Bioinformatics Advances, 2026, **6**(1), vbag001
https://doi.org/10.1093/bioadv/vbag001
Published: 27 January 2026
Application Note | Software

OXFORD

Cell type annotation using large language models (LLMs) and CytoAnalyst

Khoi Nguyen[1], Duy Tran[2], Phuong Nguyen[2], Seungil Ro[3], Phi Bya[2], Tin Nguyen[1,4,*]

[1]Department of Industrial and Systems Engineering, Wayne State University, Detroit, MI 48201, United States
[2]Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, United States
[3]Department of Physiology and Cell Biology, University of Nevada School of Medicine, Reno, NV 89557, United States
[4]Karmanos Cancer Institute, Wayne State University School of Medicine, Detroit, MI 48201, United States
*Corresponding author. Department of Industrial and Systems Engineering, Wayne State University, Detroit, MI 48201, United States. E-mail: tin@wayne.edu
Associate Editor: Alex Bateman

Abstract

Motivation: **Cell annotation is fundamental for single-cell data interpretation. Accurate annotation allows us to identify cell types, understand their functions, trace developmental trajectories, and pinpoint alterations associated with a condition of interest. However, this complex process demands extensive manual curation, domain expertise, and proficiency across diverse bioinformatics tools. These challenges impede reproducibility and consistency.**

Results: **We have developed a new approach for semi-automatic cell type annotation, powered by large language models (LLMs). Given the input single-cell data, we first perform dimension reduction, clustering, and differential analysis to identify distinct cell groups and their respective markers. Next, we utilize Meta's Llama and structured prompting to infer potential cell types. This approach greatly reduces manual labor from researchers while maintaining biological accuracy through enforced ontology, tissue context, and marker gene signatures. Our solution is freely accessible through our web-based platform named CytoAnalyst, hosted on a high-performance infrastructure with optimized networking and storage capabilities. CytoAnalyst also offers capabilities for quality control, embedding analysis, clustering, differential analysis, gene set analysis, cell enrichment, cell type annotation, and pseudo-time trajectory inference.**

Availability and implementation: **CytoAnalyst is freely available at https://cytoanalyst.tinnguyen-lab.com/. The CytoAnalyst handbook, including step-by-step tutorials and example case studies, is available at https://cytoanalyst.tinnguyen-lab.com/docs/.**

1 Introduction

Cell type annotation is a crucial yet challenging step in the analysis of single-cell RNA sequencing data (scRNA-Seq) (Pasquini *et al.* 2021, Hou and Ji 2024). Despite advances in single-cell technologies and method development, accurate identification of cell types using scRNA-Seq remains complex. The process typically involves a sequence of steps: quality control, normalization, dimensionality reduction, clustering, and differential analysis to pinpoint biomarkers. The final step—cell type identification—heavily relies on expert knowledge and manual curation using biomarkers (Stuart *et al.* 2019, Cheng *et al.* 2023). Such a process is inherently time-consuming, subjective, and prone to inconsistencies across different studies (Clarke *et al.* 2021, Hao *et al.* 2021, Quan *et al.* 2023).

There exist tools that attempt to automate certain steps in the annotation process, but they lack the flexibility needed to handle diverse experimental contexts, or cannot identify novel cell types (Aran *et al.* 2019, Pliner *et al.* 2019, Pasquini *et al.* 2021, Ji *et al.* 2023). Additionally, many tools focus on specific analytical aspects but lack integrated workflows that smoothly connect data processing, visualization, and annotation in a unified framework (Heumos *et al.* 2023).

With the advancement of AI, especially Large Language Models (LLMs), we have the opportunity to automate many steps in the interpretation of complex biomedical data (Jumper *et al.* 2021, Lu *et al.* 2024, Boehm *et al.* 2025). Foundation LLMs, trained on massive data corpora and biomedical knowledge, can be fine-tuned or adjusted for the inference of cell types using tissue context, biomarkers, and cell ontology hierarchy. Notably, GPTCelltype represents the first methodology to leverage GPT-4 for inferring cell type names from marker genes (Hou and Ji 2024). However, the effective integration of LLM-based inference with traditional bioinformatics workflows for single-cell annotation is still largely unexplored.

In this manuscript, we introduce a comprehensive annotation workflow that leverages the power of LLMs to identify potential cell types from scRNA-Seq. The workflow is freely accessible

Received: 15 August 2025. **Revised:** 4 November 2025. **Accepted:** 23 December 2025

through CytoAnalyst (Bya *et al.* 2025), a web-based platform hosted on our internal servers. The platform provides a complete single-cell data analysis pipeline, as well as free storage and computational resources for all researchers, supporting comprehensive analyses of large single-cell datasets (Section 2, available as supplementary data at *Bioinformatics Advances* online and Fig. 10, available as supplementary data at *Bioinformatics Advances* online).

2 Workflow

Figure 1 shows the overall workflow of our annotation approach. The analysis begins with data upload and quality control, which filters low-quality cells and genes based on expression metrics, followed by a marker discovery to identify marker genes for LLM-powered inference and interactive annotation. The detailed pipeline is shown in Section 1, available as supplementary data at *Bioinformatics Advances* online.

2.1 Data upload and quality control

The annotation workflow begins with data upload and quality control (Fig. 1A). Accepted input includes 10X Genomics Cell Ranger output (.tar.gz or .h5) and AnnData (.h5ad) objects, along with optional metadata containing sample information and experimental conditions for cell type identification. The platform visualizes key metrics, unique molecular identifier counts, gene detection rates, and mitochondrial gene percentages (Fig. 1, available as supplementary data at *Bioinformatics Advances* online), allowing users to filter out low-quality cells and genes.

2.2 Cell segmentation and markers

Following quality control, users can transform the high-dimensional expression data into a transcriptome landscape where cells of similar types cluster together (Fig. 1B1). The workflow includes a comprehensive embedding analysis with customizable parameters for normalization, variable gene selection, and dimensionality reduction. The software supports multiple normalization methods and sample integration approaches to ensure that cell type identification reflects biological differences rather than technical batch effects (Fig. 2, available as supplementary data at *Bioinformatics Advances* online). Users can skip this step if the dataset already contains calculated embeddings, reducing computational overhead.

The next step is cluster analysis, which groups similar cells into distinct populations that correspond to potential cell types (Fig. 1B2). We implement three clustering algorithms to accommodate diverse data characteristics: Louvain (Blondel *et al.* 2008), Leiden (Traag *et al.* 2019), and k-means (Kodinariya and Makwana 2013). Users can create multiple clustering analysis instances simultaneously with different parameters to identify optimal cell type boundaries. The software enables visualization of clustering results across different embeddings and comparison of outcomes, ensuring identified populations accurately reflect biological cell types rather than analytical artifacts (Fig. 3, available as supplementary data at *Bioinformatics Advances* online).

After cluster analysis, the next step is to perform differential analysis to identify marker genes of each cell group. These marker genes represent molecular signatures essential for cell type annotation (Fig. 1B3). The platform supports five methods: Wilcoxon rank-sum test (Wilcoxon *et al.* 1970), MAST (Finak *et al.* 2015), ClusterDE (Song *et al.* 2025), DESeq2 (Love *et al.* 2014), and logistic regression (Butler *et al.* 2018), all followed by Benjamini-Hochberg adjustment for multiple comparisons (Benjamini and Hochberg 1995). It is important to note that relying on *P*-values alone to select marker genes will likely lead to many false positives. This is because each cell group often consists of many cells, meaning even a small change in expression values can yield a statistically significant *P*-value. Therefore, the software integrates interactive visualization with advanced filtering capabilities to identify marker genes that are specific for each cell group (Fig. 5, available as supplementary data at *Bioinformatics Advances* online).

Alternatively, users can perform embedding, clustering, and differential analyses externally using their preferred tools and upload the results in AnnData format (.h5ad). They can also upload external gene lists (e.g., gene markers extracted from differential analysis, gene sets from curated databases) as .gmt files using the geneset management module.

2.3 LLM-powered cell type inference and interactive annotation

We implement an analysis workflow that enables automated cell annotation yet allows for flexible customization (Fig. 1C). For automated cell type prediction, we leverage Llama 3.3, a state-of-the-art LLM developed by Meta and accessed through the Ollama API gateway (Grattafiori *et al.* 2024). We utilize the pre-trained LLM without additional training or fine-tuning. To counter potential API failures and unreliable responses, we implement an automatic 20-try retry process. We inject user-provided gene markers and tissue information into a carefully crafted prompt that enforces strict ontological compliance with the Cell Ontology and CellxGenes databases (Fig. 1C1).

We set up the prompt to generate hierarchical cell type classifications and follow the format “Cell type → A → B → C → D → E →Cells: [markers]”, which enforces at least a 4-level parent hierarchy, with “Cells” being the top ancestor. After obtaining the results from LLMs, we perform a post-processing validation using pattern matching to extract cell type hierarchy, followed by gene filtering against the input set to keep only the supported markers in the predictions. The LLM is prompted to predict cell lineages with associated marker genes, which are subsequently filtered to retain only those present in the user-provided marker list. If all predicted lineages lack supporting markers after this filtering, the system reruns the prompt until it obtains lineages with validated marker support. This approach combines the LLM’s ability to process the hierarchical prompt with domain-specific biological knowledge while maintaining strict output formatting and validation requirements. This LLM-based cell annotation greatly reduces the manual labor required for initial cell type assignment. Researchers can further validate annotation results through additional analysis and validation (Fig. 7, available as supplementary data at *Bioinformatics Advances* online).

Additionally, we provide an interactive interface where researchers can assign predicted cell types to corresponding

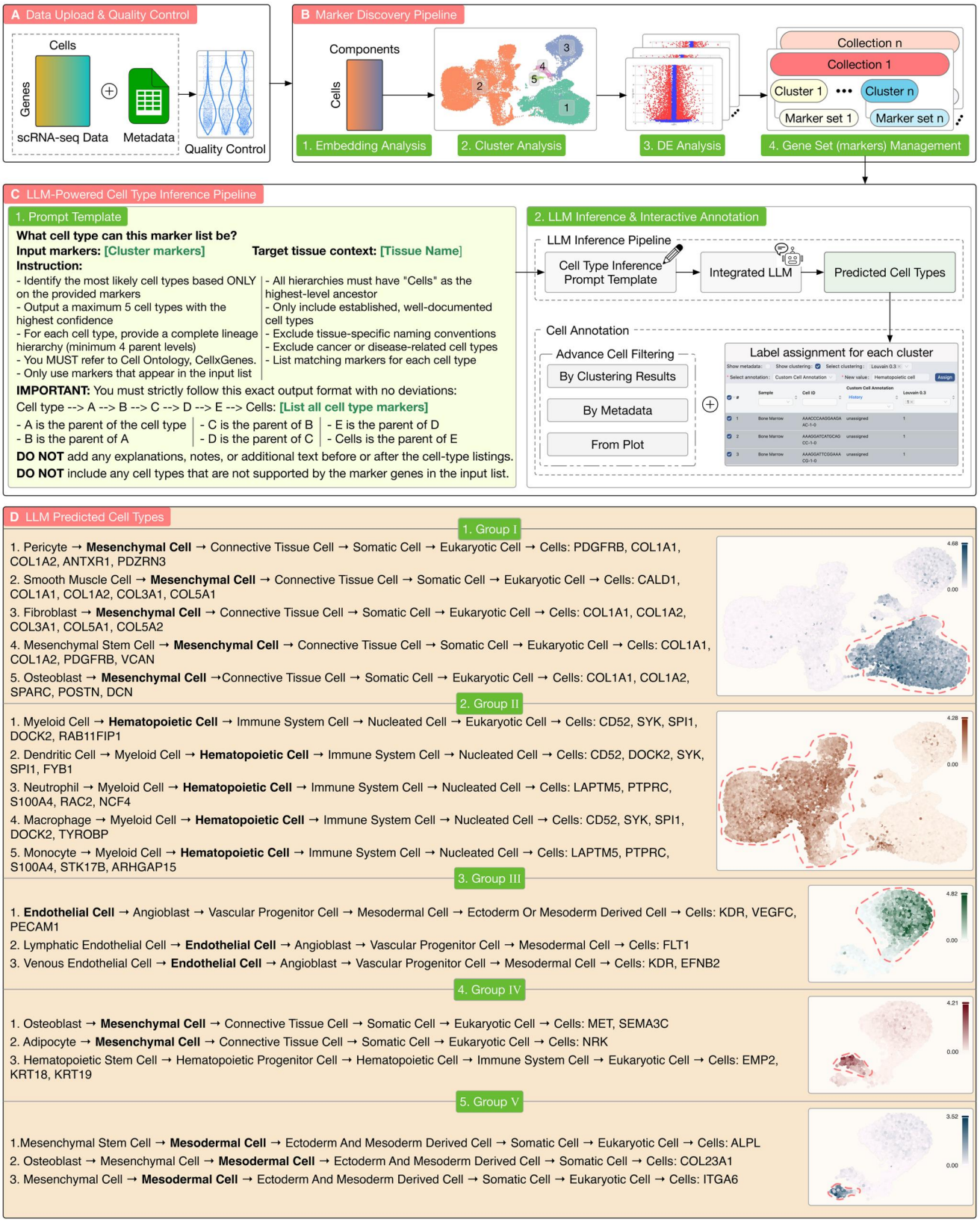

Figure 1 LLM-powered cell annotation for single-cell RNA sequencing data (scRNA-Seq). (A) Data upload and quality control. Data upload imports single-cell data (in 10X Genomics Cell Ranger or AnnData format) and metadata, while quality control filters out low-quality cells and genes. (B) Marker discovery pipeline, including: (1) embedding analysis, (2) visualization, (3) clustering, and (4) marker discovery through interactive differential analysis. (C) LLM-powered cell type inference and interactive annotation. The inference workflow uses a structured prompt template that guides the LLM to predict potential cell types using the provided gene sets, tissue information, and cell ontology, ensuring biologically meaningful predictions. The interactive annotation interface allows users to combine automatic annotation and domain expertise with advanced cell filtering capabilities. (D1–D5) Analysis results of the case study using bone marrow organoids. The left panels show the inferred lineage hierarchies and predicted cell types, while the right panels show the expression patterns of marker genes for the five cell groups identified by the platform.

clusters and edit annotations (Fig. 1C2). The process combines LLM inference with manual curation capabilities, allowing researchers to create biologically meaningful cell type labels. The software supports two manual annotation strategies: (i) a systematic approach where users can automatically assign cell types to clusters based on marker genes enrichment results (Fig. 8C, available as supplementary data at *Bioinformatics Advances* online), and (ii) a targeted approach, where users can select individual cells or clusters to assign cell types using marker genes expression and also enrichment results (Fig. 9, available as supplementary data at *Bioinformatics Advances* online). Users can combine these strategies to create a flexible annotation workflow that suits their needs. The purpose is to ensure that the final annotation reflects multiple lines of evidence, enhancing confidence in cell type assignments.

3 Results

3.1 Case study of bone marrow organoids

We demonstrate the capabilities of the proposed approach through the analysis of a single-cell dataset (31,040 cells), which contains bone marrow organoids generated from human induced pluripotent stem cells (Frenz-Wiessner *et al.* 2024). The analysis results are summarized in Fig. 1D (D1–D5), and the details are provided in Section 3, available as supplementary data at *Bioinformatics Advances* online. The dataset is available at https://cellxgene.cziscience.com/collections/59cd85c5-3b22-4035-b628-2a20810ad54b.

Following embedding analysis and visualization, we identify three potential cell populations (I, II, III) that correspond to the three cell islands (Fig. 11A, available as supplementary data at *Bioinformatics Advances* online). Louvain clustering with different parameters identifies substantially more cell clusters (Fig. 11B and C, available as supplementary data at *Bioinformatics Advances* online), with the number of clusters in the range of 10–20. However, all clustering results also confirm that the data can be divided into three main populations. Next, we perform differential analysis and visualize the expression patterns of the marker genes for each population. The patterns of the markers in population III, combined with cluster analysis, reveal that the cells in this population are not truly homogeneous and can be divided into three different subgroups (Fig. 11D–G, available as supplementary data at *Bioinformatics Advances* online). Through differential analysis, visualization, and cluster analysis verification, we identify five cell groups and their markers.

For each of the five cell groups, the software performs LLM-based inference to identify biologically meaningful cell type lineages (Table 1, available as supplementary data at *Bioinformatics Advances* online). Panels D1–D5 in Fig. 1D show the expression patterns of the markers for each cell population, along with the predicted lineages. Following our selection strategy, we choose the most frequently appearing label in the top 5 lineages as the predicted cell type. If there are multiple labels with the same frequency, we choose the cell type with the lowest order (most fine-grained) in the cell ontology.

At the end, based on the LLM inference, we label the cell groups as follows: (i) mesenchymal cells for Groups I and IV (Fig. 1D1 and D4), (ii) hematopoietic cells for Group II (Fig. 1D2), (iii) endothelial cells for Group III (Fig. 1D3), and (iv) mesodermal cells for Group V (Fig. 1D5). The LLM annotation aligns well with the original annotation provided by the authors of the single-cell dataset (Fig. 12, available as supplementary data at *Bioinformatics Advances* online). The figure shows that the two annotations share 99% similarity, with the difference being that the LLM assigns cells in group IV to mesenchymal instead of epithelial cells. We hypothesize that the authors of the dataset were able to distinguish between the two cell types using external evidence from flow cytometry data (e.g., cell size, morphology, etc.), evidence that may not be present or visible in the gene expression data that we analyze (Frenz-Wiessner *et al.* 2024).

3.2 Systematic benchmarking

We perform a comprehensive assessment of the proposed approach using four different tissues. To demonstrate the suitability of LLMs in cell type annotation, we now include four LLMs, Llama 3.3 (70B parameters) (Grattafiori *et al.* 2024), Gemma 3 (27B) (Team *et al.* 2025), TxGemma (27B) (Wang *et al.* 2025), and Qwen 3 (30B) (Yang *et al.* 2025), and three standard annotation tools, scCATCH (Shao *et al.* 2020), scType (Ianevski *et al.* 2022), and SingleR (Aran *et al.* 2019).

The four datasets collectively contain 28 distinct cell types. The Bone Marrow dataset consists of 31,040 cells from bone marrow organoids, encompassing five cell types: mesenchymal cell, hematopoietic cell, endothelial cell, epithelial cell, and mesodermal cell (Frenz-Wiessner *et al.* 2024). The Breast Cancer dataset contains 35,214 cells from 26 primary breast tumors that have five cell types: CD4-positive, CD8-positive (alpha-beta T cell), mature NK T cell, natural killer cell, and T cell (Wu *et al.* 2021). The Lobes of Liver dataset includes 16,665 cells sourced from 24 neurologically deceased donor human livers and contains eight cell types: CD4-positive, CD8-positive (alpha-beta T cell), erythroblast, hepatic pit cell, lymphocyte, mature B cell, myeloid cell, and natural killer cell (Andrews *et al.* 2024). Finally, the Zone of Skin dataset contains 15,457 cells from five healthy male donors and has the highest diversity with ten cell types: endothelial cell of lymphatic vessel, endothelial cell of vascular tree, erythrocyte, keratinocyte, macrophage, melanocyte, pericyte, skin fibroblast, stem cell of epidermis, and T cell (Solé-Boldo *et al.* 2020).

For each dataset, we group cells based on annotated cell types provided by the data sources. Assuming that the cell types are unknown, we follow the instructions of scCATCH, scType, and SingleR to identify the cell type of each cell group. The method scCATCH uses self-identified DE genes and user-provided tissue/species to query a built-in database, scoring candidates by matched markers and supporting publications. The method scType queries its own database for tissue-related markers and scores cell types based on the expression level of those markers within the group. In contrast, SingleR bypasses DE gene identification, assigning cell type labels by calculating the correlation between the expression profile of each cell group and the profiles in the reference dataset (the Human Primary Cell Atlas was used). Ultimately, all three methods assign the

Model	DE Metrics	Top DE Genes	Bone Marrow	Breast Cancer	Lobes of Liver	Zone of Skin	Row Average	DE Metric Average	Model Average	Accuracy Summary
Llama 3.3 70B	sig_weighted_logfc	10	0.600	0.600	0.750	0.600	**0.638**	**0.733**	**0.776**	
		15	0.600	0.600	0.875	0.800	**0.719**			
		20	1.000	0.800	0.875	0.700	**0.844**			
	exp_pct_diff	10	1.000	0.600	0.625	0.600	**0.706**	**0.752**		
		15	1.000	0.600	0.625	0.700	**0.731**			
		20	1.000	0.600	0.875	0.800	**0.819**			
	pct_weighted_logfc	10	1.000	0.600	0.625	0.800	**0.756**	**0.798**		
		15	0.800	0.600	0.875	0.900	**0.794**			
		20	1.000	0.600	0.875	0.900	**0.844**			
	pct_weighted_sig	10	1.000	0.600	0.750	0.900	**0.813**	**0.817**		
		15	1.000	0.600	0.875	0.800	**0.819**			
		20	1.000	0.600	0.875	0.800	**0.819**			
	mixed_logfc_sig	10	1.000	0.800	0.625	0.800	**0.806**	**0.779**		
		15	1.000	0.600	0.750	0.700	**0.763**			
		20	0.800	0.600	0.875	0.800	**0.769**			
Gemma 3 27B	sig_weighted_logfc	10	0.600	0.600	0.500	0.800	**0.625**	**0.673**	**0.709**	
		15	0.800	0.800	0.750	0.600	**0.738**			
		20	0.600	0.600	0.625	0.800	**0.656**			
	exp_pct_diff	10	0.600	0.600	0.625	0.900	**0.681**	**0.738**		
		15	1.000	0.600	0.750	0.800	**0.788**			
		20	0.800	0.600	0.875	0.700	**0.744**			
	pct_weighted_logfc	10	1.000	0.600	0.750	0.800	**0.788**	**0.744**		
		15	0.800	0.600	0.500	0.800	**0.675**			
		20	0.800	0.600	0.875	0.800	**0.769**			
	pct_weighted_sig	10	0.800	0.600	0.750	0.900	**0.763**	**0.715**		
		15	0.800	0.600	0.625	0.800	**0.706**			
		20	0.800	0.600	0.500	0.800	**0.675**			
	mixed_logfc_sig	10	0.600	0.600	0.375	0.700	**0.569**	**0.675**		
		15	0.600	0.600	0.750	0.800	**0.688**			
		20	0.800	0.600	0.875	0.800	**0.769**			
TxGemma 27B	sig_weighted_logfc	10	0.800	0.600	0.875	0.700	**0.744**	**0.688**	**0.731**	
		15	1.000	0.400	0.625	0.600	**0.656**			
		20	0.800	0.400	0.750	0.700	**0.663**			
	exp_pct_diff	10	1.000	0.400	0.875	0.800	**0.769**	**0.750**		
		15	1.000	0.400	0.875	0.700	**0.744**			
		20	1.000	0.400	0.750	0.800	**0.738**			
	pct_weighted_logfc	10	1.000	0.400	0.750	0.900	**0.763**	**0.754**		
		15	1.000	0.400	0.875	0.800	**0.769**			
		20	0.800	0.600	0.625	0.900	**0.731**			
	pct_weighted_sig	10	1.000	0.400	0.750	0.700	**0.713**	**0.729**		
		15	1.000	0.400	0.875	0.800	**0.769**			
		20	1.000	0.400	0.625	0.800	**0.706**			
	mixed_logfc_sig	10	0.800	0.600	0.625	0.800	**0.706**	**0.733**		
		15	1.000	0.600	0.750	0.900	**0.813**			
		20	1.000	0.400	0.625	0.700	**0.681**			
Qwen 3 30B	sig_weighted_logfc	10	1.000	0.400	0.750	0.900	**0.763**	**0.715**	**0.708**	
		15	0.600	0.600	0.750	0.900	**0.713**			
		20	0.800	0.600	0.375	0.900	**0.669**			
	exp_pct_diff	10	0.800	0.400	0.625	0.900	**0.681**	**0.671**		
		15	0.800	0.400	0.625	0.900	**0.681**			
		20	0.800	0.400	0.500	0.900	**0.650**			
	pct_weighted_logfc	10	0.600	0.400	0.875	0.900	**0.694**	**0.698**		
		15	0.800	0.400	0.875	0.800	**0.719**			
		20	0.800	0.600	0.625	0.700	**0.681**			
	pct_weighted_sig	10	0.800	0.600	0.875	0.600	**0.719**	**0.731**		
		15	0.800	0.400	0.875	0.900	**0.744**			
		20	0.800	0.600	0.625	0.900	**0.731**			
	mixed_logfc_sig	10	0.800	0.400	0.625	0.800	**0.656**	**0.723**		
		15	0.800	0.600	0.625	1.000	**0.756**			
		20	1.000	0.600	0.625	0.800	**0.756**			
scCATCH			0.600	0.600	0.000	0.300	**0.375**	**0.375**	**0.375**	
scType			0.800	0.800	0.000	0.100	**0.425**	**0.425**	**0.425**	
SingleR			0.600	0.200	0.500	0.400	**0.425**	**0.425**	**0.425**	

Figure 2 Benchmarking results comparing four LLMs (Llama 3.3, Gemma 3, TxGemma, and Qwen 3) and three standard annotation tools (scCATCH, scType, and SingleR) using four single-cell datasets. For each cell group in each dataset, we use five different metrics to rank the genes, selecting the top 10, 15, and 20 genes as marker genes. These marker genes are the input of the LLMs to identify the potential cell type of each cell group. For standard tools (scCATCH, scType, and SingleR), we follow the authors' published instructions to infer the cell type name of each cell group. After obtaining the results, we use Gemini 2.5 Pro to assess the accuracy of cell type inference. A score of 1 was assigned if the predicted cell type was an exact match, a subtype, or a descendant of the ground-truth cell type; otherwise, a score of 0 was assigned. A method's final score for a dataset is the average score across all cell types. Overall, all four LLMs greatly outperform the standard tools by a significant margin, irrespective of the ranking methods or the number of differentially expressed (DE) genes chosen. Llama 3.3, which has the highest number of parameters among the four LLMs, has the highest score. Specifically, the average scores of Llama 3.3, Gemma 3, TxGemma, and Qwen 3 are 0.776, 0.709, 0.731, and 0.708, respectively. These LLM scores are substantially higher than the average scores of scCATCH (0.375), scType (0.425), and SingleR (0.425).

highest-scoring or most correlated cell type as the prediction for the respective group.

To demonstrate that LLMs do not overfit, we apply five different metrics to choose marker genes. Denoting $\log_2 FC$ as the log-fold-change, $\log_{10} p$ as the log10 *P*-value, $\log_{10} p_{adj}$ as the log10 adjusted *P*-value, and *pct* as the percentage of cells expressing the gene, the metrics are as follows:

- Significance-weighted log-fold-change (*sig_weighted_logfc*): The first metric is the log2 fold-change weighted by statistical

significance (*P*-value), which is calculated as $sig\_weighted\_logfc = -\log_{10}p \cdot \log_2 FC$. This approach ranks genes highly only when they show both significant expression changes and strong statistical support.

- Expression percentage difference (*exp_pct_diff*): This metric is the difference in the percentage of cells expressing the gene between the underlying group and other groups. This metric assigns higher ranks to genes that are expressed in a larger proportion of cells in the underlying group (compared to other groups).
- Percentage-weighted log-fold-change (*pct_weighted_logfc*): The third metric is the log2 fold-change weighted by the absolute expression percentage difference, which is calculated as $pct\_weighted\_logfc = |\exp\_pct\_diff| \cdot \log_2 FC$. This metric assigns higher ranks to genes that have both higher log-fold-change and more cells expressing the gene.
- Percentage-weighted significance (*pct_weighted_sig*): The fourth metric is the minus log10 adjusted *P*-value, weighted by expression percentage difference. In other words, $pct\_weighted\_sig = -\log_{10}p_{adj} \cdot \text{sign}(\log_2 FC) \cdot |\exp\_pct\_diff|$. In this formula, we multiply the weighted significance score by the sign of log-fold-change to preserve direction. Genes are ranked highly when they show both strong statistical significance and significant differences in detection frequency between groups.
- Mixture of log-fold-change and statistical significance (*mixed_logfc_sig*): The fifth metric is a weighted average between log-fold-change and minus log10 *P*-value. The metric is calculated as: $mixed\_logfc\_sig = 0.7 \cdot \log_2 FC - 0.3 \cdot \text{sign}(\log_2 FC) \cdot \log_{10}p$. High-ranking genes show both high log-fold-changes and significant *P*-values.

We sort the genes according to each metric and then choose the top 10, 15, and 20 DE genes as markers, resulting in 15 sets of markers for each cell group in each dataset. These markers serve as the input of the LLMs to predict cell types.

In total, we analyze four datasets using 63 approaches: 60 LLM-based (4 LLMs × 5 DE metrics × 3 cutoffs) and 3 standard annotation tools. After obtaining the results, we use Gemini 2.5 Pro (Comanici *et al.* 2025) to compare the predicted cell types against the ground truth to assess the performance of each approach. We assign a score of 1 when the predicted cell type is an exact match, a subtype, or a descendant of the ground-truth cell type. Otherwise, we assign a score of 0 to the predicted cell type. The score of a method for a dataset is the average score of its predicted cell types.

Figure 2 shows the benchmarking results of all 63 approaches. Overall, all LLMs greatly outperform standard tools by having higher scores by large margins, regardless of ranking metrics and cutoff thresholds. As shown in the column DE Metric Average, the scores of Llama 3.3 across the five metrics (column DE Metric Average) are in the range of $[0.733, 0.817]$. Gemma 3, TxGEmma, and Qwen 3, being smaller LLMs than Llama 3.3, also achieved outstanding results with scores in the range of $[0.673, 0.744]$, $[0.688, 0.754]$, and $[0.671, 0.731]$, respectively. The average scores of Llama 3.3, Gemma 3, TxGemma, and Qwen 3 are 0.776, 0.709, 0.731, and 0.708, respectively. These LLM scores are substantially higher than the average scores of scCATCH (0.375), scType (0.425), and SingleR (0.425). Llama 3.3, which has the highest number of parameters among the four LLMs, has the highest score. Llama 3.3 with the *pct_weighted_sig* metric achieves an outstanding score of 0.817 across all DE cutoffs and datasets.

4 Conclusion

We introduce a novel LLM-based cell type annotation workflow within CytoAnalyst, a unified platform designed for comprehensive scRNA-Seq analysis. Our approach seamlessly integrates analytical methods with advanced LLMs to accelerate and enhance cell type annotation. The integration of Meta's Llama 3.3 with embedding, clustering, and differential analysis methods for automated cell type inference marks a major leap in single-cell data analysis. The added graphical interface and interactive visualization system make sophisticated annotation tools accessible to researchers without a computational background, greatly reducing the burden of manual annotation for life scientists and bioinformaticians alike. The analysis results show promise for the application of advanced AI techniques in biomedical data analysis across diverse tissues and experimental conditions. Despite their promise, LLMs are prone to limitations such as overfitting and hallucination. One future direction to address this issue is Retrieval-Augmented Generation (RAG), where LLMs retrieve information from external, authoritative databases before generating a response. For future work, we will combine LLMs with systems-level analysis (Nguyen *et al.* 2021a, 2024b) and deep learning models (Tran *et al.* 2021, Nguyen *et al.* 2021b, 2024a) for practical applications in biomedical research. We continue to maintain and update the platform with the latest technologies.

Supplementary material

Supplementary material is available at *Bioinformatics Advances* online.

Conflicts of interest

None declared.

Funding

This work was partially supported by National Science Foundation [2343019, 2203236], National Institute of General Medical Sciences [R44GM152152], and National Cancer Institute [U01CA274573]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

Data availability

The data underlying this article are available in the CellxGene data portal at https://cellxgene.cziscience.com/. The datasets were derived from sources in the public domain: Frenz-Wiessner *et al.* 2024 (https://cellxgene.cziscience.com/collections/59cd85c5-3b22-4035-b628-2a20810ad54b), Wu *et al.* 2021 (https://cellxgene.

cziscience.com/collections/dea97145-f712-431c-a223-6b5f565f362a), Andrews *et al.* 2024 (https://cellxgene.cziscience.com/collections/0c8a364b-97b5-4cc8-a593-23c38c6f0ac5), and Solé-Boldo *et al.* 2020 (https://cellxgene.cziscience.com/collections/c353707f-09a4-4f12-92a0-cb741e57e5f0).

References

Andrews TS, Nakib D, Perciani CT *et al.* Single-cell, single-nucleus, and spatial transcriptomics characterization of the immunological landscape in the healthy and PSC human liver. *J Hepatol* 2024;**80**:730–43.

Aran D, Looney AP, Liu L *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;**20**:163–72.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B (Methodol)* 1995;**57**:289–300.

Blondel VD, Guillaume J-L, Lambiotte R *et al.* Fast unfolding of communities in large networks. *J Stat Mech* 2008; **2008**:P10008.

Boehm KM, El Nahhas OSM, Marra A *et al.* Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. *Nat Commun* 2025;**16**:2106.

Butler A, Hoffman P, Smibert P *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20.

Bya P, Tran D, Nguyen K *et al.* CytoAnalyst web platform facilitates comprehensive single cell RNA sequencing analysis. *Sci Rep* 2025;**15**:28736.

Cheng C, Chen W, Jin H *et al.* Review of single-cell RNA-seq annotation, integration, and cell–cell communication. *Cells* 2023; **12**:1970.

Clarke ZA, Andrews TS, Atif J *et al.* Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat Protoc* 2021;**16**:2749–64.

Comanici G, Bieber E, Schaekerm M *et al.* Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv, arXiv:2507.06261, 2025, preprint: not peer reviewed.

Finak G, McDavid A, Yajima M *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 2015;**16**:278.

Frenz-Wiessner S, Fairley SD, Buser M *et al.* Generation of complex bone marrow organoids from human induced pluripotent stem cells. *Nat Methods* 2024;**21**:868–81.

Grattafiori A, Dubey A, Jauhri A *et al.* The llama 3 herd of models. arXiv, arXiv:2407.21783, 2024, preprint: not peer reviewed.

Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–87.e29.

Heumos L, Schaar AC, Lance C *et al.*; Single-cell Best Practices Consortium. Best practices for single-cell analysis across modalities. *Nat Rev Genet* 2023;**24**:550–72.

Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods* 2024;**21**:1462–5.

Ianevski A, Giri AK, Aittokallio T *et al.* Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* 2022; **13**:1246.

Ji X, Tsao D, Bai K *et al.* Scannotate: an automated cell-type annotation tool for single-cell RNA-sequencing data. *Bioinform Adv* 2023;**3**:vbad030.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.

Kodinariya T, Makwana P. Review on determining number of cluster in K-Means clustering. *Int J Adv Res Comput Sci Management Stud* 2013;**1**:01.

Love MI, Huber W, Anders S *et al.* Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.

Lu MY, Chen B, Williamson DFK *et al.* A visual-language foundation model for computational pathology. *Nat Med* 2024;**30**:863–74.

Nguyen H, Tran D, Galazka JM *et al.* CPA: a web-based platform for consensus pathway analysis and interactive visualization. *Nucleic Acids Res* 2021a;**49**:W114–24.

Nguyen H, Tran D, Tran B *et al.* A comprehensive survey of regulatory network inference methods using single-cell RNA sequencing data. *Brief Bioinform* 2021b;**22**:1–15.

Nguyen H, Nguyen H, Tran D *et al.* Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Res* 2024a; **52**:4761–83.

Nguyen H, Pham V-D, Nguyen H *et al.* CCPA: cloud-based, self-learning modules for consensus pathway analysis using GO, KEGG and reactome. *Brief Bioinf* 2024b;**25**:bbae222.

Pasquini G, Rojo Arias JE, Schäfer P *et al.* Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J* 2021;**19**:961–9.

Pliner HA, Shendure J, Trapnell C *et al.* Supervised classification enables rapid annotation of cell atlases. *Nat Methods* 2019; **16**:983–6.

Quan F, Liang X, Cheng M *et al.* Annotation of cell types (ACT): a convenient web server for cell type annotation. *Genome Med* 2023;**15**:91.

Shao X, Liao J, Lu X *et al.* scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data. *iScience* 2020;**23**:100882.

Solé-Boldo L, Raddatz G, Schütz S *et al.* Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. *Commun Biol* 2020;**3**:188.

Song D, Chen S, Lee C *et al.* Synthetic control removes spurious discoveries from double dipping in single-cell and spatial transcriptomics data analyses. In: *International Conference on Research in Computational Molecular Biology*. Switzerland, Cham: Springer Nature, 2025, 400–4.

Stuart T, Butler A, Hoffman P *et al.* Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–902.e21.

Team G, Kamath A, Ferret J *et al.* Gemma 3 Technical Report. arXiv, arXiv:2503.19786, 2025, preprint: not peer reviewed.

Traag VA, Waltman L, van Eck NJ *et al.* From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019; **9**:5233–12.

Tran D, Nguyen H, Tran B *et al.* Fast and precise single-cell data analysis using hierarchical autoencoder. *Nat Commun* 2021; **12**:1029.

Wang E, Schmidgall S, Jaeger PF *et al.* Txgemma: Efficient and Agentic LLMs for Therapeutics. arXiv, arXiv:2504.06196, 2025, preprint: not peer reviewed.

Wilcoxon F, Katti SK, Wilcox RA. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. *Selected Tables Math Stat* 1970;**1**:171–259.

Wu SZ, Al-Eryani G, Roden DL *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* 2021; **53**:1334–47.

Yang A, Li A, Yang B *et al.* Qwen3 Technical Report. arXiv, arXiv:2505.09388, 2025, preprint: not peer reviewed.

Bioinformatics Advances, 2026, 6, 1–8
https://doi.org/10.1093/bioadv/vbag001
Application Note