforward to annotate regardless of the strategy employed (manually investigating a list of marker genes, querying an LLM, or running a specialized cell type annotation model). On the contrary, other cell types are infrequent across donors, their functionality is ambiguous, and their gene expression profiles are non-specific. Annotating such cells turns out to be difficult regardless of the approach used. Moreover, researchers often disagree on the most appropriate ground truth annotation label, or even whether such groups of cells indeed represent a novel cell type or are better characterized as an alternative cellular state of an already-existing cell type (Domcke & Shendure, 2023). Nevertheless, rare and ambiguous cell types are crucial for novel biological discoveries (Alečković et al., 2022). Such aspects turn cell type annotation into a complex problem for which even evaluating the quality of a given solution, is challenging and subject of on-going debate (Domcke & Shendure, 2023). Therefore, the ideal real-world cell type annotation procedure should not only be accurate, but also versatile and adaptable to different scenarios, some of which are hard to capture *a priori* as a set of pre-defined rules.

Building on this rationale, we hypothesized that recently developed general-purpose reasoning LLMs could have the potential to positively contribute to the scRNAseq cell type annotation toolbox, by striking an interesting balance between accuracy and discovery. To this end, we examined the feasibility of employing DeepSeek-R1-0528 to perform both cluster-level and single-cell level annotations in scRNAseq data in a zero-shot setting, without specialized fine-tuning. Running the LLMs zero-shot is key, as fine-tuning requires both existing labeled data and expert bioinformatics expertise, which can be a bottleneck in real-world situations.

By prompting DeepSeek-R1 with a list of ranked marker genes, we assessed its capacity to identify cell types through interpretable CoT reasoning that captures canonical markers, biological functions, and tissue-specific knowledge. We found that LLM reasoning enhanced cell type label prediction at both cluster and single-cell levels. In single-cell datasets, our results revealed that running DeepSeek-R1 in classifier mode, with its prompt contextualizing a large set of cell type labels to choose from, improved its performance, leading to overall superior adaptability and generalizability across tissues and datasets. When comparing DeepSeek-R1 and its classifier variants with the three specialized models scTab, scGPT and C2S-Scale-1B on the curated scTab *in-domain* data, general LLMs performed better than scGPT, while lagging behind scTab and C2S-Scale-1B. However, on random *OOD* data unseen by the expert models, the DeepSeek-R1 scGPT classifier performed on par with scTab, outperforming the other models. The expert models faced generalization challenges on unseen data (di Montesano et al., 2025) and were outperformed by DeepSeek-R1 and its classifier versions on a separate *OOD* dataset consisting of cells from eight relatively common tissues. Notably, half of the cells in this *OOD* balanced tissue dataset fell outside scTab's 164-label vocabulary, highlighting how strict curation criteria can inadvertently exclude biologically relevant but less frequent cell types, leading to decreased generalization capabilities.

Our study demonstrated that annotating scRNAseq data with general reasoning LLMs is reliable and interpretable. Employing such models for scRNAseq cell type annotation will allow the single-cell community to directly benefit from the AI foundation models and biological reasoning innovation wave (Fallahpour et al., 2025), with new SOTA general LLMs released very frequently. In addition, DeepSeek-R1's interpretability and the biological context of its predictions can provide a transparent rationale to cell type annotation similar to manual marker-based annotations. In some situations, DeepSeek-R1 outputted correct predictions of higher granularity than ground-truth labels, reframing the cell type annotation process as a learning opportunity for researchers. Its adaptability counteracted the traditional limitation of tissue-specific training of expert models, allowing on-the-fly generalization to real-world situations and potentially revealing novel cellular biology.

At the same time, our study demonstrated that scRNAseq cell type annotation remains a challenging problem to evaluate and solve. Peak cluster-level and single-cell-level performance was overall modest for any dataset which was not heavily curated by removing infrequent cell types. This happened because real-world randomly-drawn unseen data remains difficult to annotate at the level of granularity proposed by the original study. Some proposed cell type labels are by design very specific, with potentially highly similar expression profiles to other cell types, making accurate differentiation challenging. Additionally, the evaluation criterion employed here was stringent: we required the predicted label to be at least as granular as the ground-truth label, leading to lower performance measures than reported elsewhere.

As recent AI agentic workflows that use LLMs for cell type annotation have been shown to perform better than the models alone (Mao et al., 2025; Gao et al., 2024), we hypothesize that the entire scRNAseq cell type annotation process can be automated with a suite of specialized agents orchestrated by an LLM reasoning "brain" (Fig. 5). For example, the Data Ingestion Agent can take raw scRNAseq data and run basic quality checks. Next, the Clustering Agent can sort cells into subclusters, while an Annotation Agent can assign preliminary labels at either the cluster level or individual cell level. After that, a Quality Control (QC) Agent, backed by an Ontology Agent, a Label Verifier Agent, and a Reasoning Verifier Agent, can examine the given labels against known marker references and ontology databases. At any point, the orchestrator brain can call for deeper checks or bring in human expertise. Finally, an Output Agent can package the refined annotations, key performance metrics, and a summary of how the decisions were made for researchers to evaluate. Because DeepSeek-R1 is flexible and requires no domain-specific training, it could serve as the central orchestrator brain, iteratively integrating and verifying knowledge from the individual sub-agents with minimal human intervention. This multi-agent system is both interpretable and adaptable: the actions and rationale of each sub-agent remain clearly documented, preserving the transparency associated with manual marker gene review, and the orchestrator's reasoning capabilities allow it to readily handle new *OOD* cell types. We anticipate that such automations could substantially accelerate the rapidly evolving landscape of single-cell research, boost efficiency, and allow scientists to focus on deeper biological questions.

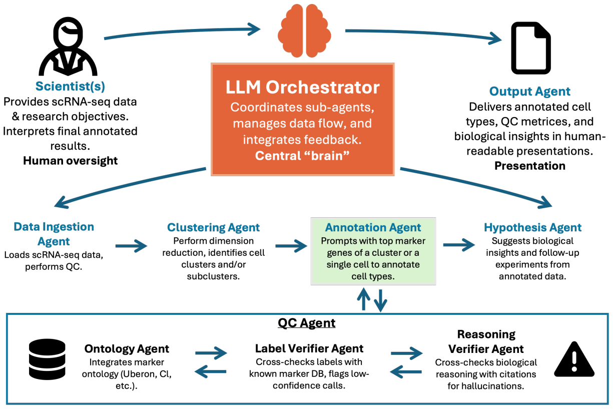

Scientist(s)
Provides scRNA-seq data & research objectives. Interprets final annotated results.
Human oversight
LLM Orchestrator
Coordinates sub-agents, manages data flow, and integrates feedback.
Central "brain"
Output Agent
Delivers annotated cell types, QC metrices, and biological insights in human-readable presentations.
Presentation
Data Ingestion Agent
Loads scRNA-seq data, performs QC.
Clustering Agent
Perform dimension reduction, identifies cell clusters and/or subclusters.
Annotation Agent
Prompts with top marker genes of a cluster or a single cell to annotate cell types.
Hypothesis Agent
Suggests biological insights and follow-up experiments from annotated data.
QC Agent
Ontology Agent
Integrates marker ontology (Uberon, Cl, etc.).
Label Verifier Agent
Cross-checks labels with known marker DB, flags low-confidence calls.
Reasoning Verifier Agent
Cross-checks biological reasoning with citations for hallucinations.

Figure 5. **Proposed multi-agent LLM-based workflow for scRNAseq cell type annotation.** The LLM Orchestrator *brain* coordinates specialized sub-agents, each handling a distinct step of the scRNAseq bioinformatics pipeline from data ingestion and clustering to annotation and quality control. The Data Ingestion Agent loads raw scRNAseq data and performs initial checks; the Clustering Agent partitions cells into groups; and the Annotation Agent uses marker-gene prompts to assign preliminary labels. A QC layer, supported by the Ontology Agent and various verifier agents, cross-references known marker databases to flag potential hallucinations. Finally, the Output Agent consolidates refined annotations, metrics, and summary reports for researchers' review. This orchestrated setup ensures transparency, modularity, and flexibility in automating the cell type annotation process.

References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., and Mahfouz, A. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome Biology*, 20(1):194, 2019. doi: 10.1186/s13059-019-1795-z.

Alečković, M., Cristea, S., Gil Del Alcazar, C. R., Yan, P., Ding, L., Krop, E. D., Harper, N. W., Rojas Jimenez, E., Lu, D., Gulvady, A. C., Foidart, P., Seehawer, M., Diciaccio, B., Murphy, K. C., Pyrdol, J., Anand, J., Garza, K., Wucherpfennig, K. W., Tamimi, R. M., Michor, F., and Polyak, K. Breast cancer prevention by short-term inhibition of tgfβ signaling. *Nature Communications*, 13 (1):7558, 2022. doi: 10.1038/s41467-022-35043-5.

Andreatta, M., Hérault, L., Gueguen, P., Gfeller, D., Berenstein, A. J., and Carmona, S. J. Semi-supervised integration of single-cell transcriptomics data. *Na-*

ture Communications, 15(1):872, 2024. doi: 10.1038/s41467-024-45240-z.

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J., and Bhattacharya, M. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163–172, 2019. doi: 10.1038/s41590-018-0276-y.

Balhoff, J. P., Bayindir, U., Caron, A. R., Matentzoglu, N., Osumi-Sutherland, D., and Mungall, C. J. Ubergraph: Integrating obo ontologies into a unified semantic graph. In *Proceedings of the International Conference on Biomedical Ontology (ICBO 2022)*, pp. 1–9, 2022. doi: 10.5281/zenodo.7249759.

Ballon, M., Algaba, A., and Ginis, V. The relationship between reasoning and performance in large language models – o3 (mini) thinks harder, not longer. *arXiv preprint arXiv:2502.15631*, 2025.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., , et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Chen, Y. T. and Zou, J. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature Biomedical Engineering*, pp. 1–11, 2024. doi: 10.1038/s41551-024-01284-6.

Choi, H., Park, J., Kim, S., Kim, J., Lee, D., Bae, S., Shin, H., and Lee, D. Cellama: Foundation model for single-cell and spatial transcriptomics by cell embedding leveraging language model abilities. *bioRxiv preprint 2024.05.08.593094*, 2024. doi: 10.1101/2024.05.08.593094.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: Toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Côté, R. G., Jones, P., Apweiler, R., and Hermjakob, H. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7(1):97, 2006. doi: 10.1186/1471-2105-7-97.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., , et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

di Montesano, S. C., D'Ascenzo, D., Raghavan, S., Amini, A. P., Winter, P. S., and Crawford, L. Hierarchical cross-entropy loss improves atlas-scale single-cell annotation models. *bioRxiv preprint 2025.04.23.650210*, 2025. doi: 10.1101/2025.04.23.650210.

Domcke, S. and Shendure, J. A reference cell tree will serve science better than a reference cell atlas. *Cell*, 186(6):1103–1114, 2023. doi: 10.1016/j.cell.2023.02.016.

Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Ergen, C., Xing, G., Xu, C., Kim, M., Jayasuriya, M., McGeever, E., Oliveira Pisco, A., Streets, A., and Yosef, N. Consensus prediction of cell type labels in single-cell data with popv. *Nature Genetics*, 56(12):2731–2738, 2024. doi: 10.1038/s41588-024-01993-3.

Fallahpour, A., Magnuson, A., Gupta, P., Ma, S., Naimer, J., Shah, A., Duan, H., Ibrahim, O., Goodarzi, H., Maddison, C. J., and Wang, B. Bioreason: Incentivizing multimodal biological reasoning within a {DNA-LLM} model. *arXiv Preprint*, 2025. URL https://arxiv.org/abs/2505.23579.

Fischer, F., Fischer, D. S., Mukhin, R., Isaev, A., Biederstedt, E., Villani, A.-C., and Theis, F. J. sctab: Scaling cross-tissue single-cell annotation models. *Nature Communications*, 15(1):6611, 2024. doi: 10.1038/s41467-024-51059-5.

Fu, Y., Chen, J., Zhu, S., Fu, Z., Dai, Z., Qiao, A., and Zhang, H. Efficiently serving llm reasoning programs with certaindex. *arXiv preprint arXiv:2412.20993*, 2024.

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., Ektefaie, Y., Kondic, J., and Zitnik, M. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024. doi: 10.1016/j.cell.2024.09.022.

Han, T., Wang, Z., Fang, C., Zhao, S., Ma, S., and Chen, Z. Token-budget-aware llm reasoning. *arXiv Preprint*, 2025. URL https://arxiv.org/abs/2412.18547v4. First posted 2024-12-30; cited version 4, revised 2025-02-17.

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., , et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023. doi: 10.1038/s41576-023-00586-w.

Hou, W. and Ji, Z. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, 21(8):1462–1465, 2024. doi: 10.1038/s41592-024-02235-4.

Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., Li, X., Wang, P., and Zhang, Y. Cellmarker 2.0: An updated database of manually curated cell markers in human/mouse and web tools based on {scRNA-seq} data. *Nucleic Acids Research*, 51(D1): D870–D876, 2023. doi: 10.1093/nar/gkac947.

Hwang, B., Lee, J. H., and Bang, D. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, 50(8):1–14, 2018. doi: 10.1038/s12276-018-0071-8.

Ianevski, A., Giri, A. K., and Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nature Communications*, 13(1):1246, 2022. doi: 10.1038/s41467-022-28803-w.

Lee, A., Che, E., and Peng, T. How well do llms compress their own chain-of-thought? a token complexity approach. *arXiv preprint arXiv:2503.01141*, 2025.

Levine, D., Lévy, S., Rizvi, S. A., Pallikkavaliyaveetil, N., Chen, X., Zhang, D., Ghadermarzi, S., Wu, R., Zheng, Z., Vrkic, I., Zhong, A., Raskin, D., Han, I., Fonseca, A. H. d. O., Caro, J. O., Karbasi, A., Dhodapkar, R. M., van Dijk, D., , et al. Cell2sentence: Teaching large language models the language of biology. *bioRxiv preprint 2023.09.11.557287*, 2024. doi: 10.1101/2023.09.11.557287.

Liu, T., Chen, T., Zheng, W., Luo, X., and Zhao, H. scelmo: Embeddings from language models are good learners for single-cell data analysis. *bioRxiv preprint 2023.12.07.569910*, 2024. doi: 10.1101/2023.12.07.569910.

Lu, Y.-C., Varghese, A., Nahar, R., Chen, H., Shao, K., Bao, X., and Li, C. scchat: A large language model–powered co-pilot for contextualized single-cell rna sequencing analysis. *bioRxiv preprint 2024.10.01.616063*, 2024. doi: 10.1101/2024.10.01.616063.

Luecken, M. D. and Theis, F. J. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019. doi: 10.15252/msb.20188746.

Mao, Y., Mi, Y., Liu, P., Zhang, M., Liu, H., and Gao, Y. scagent: Universal single-cell annotation via a llm agent. *arXiv preprint arXiv:2504.04698*, 2025.

McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M. A., Carlotti, F., de Koning, E. J. P., and van Oudenaarden, A. A single-cell transcriptome atlas of the human pancreas. *Cell Systems*, 3(4):385–394.e3, 2016. doi: 10.1016/j.cels.2016.09.002.

Nayab, S., Rossolini, G., Simoni, M., Saracino, A., Buttazzo, G., Manes, N., and Giacomelli, F. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825*, 2025.

OpenAI, Achiam, J., Adler, S., , et al. Gpt-4 technical report. Technical report, OpenAI, 2024.

Papalexi, E. and Satija, R. Single-cell rna sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35–45, 2018. doi: 10.1038/nri.2017.76.

Park, J. H., Feroze, A. H., Emerson, S. N., Mihalas, A. B., Keene, C. D., Cimino, P. J., Lopez Garcia de Lomana, A., Kannan, K., Wu, W., Turkarslan, S., Baliga, N. S., and Patel, A. P. A single-cell based precision medicine approach using glioblastoma patient-specific models. *NPJ Precision Oncology*, 6(1):1–13, 2022. doi: 10.1038/s41698-022-00294-4.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., and Bernstein, B. E. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014. doi: 10.1126/science.1254257.

Program, C. C. S., Abdulla, S., Aevermann, B., Assis, P., Badajoz, S., Bell, S. M., Bezzi, E., Cakir, B., Chaffer, J., Chambers, S., Cherry, J. M., Chi, T., Chien, J., Dorman, L., Garcia-Nieto, P., Gloria, N., Hastie, M., Hegeman, D., Hilton, J., Huang, T., Infeld, A., Istrate, A.-M., Jelic, I., Katsuya, K., Kim, Y. J., Liang, K., Lin, M., Lombardo, M., Marshall, B., Martin, B., McDade, F., Megill, C., Patel, N., Predeus, A., Raymor, B., Robatmili, B., Rogers, D., Rutherford, E., Sadgat, D., Shin, A., Small, C., Smith, T., Sridharan, P., Tarashansky, A., Tavares, N., Thomas, H., Tolopko, A., Urisko, M., Yan, J., Yeretssian, G., Zamanian, J., Mani, A., Cool, J., and Carr, A. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 11 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1142.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_ models_are_unsupervised_multitask_learners.pdf. Technical report.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., et al. The human cell atlas. *eLife*, 6: e27041, 2017. doi: 10.7554/eLife.27041.

Rizvi, S. A., Levine, D., Patel, A., Zhang, S., Wang, E., He, S., Zhang, D., Tang, C., Lyu, Z., Darji, R., Li, C., Sun, E., Jeong, D., Zhao, L., Kwan, J., Braun, D., Hafler, B., Ishizuka, J., Dhodapkar, R. M., Chung, H., , et al. Scaling large language models for next-generation single-cell analysis. *bioRxiv preprint 2025.04.14.648850*, 2025. doi: 10.1101/2025.04.14.648850.

Sandmann, S., Hegselmann, S., Fujarski, M., Bickmann, L., Wild, B., Eils, R., and Varghese, J. Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, 31:1–1, 2025. doi: 10.1038/s41591-025-03727-2.

Seabold, S. and Perktold, J. {Statsmodels}: Econometric and statistical modeling with {Python}. In *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, pp. 92–96, 2010. doi: 10.25080/Majora-92bf1922-011.

Skarlinski, M. D., Cox, S., Laurent, J. M., Braza, J. D., Hinks, M., Hammerling, M. J., Ponnapati, M., Rodriques, S. G., and White, A. D. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Stuart, T. and Satija, R. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, 2019. doi: 10.1038/s41576-019-0093-7.

Szałata, A., Hrovatin, K., Becker, S., Tejada-Lapuerta, A., Cui, H., Wang, B., and Theis, F. J. Transformers in single-cell omics: a review and new perspectives. *Nature Methods*, 21(8):1430–1443, 2024. doi: 10.1038/s41592-024-02353-z.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009. doi: 10.1038/nmeth.1315.

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth II, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G. K., Kazer, S. W., Gaillard, A., Kolb, K. E., Villani, A.-C., Johannessen, C. M., Andreev, A. Y., Van Allen, E. M., Bertagnolli, M., Sorger, P. K., Sullivan, R. J., Flaherty, K. T., Frederick, D. T., Jané-Valbuena, J., Yoon, C. H., Rozenblatt-Rosen, O., Shalek, A. K., Regev, A., and Garraway, L. A. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016. doi: 10.1126/science.aad0501.

Tordjman, M., Liu, Z., Yuce, M., Fauveau, V., Mei, Y., Hadjadj, J., Bolger, I., Almansour, H., Horst, C., Parihar, A. S., Geahchan, A., Meribout, A., Yatim, N., Ng, N., Robson, P., Zhou, A., Lewis, S., Huang, M., Deyer, T., , et al. Comparative benchmarking of the deepseek large language model on medical tasks and clinical reasoning. *Nature Medicine*, 31:1–1, 2025. doi: 10.1038/s41591-025-03726-3.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 5998–6008, 2017. doi: 10.48550/arXiv.1706.03762.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2023.

Wolf, F. A., Angerer, P., and Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

Zhang, Z., Zhang, A., Li, M., and Smola, A. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

A. Methods

A.1. Benchmarking datasets

A.1.1. Cluster-level data

For the cluster-level analysis, we utilized the data curated by Hou & Ji (2024), a study which assessed GPT-4's ability to annotate single-cell clusters based on marker-gene information. The data had been derived from a comprehensive list of human and mouse scRNAseq datasets spanning a wide range of tissues (Fig. 2A). We downloaded the annotated data containing human and mouse clusters (excluding non-model mammals) through the GitHub repo associated with the paper: `https://github.com/Winnie09/GPTCelltype_Paper/blob/master/anno/compiled/all.csv`. In the study, each cluster was assigned a set of 10 marker genes identified by a Wilcoxon-based differential-expression analysis. We used these top 10 marker genes per cluster to prompt the LLMs in a zero-shot manner. If metadata (*e.g.*, tissue name and/or disease) was available, we appended it to the prompt, as described below.

A.1.2. Single-cell level data

In-Domain scTab datasets We obtained the dataset from the scTab publication (Fischer et al., 2024), encompassing both validation and test pools of curated single cells. Specifically, we downloaded the data through the training-data checkpoint from `https://github.com/theislab/scTab/tree/devel` using this specific link: `https://pklab.med.harvard.edu/felix/data/merlin_cxg_2023_05_15_sf-log1p.tar.gz`. To limit computational and API overhead, we randomly subsampled $10,000$ cells from both the validation and the test sets for our benchmarking. We call these two pools of $10,000$ cells *in-domain*.

Out-Of-Domain (OOD) random dataset To evaluate generalizability beyond the training distribution of scTab, we curated a subset of scRNAseq data from CellxGene (Program et al., 2024) with the following characteristics: human, primary (the study that originally generated the dataset), non-diseased, and added to the database after May 15th 2023 – the training cutoff date for both scTab and scGPT development. We call such data *Out-Of-Domain (OOD)*. A custom Python script first built an index of new unique cell-specific identifiers absent from the 2023-05-15 release and present only in the 2024-07-01 release, then downloaded those cells in chunked form, storing the data in partitioned `.h5ad` files. After assembling all downloaded cells, we filtered out the cells labeled as *unknown* cell type and randomly subsampled $10,000$ cells, reflecting the random *OOD* pool of normal tissues from CellxGene that scTab's training and testing had not yet encountered. All cells passed our QC assessment.

OOD balanced tissue dataset Beyond random sampling, we also built a *balanced OOD* dataset of $10,000$ cells that more evenly covered multiple tissue types. Specifically, from the same post–2023-05-15 normal data obtained as explained above, we chose eight representative datasets of interest, as assessed by three criteria: (i) a relatively high number of unique cell types profiled; (ii) a relatively low fraction of cells labeled as unknown; (iii) a relatively well-studied tissue. We randomly drew $1,250$ cells from each tissue dataset, totaling $10,000$ cells (see Supplementary Table S2) for an overview of the datasets we chose from). The tissues we chose were blood, peripheral region of retina, breast, lung, trachea, kidney, pancreas and cerebellum. For the lung dataset, 382 of the $1,250$ cells had no match of the ground-truth label in Cell Ontology and were therefore excluded from evaluations.

A.2. Prompting

We crafted prompt templates that encode key metadata (*e.g.*, tissue of origin, disease status) and a ranked gene list (either cluster-level marker genes obtained via differential expression or single-cell–level top highly expressed genes). We tested both a long prompt, which encouraged a thorough rationale, as well as a shorter variant, which requested a briefer answer.

A.2.1. Cluster-level

We used a Python script to create prompts taking as input a CSV file in which each row described one cluster, including the top 10 marker genes determined by the differential expression analysis performed by Hou & Ji (2024), as well as relevant metadata (*e.g.* tissue, disease status, etc.). The script outputted a line-delimited JSON file suitable for ingestion by downstream LLM pipelines.

We constructed both a long prompt and a short prompt. Below is the exact text used for the long prompt:

```
"You are an expert in single-cell biology.\n\n"
"Below is metadata for one cluster of cells along with its marker genes:\n"
"Tissue: [TISSUE_PLACEHOLDER]\n"
"Disease: [DISEASE_PLACEHOLDER]\n"
"Development stage: [DEVELOPMENT_STAGE_PLACEHOLDER]\n"
"Marker genes: [GENE_LIST]\n\n"
"Please identify what cell type this might be, as granular and accurate as possible.\n"
"At the end of your response, strictly place the final lines in this format:\n\n"
"Cell type: X\n"
```

Conversely, the short prompt added the instruction to "keep your response concise and clear", while otherwise preserving the exact same structure:

```
"Please identify what cell type this might be, as granular and accurate as possible.\n"
"Keep your response concise and clear.\n"
"At the end of your response, strictly place the final lines in this format:\n\n"
"Cell type: X\n"
```

A.2.2. Single-cell level

Below is the skeleton of our long prompt for querying LLMs to annotate a single cell:

```
"You are an expert in single-cell biology.\n\n"
"Below is metadata for one cell, followed by a list of its genes in descending expression:\n"
"Tissue: [TISSUE]\n"
"Disease: [DISEASE]\n"
"Development stage: [DEVELOPMENT_STAGE]\n"
"Genes: [GENE_LIST]\n\n"
"Please identify what cell type this might be, as granular and accurate as possible.\n"
"At the end of your response, strictly place the final lines in this format:\n\n"
"Cell type: X\n"
```

As in the case of the cluster-level annotations, the short prompt contained the additional instruction: "Keep your response concise and clear."

A third variant of prompt incorporated a user-provided list of cell-type labels from which the LLM was tasked to choose, effectively enforcing the model to act like a multi-class classifier. In our experiments, we used two different sets of classifier labels: the labels curated by scTab (Fischer et al., 2024), and the labels provided by scGPT (Cui et al., 2024). The relevant lines added to the prompt were:

```
"Here is a list of all possible cell types you must choose from:\n"
"[MULTILINE_STRING_OF_CELL_TYPES]\n\n"
"Please pick the single best matching cell type from this list.\n"
```

We tested multiple values of N (the number of top expressed genes) for prompting: 5, 10, 25, 50, 100, 200, 500, $1,000$, $2,000$, as well as all the non-zero expressing genes. We settled on top 500 most highly expressed genes for final input prompts, as they achieved top accuracy and Macro-F1 score in our testing setting.

A.3. Performance evaluation

To evaluate the performance of the tested models, we considered as ground truth the labels curated and proposed by the original dataset. In order to compare model predictions with ground-truth labels, we first harmonized both predicted and ground-truth labels to official Cell Ontology (CL) terms, as done by Hou & Ji (2024). This step was necessary because, unlike multi-class classifier models like scTab, where the predicted cell-type labels are standardized to CL terms, LLM outputs could vary by case, wording, or description, even if conceptually in agreement with the ground-truth label. After the harmonization process, we removed the cells with no match (empty match) of the ground-truth label with the OLS database, specifically 94 cells from the *in-domain* validation pool, 87 cells from the *in-domain* test pool, 31 cells from the *OOD* random dataset, and 382 lung cells from the *OOD* balanced tissue dataset. Next, given the unambiguous ground-truth label and unambiguous predicted label, we applied the scTab framework (Fischer et al., 2024) as described in their GitHub page (`https://github.com/theislab/scTab/tree/devel`). Specifically, a match was achieved (return TRUE) if the predicted label was either identical (exact match) or a Cell Ontology descendant (predicted subtype) of the ground-truth label. Conversely, if the ground-truth label was a child of the predicted label or the classes shared no direct parent–child relationship, the result was incorrect (return FALSE). In other words, the predicted labels were not penalized for higher granularity than ground-truth labels. Finally, we compiled the cell-level binary TRUE/FALSE summary statistics to compute overall prediction accuracies and Macro-F1 scores for each dataset and for each model, as detailed below.

A.3.1. Accuracy

We computed the accuracy score as follows:

$$\text{Accuracy} = \frac{\text{number of cells correctly predicted}}{\text{total number of cells}}$$

A.3.2. Macro-F1

To obtain the macro-averaged F1 score, for every ground-truth cell-type label ℓ, we counted:

- $\mathbf{TP}_\ell$: cells whose ground-truth label is ℓ and were correctly predicted as either ℓ or an ontology descendant of ℓ (TRUE),
- $\mathbf{FN}_\ell$: cells whose ground-truth label is ℓ and were incorrectly predicted (FALSE),
- $\mathbf{FP}_\ell$: cells whose ground-truth label is not ℓ but whose predicted label is ℓ (FALSE).

We compute the per-label precision and recall metrics:

$$\text{Precision}_\ell = \frac{TP_\ell}{TP_\ell + FP_\ell}, \qquad \text{Recall}_\ell = \frac{TP_\ell}{TP_\ell + FN_\ell}$$

And the per-label F1 score:

$$F1_\ell = 2 \times \frac{\text{Precision}_\ell \times \text{Recall}_\ell}{\text{Precision}_\ell + \text{Recall}_\ell}$$

The Macro-F1 score is the average over all L ground-truth labels:

$$\text{Macro-F1} = \frac{1}{L} \sum_{\ell=1}^{L} F1_\ell$$

A.3.3. Weighted-F1

Weighted-F1 score gives each cell type label a weight proportional to its relative frequency in the dataset. Therefore, if for every ground-truth label ℓ we define $n_\ell = TP_\ell + FN_\ell$ as the number of cells with the true label ℓ, then

$$\text{Weighted-F1} = \frac{\sum_\ell n_\ell F1_\ell}{\sum_\ell n_\ell}$$

A.3.4. Bootstrapped Confidence Intervals

To complement the point estimates for the single-cell *in-domain* and random *OOD* evaluations, we report 95% confidence intervals (CI) around accuracy and Macro-F1 values, obtained through a percentile bootstrap procedure. For each dataset consisting of n cells (*in-domain* validation 9,906 cells, *in-domain* test 9,913 and *OOD* random 9,969), we drew 10,000 bootstrap resamples, each consisting of n cell indices selected with replacement from the original index set. The same sampled indices were applied to the ground-truth vector and to the prediction vectors, to generate the bootstrapped evaluation vectors for all models. On every bootstrapped evaluation vector, we recomputed the metric of interest (accuracy and Macro-F1) for each model exactly as described above. For each model, we took the 2.5-th and 97.5-th percentiles of the $10,000$ bootstrap replicates as the lower and upper bounds of the 95% CI. As the *OOD* balanced tissue dataset was composed specifically of eight tissues with only $1,250$ cells/tissue, we didn't bootstrap CIs for this dataset.

A.3.5. Pair-wise statistical significance testing

To determine whether performance differences between models were significantly different than expected by sampling variability, we carried out pair-wise significance tests separately for each dataset and for each metric (accuracy and Macro-F1). All p-values were adjusted for multiple testing by the Benjamini–Hochberg (Benjamini & Hochberg, 1995) false-discovery-rate (FDR) procedure with $\alpha = 0.05$.

Accuracy For every pair of models, we formed a 2×2 contingency table:

	model B	
	correct	wrong
model A correct	n_{11}	n_{10}
model A wrong	n_{01}	n_{00}

where n_{10} denotes cells called as correct by model A but wrong by model B and *vice versa* for n_{01}. Under the null hypothesis that the two models have identical accuracy, the number of discordant pairs favouring model A, n_{10}, follows a $\mathrm{Binomial}\big(N, 0.5\big)$ distribution with $N = n_{01} + n_{10}$. We therefore applied the exact McNemar test (McNemar, 1947) (two-sided, binomial formulation) implemented in statsmodels (Seabold & Perktold, 2010) 0.14.4. Given a total of k models tested, the resulting exact p-values were FDR-adjusted across the $\binom{k}{2}$ model pairs within the same dataset.

Macro-F1 Because Macro-F1 is not a per-cell statistic, we assessed significance via a paired non-parametric bootstrap (Efron, 1979) with $10,000$ resamples. For each resample, we first sampled n cell indices with replacement from the dataset (where n is the number of evaluation cells) as described at the beginning of this subsection. We then recomputed the Macro-F1 score for every model on the resample and recorded the difference for each pair of models A and B as $\Delta = \text{Macro–F1}_A - \text{Macro–F1}_B$. The empirical distribution of Δ over the $10,000$ replicates approximates its sampling distribution. A two-sided p-value was obtained as

$$p \;=\; 2 \times \min\big\{\Pr(\Delta \le 0),\, \Pr(\Delta \ge 0)\big\}.$$

Bootstrap p-values for all model pairs were then FDR-corrected as described above.

A.4. scGPT

We employed the scGPT (Cui et al., 2024) single-cell foundation model for benchmarking cell-type annotation at the single-cell level. Our approach utilized the scGPT package and followed established procedures from the scGPT GitHub tutorial for zero-shot annotation via embedding-similarity search, available here: `https://github.com/bowang-lab/scGPT/blob/main/tutorials/Tutorial_Reference_Mapping.ipynb`. The input data was provided as AnnData objects (.h5ad files) containing raw gene-expression counts. Initial preprocessing involved filtering the genes present in the input data to retain only those included in the scGPT model's vocabulary, ensuring compatibility between the input-data features and the pre-trained scGPT model. Subsequently, we identified the top 3,000 highly variable genes (HVGs) within the filtered gene set using the *scanpy.pp.highly_variable_genes* function with the flavor='seurat_v3' setting from the scanpy (Wolf et al., 2018) Python package. The AnnData object was then subsetted to include only these 3,000 HVGs for downstream processing.

Cell embeddings were generated using the scgpt.tasks.embed_data function from the scGPT library, applying the pre-trained scGPT whole-human model (checkpoint-0623 from scGPT GitHub) to obtain a low-dimensional representation for each cell.

Zero-shot cell-type annotation was performed using a nearest-neighbor approach based on these cell embeddings, which is the method recommended in scGPT tutorials. We utilized a pre-computed FAISS index built from reference cell embeddings derived from the CellXGene atlas (downloaded following scGPT GitHub instructions). For each cell in our input dataset, we queried the FAISS index to find the 50 nearest neighbors ($k = 50$) within the reference-atlas embedding space. The cell-type labels associated with these 50 neighbors were retrieved from the reference metadata. A final cell-type prediction for each input cell was determined by majority voting among the labels of its 50 nearest neighbors, using the voting function adapted from the scGPT repository's utility scripts. The resulting single-cell-level predictions were saved alongside ground-truth labels for subsequent evaluation.

A.5. scTab

We also employed the scTab (Fischer et al., 2024) multi-class classification model for benchmarking cell-type annotation at the single-cell level, following the scTab GitHub tutorials available here: https://github.com/theislab/scTab. Input data consisted of AnnData objects (.h5ad files) with raw gene-expression counts in the .X attribute. Consistent with the scTab protocol, the feature space was first streamlined. Genes present in the input data were filtered and ordered to match the reference gene set used for model training. This step ensured that the input-matrix dimensions and gene order align precisely with the model's expectations. The streamline_count_matrix function from the cellnet library was utilized for this alignment (Fischer et al., 2024). Following gene alignment, the count matrix underwent normalization. Each cell's expression profile was scaled to a total count of $10,000$, followed by a $\log(x + 1)$ transformation. This normalization method is standard practice for the scTab model.

For inference, we used the same pre-trained scTab checkpoint from their official GitHub tutorial (scTab-checkpoints/scTab/run5/val_f1_macro_epoch=41_val_f1_macro=0.847.ckpt). Corresponding model hyperparameters were loaded from the associated hparams.yaml file within the same directory. The TabNet classifier architecture was initialized using these parameters and loaded with the extracted state dictionary from the checkpoint. Inference was performed on the normalized data using a PyTorch DataLoader for batch processing, leveraging GPU acceleration. The model outputs raw logits, from which the predicted class index (representing the cell type) was determined using an argmax operation. Finally, these numeric prediction indices were mapped to human-readable cell-type labels using the official mapping file (merlin_cxg_2023_05_15_sf-log1p_minimal/categorical_lookup/cell_type.parquet). The resulting single-cell-level predictions were saved alongside ground-truth labels for subsequent evaluation.

A.6. Cell2Sentence-Scale-1B

We followed the exact instructions on the cell2sentence official GitHub-page tutorials for zero-shot cell-type annotations, especially tutorials 0 and 6 at the following link as of 10 June 2025: https://github.com/vandijklab/cell2sentence/tree/master/tutorials. Explicitly, we did not perform the simple QC steps recommended in tutorial 0, as the data we had were already QCed. In addition, we did not have ”batch_condition” metadata for our samples, and we used the "sex" metadata information when available. We used the newest release of the cell2sentence-scale model C2S-Scale-Pythia-1b-pt (https://huggingface.co/vandijklab/C2S-Scale-Pythia-1b-pt) in our evaluation, as this was the only new model released by the cell2sentence-scale team at the time of testing. We used the default number of 200 genes in the prompt, as shown in the GitHub tutorials.

B. Impact Statement

This paper presents work whose goal is to advance the field of Bioinformatics through advances in GenAI technologies. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

C. Supplementary Figures

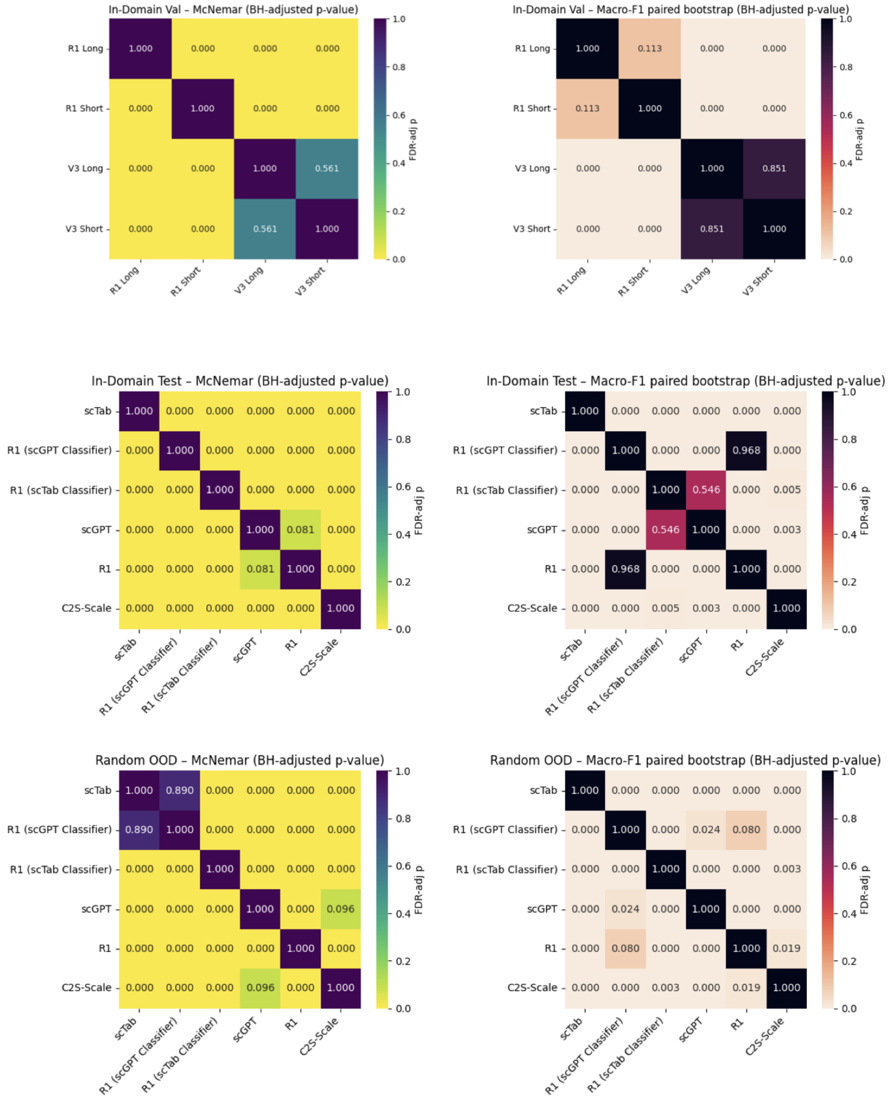


Figure S1. Heatmaps of statistical significance tests for the *in-domain* validation (top), *in-domain* test (middle), and *OOD* random (bottom) datasets for accuracy and Macro-F1 score, as described in Methods. For each dataset, we report two heatmaps: 1) Left-hand panels (yellow-green colormap): Benjamini-Hochberg-adjusted (Benjamini & Hochberg, 1995) p-values from the exact McNemar (McNemar, 1947) test applied to per-cell accuracy. 2) Right-hand panels (orange–purple colormap): Benjamini-Hochberg-adjusted p-values from a paired, non-parametric bootstrap (Efron, 1979) test on Macro-F1 score with $10,000$ bootstrap resamples.

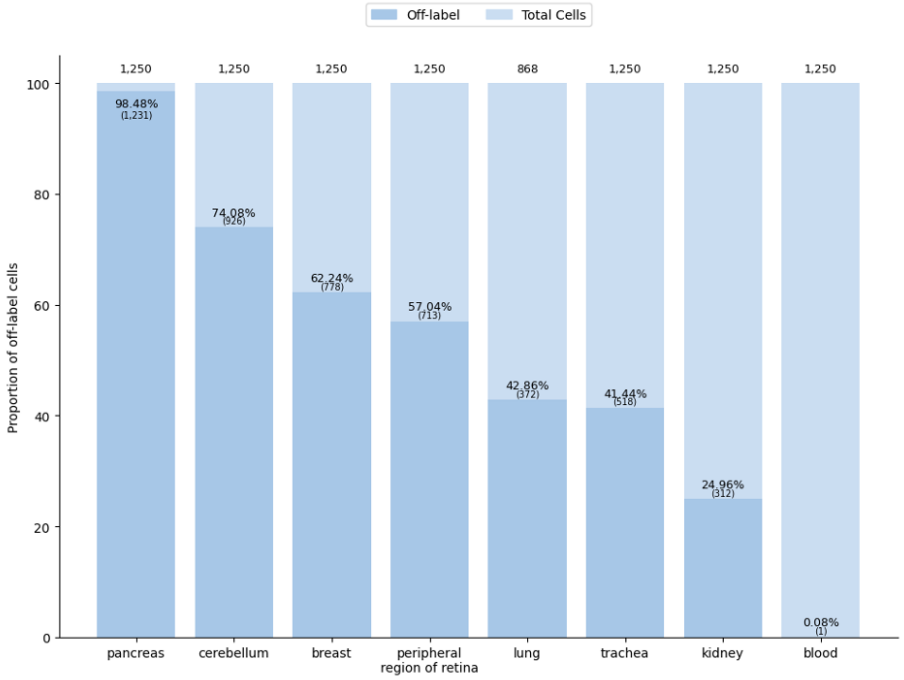


Figure S2. Bar plot showing the proportion of off-scTab-label (Fischer et al., 2024) cells by tissue types in the *OOD* balanced tissue dataset, with shades of color showing whether they were outside the labels employed by scTab (darker color indicates off-label). Percentages show the proportion of off-label cells for each tissue type, and absolute numbers show the number of cells in each category accordingly.

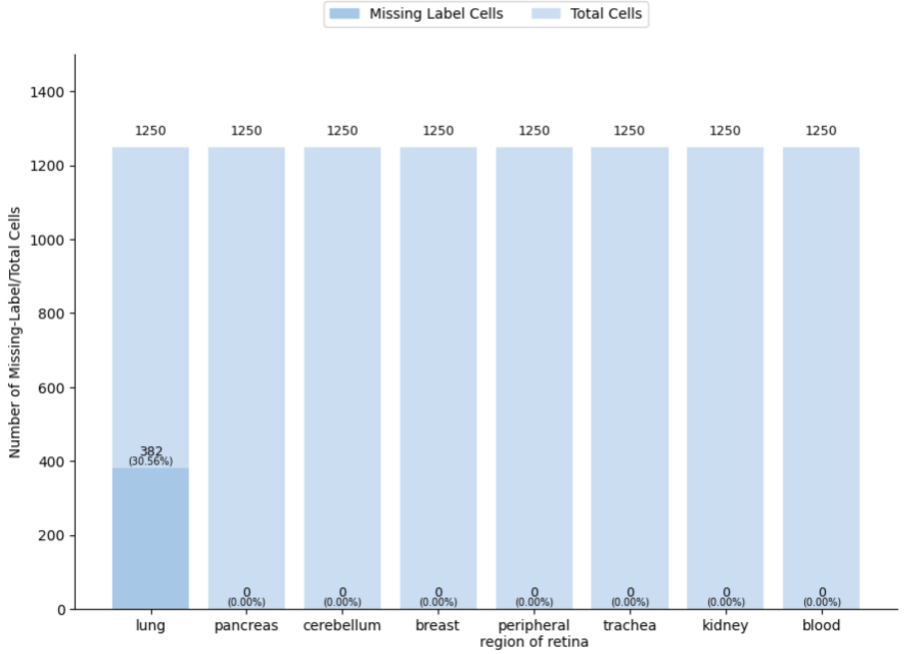


Figure S3. Bar plot showing the proportion of missing-label cells by tissue types in the *OOD* balanced tissue dataset, with shades of color showing whether they don’t have a ground truth cell type from the Cell Ontology Database (Côté et al., 2006) (darker color indicates missing-label). Percentages show the proportion of missing-label cells for each tissue type, and absolute numbers show the number of cells in each category accordingly.

D. Supplementary Tables

Table S1: Detailed Performance of Models

Dataset	Model and Prompt Mode	Accuracy	Macro-F1	Weighted F1	Acc. CI$_{lo}$	Acc. CI$_{hi}$	Macro-F1$_{lo}$	Macro-F1$_{hi}$
Cluster-level	R1 long	0.4628	0.3069	0.4876				
Cluster-level	R1 short	0.4584	0.2914	0.4811				
Cluster-level	V3-0324 long	0.4177	0.2996	0.4575				
Cluster-level	V3-0324 short	0.3823	0.2656	0.407				
Cluster-level	GPT-4o-0326 short	0.3699	0.1991	0.3838				
sc in-domain val	R1 long 100genes	0.4180	0.2492	0.4241	0.4082	0.4276	0.2357	0.2648
sc in-domain val	R1 short 100genes	0.4044	0.2309	0.4050	0.3947	0.4142	0.2184	0.2458
sc in-domain val	V3 long 100genes	0.3279	0.1501	0.3417	0.3185	0.3373	0.1392	0.1619
sc in-domain val	V3 short 100genes	0.3273	0.1498	0.3411	0.3178	0.3366	0.1385	0.1616
sc in-domain test	R1 500genes (unconstrained)	0.4596	0.2792	0.4988	0.4497	0.4691	0.2680	0.2970
sc in-domain test	R1 500genes (scGPT classifier)	0.5517	0.2915	0.5676	0.5420	0.5614	0.2835	0.3091
sc in-domain test	R1 500genes (scTab classifier)	0.5926	0.3083	0.5948	0.5827	0.6024	0.2978	0.3275
sc in-domain test	scTab	0.8837	0.7900	0.8852	0.8773	0.8900	0.7726	0.8083
sc in-domain test	scGPT	0.4708	0.2881	0.4684	0.4612	0.4808	0.2695	0.3027
sc in-domain test	C2S-Scale-200genes	0.6188	0.2770	0.5740	0.6090	0.6281	0.2676	0.2943
OOD-random	R1 500genes (unconstrained)	0.3362	0.2073	0.3572	0.3269	0.3457	0.2036	0.2290
OOD-random	R1 500genes (scGPT classifier)	0.4721	0.2230	0.4829	0.4623	0.4820	0.2092	0.2370
OOD-random	R1 500genes (scTab classifier)	0.4504	0.1667	0.4369	0.4405	0.4602	0.1663	0.1860
OOD-random	scTab	0.4730	0.2158	0.4544	0.4630	0.4829	0.2105	0.2351
OOD-random	scGPT	0.3817	0.1677	0.3895	0.3720	0.3914	0.1636	0.1855
OOD-random	C2S-Scale-200genes	0.3736	0.1407	0.3434	0.3640	0.3833	0.1350	0.1551
OOD balanced-tissue	R1 500genes (unconstrained)	0.2818	0.2327	0.3016				
OOD balanced-tissue	R1 500genes (scGPT classifier)	0.3771	0.2101	0.3728				
OOD balanced-tissue	R1 500genes (scTab classifier)	0.2375	0.1674	0.2493				
OOD balanced-tissue	scGPT	0.2167	0.1371	0.2339				
OOD balanced-tissue	scTab	0.2097	0.1598	0.1874				
OOD balanced-tissue	C2S-Scale-200genes	0.2262	0.1121	0.2245				

Table S2: Dataset Tissue Cell Counts

tissue	cell count	unknown	unique cell types	dataset total
peripheral region of retina	1062127	0	29	1378557
macula lutea	300870	0	29	1378557
macula lutea proper	15560	0	23	1378557
dorsolateral prefrontal cortex	741446	0	18	741446
blood	685024	0	25	685024
frontal cortex	448964	8200	5	684621
cingulate cortex	72550	3257	5	684621
temporal cortex	64590	2312	5	684621
cerebral cortex	39001	255	4	684621
insular cortex	32659	1494	5	684621
primary motor cortex	8259	141	5	684621
lateral ganglionic eminence	6383	65	4	684621
caudal ganglionic eminence	4669	29	4	684621
ganglionic eminence	4120	6	3	684621
medial ganglionic eminence	3426	11	3	684621
breast	525298	26472	25	525298
decidua	121211	2373	21	283558
decidua basalis	85254	0	22	283558
placenta	66104	395	21	283558
blood	10989	0	12	283558
primary visual cortex	241077	0	18	241077
skin epidermis	99508	0	20	195739
dermis	96231	0	20	195739
lung	46325	0	59	193108
lower lobe of left lung	46091	0	57	193108
bronchus	42317	0	60	193108
upper lobe of left lung	29983	0	56	193108
trachea	28392	0	58	193108
dorsolateral prefrontal cortex	172120	0	25	172120
subcutaneous adipose tissue	86064	0	15	166149
omental fat pad	80085	0	16	166149
blood	158726	0	10	158726
cerebellum	120042	8653	24	153789
hemisphere part of cerebellar posterior lobe	15503	77	12	153789
cerebellar cortex	9551	98	15	153789
dentate nucleus	8693	666	11	153789
primary somatosensory cortex	153159	0	18	153159
primary auditory cortex	139054	0	18	139054
anterior cingulate cortex	135462	0	18	135462
entorhinal cortex	104240	0	10	124917
ganglionic eminence	20677	0	9	124917
cortex of kidney	70118	0	26	122444
renal medulla	34019	0	26	122444
renal papilla	17958	0	23	122444
kidney	349	0	21	122444
breast	117346	0	10	117346
angular gyrus	110752	0	18	110752
skin of forearm	32138	0	1	87732

Continued on next page

Table S2: Dataset Tissue Cell Counts

tissue	cell count	unknown	unique cell types	dataset total
skin of body	27759	0	1	87732
skin of pes	9887	0	1	87732
skin of leg	3961	0	1	87732
hindlimb skin	3443	0	1	87732
skin of abdomen	2801	0	1	87732
skin of hip	2272	0	1	87732
skin of chest	1526	0	1	87732
lower leg skin	1475	0	1	87732
skin of cheek	1235	0	1	87732
skin of trunk	1161	0	1	87732
skin of shoulder	37	0	1	87732
skin of face	37	0	1	87732
skin of forehead	32099	0	25	82259
skin of cheek	21198	0	24	82259
skin of external ear	9366	0	23	82259
nose skin	6618	0	24	82259
arm skin	6529	0	25	82259
skin of temple	6449	0	23	82259
substantia nigra pars compacta	80576	0	1	80576
blood	58480	1604	29	77358
kidney	10437	1401	42	77358
adrenal tissue	6439	203	31	77358
perirenal fat	2002	119	30	77358
placenta	74685	0	7	74685
cerebellum	69174	0	18	69174
gingiva	60811	0	27	60811
breast	52681	0	6	52681
middle temporal gyrus	15519	636	18	47432
primary visual cortex	7851	201	18	47432
primary auditory cortex	6470	233	18	47432
primary motor cortex	5955	280	18	47432
anterior cingulate gyrus	5939	340	18	47432
primary somatosensory cortex	5698	295	18	47432
blood	45787	0	20	45787
cervical spinal cord white matter	19650	0	15	45528
white matter of cerebellum	15963	0	15	45528
Brodmann (1909) area 4	9915	0	15	45528
caudate lobe of liver	45186	0	3	45186
caudate nucleus	23999	0	10	44449
putamen	20450	0	10	44449
neural tube	43462	23	9	43462
kidney	43380	0	25	43380
frontal cortex	43033	0	6	43033
lamina propria of small intestine	21273	0	1	32926
lamina propria of large intestine	11653	0	1	32926
frontal cortex	30430	0	6	30430
perifoveal part of retina	25356	0	12	30401
fovea centralis	5045	0	12	30401
lamina propria of large intestine	19321	0	1	28758

Continued on next page

Table S2: Dataset Tissue Cell Counts

tissue	cell count	unknown	unique cell types	dataset total
lamina propria of small intestine	9437	0	1	28758
occipital cortex	22835	0	6	22835
entorhinal cortex	16416	0	9	20470
ganglionic eminence	4054	0	8	20470
blood	19631	0	9	19631
occipital cortex	19270	0	6	19270
dorsolateral prefrontal cortex	18400	0	18	18400
lamina propria of large intestine	9289	0	1	17706
lamina propria of small intestine	8417	0	1	17706
submucosa of ileum	9200	0	1	16338
submucosa of ascending colon	7138	0	1	16338
caudate lobe of liver	13719	1497	8	13719
bone marrow	13299	19631	16	13299
caudate lobe of liver	11982	0	6	11982
tendon of semitendinosus	10533	0	11	10533
nasopharynx	8874	0	17	8874
caudate lobe of liver	8855	660	6	8855
caudate lobe of liver	8219	0	6	8219
bone marrow	6701	0	6	6701
brain white matter	5230	0	9	6591
brain	1361	0	9	6591
peripheral region of retina	4335	0	13	6061
fovea centralis	1726	0	11	6061
scalp	3029	0	5	3029
islet of Langerhans	2282	262	6	2282
pancreas	2126	0	10	2126
caudate lobe of liver	1073	168	3	1073
caudate lobe of liver	1051	116	3	1051
caudate lobe of liver	474	358	1	474
right cardiac atrium	1	0	1	1

Table S3: Number of Cell Types of Tissues in scTab Dataset

tissue general label	n unique cell types
abdomen	8
abdominal wall	7
adipose tissue	14
adrenal gland	8
ascitic fluid	8
axilla	8
blood	75
bone marrow	48
brain	54
breast	37
colon	54
digestive system	3
endocrine gland	41
esophagogastric junction	20
esophagus	39
exocrine gland	14
eye	27
fallopian tube	15
heart	31
immune system	16
intestine	8
kidney	43
lamina propria	25
large intestine	40
liver	54
lung	81
lymph node	65
mucosa	10
musculature	35
nose	46
omentum	29
ovary	15
pancreas	6
paracolic gutter	8
parietal peritoneum	8
peritoneum	8
placenta	11
pleural fluid	17
prostate gland	24
reproductive system	18
respiratory system	59
saliva	10
scalp	1
skeletal system	7
skin of body	40
small intestine	68
spinal cord	11
spleen	59

Continued on next page

Table S3: Number of Cell Types of Tissues in scTab Dataset

tissue general label	n unique cell types
stomach	29
tongue	8
trunk	1
uterus	9
vasculature	1
yolk sac	19