OXFORD

Briefings in Bioinformatics, 2025, **26(6)**, bbaf622
https://doi.org/10.1093/bib/bbaf622
Problem Solving Protocol

Single-cell omics arena: evaluation of large language models for automatic cell-type annotations on single-cell omics data via RNA-seq bridging

Junhao Liu [1], Siwei Xu [1], Yongxian Wu [2], Jing Zhang [1,*]

[1]Department of Computer Science, University of California, Irvine, 6210 Donald Bren Hall, Irvine, CA 92697, United States
[2]Department of Chemical and Biomolecular Engineering, University of California, Irvine, 5200 Engineering Hall, Irvine, CA 92697, United States

*Corresponding author. E-mail: zhang.jing@uci.edu

Abstract

The single-cell sequencing revolution enables simultaneous molecular profiling of various modalities across thousands of individual cells, allowing scientists to investigate the diverse functions of complex tissues. Among all the analysis steps, assigning individual cells to specific types is fundamental for understanding cellular heterogeneity. However, this process is labor-intensive and requires extensive expert knowledge. Recent advances in large language models (LLMs) have demonstrated their ability to automatically extract biological knowledge, such as marker genes, promoting efficient, and automated cell-type annotations. To evaluate the capability of modern LLMs in automating the cell-type identification process, we first introduce an automated cell-type annotation method with comprehensive benchmark: Single-cell Omics Arena). Specifically, we began by compiling 11 publicly available single-cell RNA sequencing (scRNA-seq) datasets and evaluating eight LLMs across 1226 cell-type annotation-related tasks. This effort established a foundation for automated cell-type annotation from scRNA-seq data using interpretable features such as gene names. Building upon this benchmark, we introduced domain-specific chain-of-thought prompting techniques to enhance the accuracy of cell-type annotation and facilitate the extraction of relevant biological insights. Finally, to accommodate non-interpretable features, we proposed to leverage a pretrained VAE-based cross-modality translation module to convert features such as epigenetic marks into interpretable representations, which enables the seamless extension of LLM-based cell-type annotation to non-RNA-based sequencing technologies. In summary, our benchmark provides key insights into automated cell-type annotation from scRNA-seq data and demonstrates the potential of cross-modality translation for handling non-interpretable features.

Keywords: large language model; cell-type annotation; natural language processing for bioinformatics

Introduction

The recent advancements in single-cell technologies [1–5] enable simultaneous molecular profiling of diverse modalities across tens of thousands of individual cells, allowing researchers to explore the heterogeneity and functionality within complex tissues by uncovering rare or previously unidentified cell types that would otherwise be obscured by traditional bulk tissue sequencing methods. Among the various tasks of single-cell analyses, the classification of cells into canonical or novel cell types—referred to as cell-type annotation—serves as the primary and most fundamental step [6]. This is crucial because each cell type performs distinct roles, and their accurate identification facilitates the study of their specific contributions to biological processes, development, and disease mechanisms [7]. Unfortunately, this task is usually computationally demanding, labor-intensive, and requires extensive manual labeling, as traditional methods rely heavily on expert knowledge of gene functions and cell biology to ensure annotation accuracy. Consequently, there is a pressing need to develop efficient and precise cell-type annotation methods to automate and streamline this process.

Over the past decades, there have been significant transformations in the acquisition and utilization of domain-specific knowledge required for cell-type annotation, largely driven by advancements in artificial intelligence (AI) and natural language processing (NLP) [8]. In particular, large language models (LLMs) have emerged as powerful tools to efficiently process and synthesize extensive text corpora, including scientific literature, expert discussions, and technical documents, to accurately associate key features, such as marker genes, with specific cell types [9–13]. This has made automated cell annotation increasingly feasible with minimal expert involvement. Moreover, LLMs are capable of integrating diverse and complex data types, including genomic datasets, biological knowledge, and previous annotations, to further enhance the accuracy and efficiency of cell-type classification [8]. Consequently, several recent pioneering studies [8, 14–16] started to develop automated cell-type annotation methods

Received: May 17, 2025. **Revised:** September 3, 2025. **Accepted:** October 27, 2025

based on LLM for single-cell RNA sequencing (scRNA-seq) data, which already demonstrated strong concordance with traditional manual annotations [17–19].

Despite its promise, the automation of cell-type annotation from single-cell omics data using LLMs remains in its early stages and faces significant challenges. First, although LLMs have been extensively benchmarked across various domains [20, 21], no comprehensive evaluation of their performance on the specific task of cell-type labeling has been conducted. Benchmarking LLMs on existing datasets could provide valuable insights into their effectiveness. Second, current methodologies [8, 14, 15] predominantly rely on single-step prompt learning without incorporating intermediate reasoning processes [22], which limits their capacity to capture a deeper understanding of gene function and cell biology during the cell-type annotation process. Finally, unlike the immediately interpretable features of scRNA-seq data, such as gene names [3], many single-cell sequencing technologies generate molecular measurements that are not directly accessible to LLMs. For instance, single-cell ATAC-seq (scATAC-seq) assays [23] profile open chromatin regions, while single-cell Hi-C (scHi-C) experiments [24] capture chromatin contact probabilities. These features are rarely represented in the textual corpora used to train LLMs, making it highly challenging for LLMs to directly perform cell-type annotation.

To address these critical challenges, we introduce Single-cell Omics Arena (***SOAR***), an LLM-based framework for automatic cell-type annotation from single-cell omics data. Our contributions are three-fold, each directly aligned with the key limitations identified above. **(1)** To tackle the absence of systematic evaluation of LLMs in this domain, we compiled 11 publicly available datasets and assessed the performance of eight instruction-tuned LLMs across 1226 cell-type annotation tasks. This benchmarking effort establishes a comprehensive foundation for evaluating LLMs' ability to annotate cell types from interpretable features such as gene names in scRNA-seq. **(2)** To overcome the limitations of single-step prompting, we introduced domain-specific chain-of-thought (CoT) prompting strategies that enable LLMs to engage in intermediate reasoning. This improves their capacity to capture biological insights and enhances annotation accuracy. **(3)** To bridge the gap between LLMs and non-interpretable modalities, we further leveraged a pretrained cross-modality translation module that converts uninterpretable features (e.g. epigenetic marks and chromatin structures) into interpretable ones (e.g. gene names), allowing seamless extension of LLM-based cell-type annotation to non-RNA-based single-cell sequencing technologies.

Related work

The LLM field has made substantial progress in recent years. Models such as GPT [9, 25], Qwen [10, 26], and the Llama series [11, 27, 28] have already demonstrated remarkable performance across a wide range of tasks, including question answering [29], image captioning [30], and text summarization [31]. Training an LLM to follow human instructions typically involves two key phases: pretraining on a large text corpus and subsequent instruction tuning [32]. Once tuned, various evaluation frameworks have been developed to measure the models' adherence to human instructions. For example, MMLU [20] is widely used to assess a model's general knowledge and reasoning abilities across diverse subjects, while GSM8K [21] serves as a benchmark for mathematical problem-solving. Despite these advancements, much of the training and evaluation of LLMs continues to focus on NLP, limiting their potential to address more complex scientific challenges. Single-cell genomics is a groundbreaking technique that enables researchers to quantify molecular features from individual cells, facilitating the study of cellular heterogeneity and functionality within complex tissues. However, applying LLMs to automate the cell-type labeling process remains challenging, primarily due to the uncertainty surrounding LLMs' ability to accurately interpret the specialized domain knowledge inherent in these datasets.

Several pioneering works have explored the application of pretrained models to genomic datasets, such as Geneformer [14] and scGPT [15]. However, these studies focus on training bidirectional contextual genomics embedding models rather than developing general instruction-following LLMs capable of reasoning through genomics analysis tasks based on human commands. Consequently, these models require additional post-finetuning to perform specific downstream tasks, aligning with the pretrain-then-finetune paradigm introduced by BERT [33], rather than leveraging general-purpose instruction-tuned LLMs in a zero-shot setting without supervision from downstream tasks.

Recently, Cell2Sentence [8] introduced a post-pretraining strategy that integrates natural language-based LLMs with transcriptomic knowledge, enabling LLMs to follow human instructions to complete various genomics analysis tasks. Meanwhile, Hou [16] curated an scRNA-seq dataset to evaluate the cell-type annotation capabilities of GPT models. However, their experiments relied on manual assessments of annotation quality using a coarse scoring rubric that categorized matches as full, partial, or absent, thereby introducing human bias. Moreover, their evaluations were limited to the zero-shot and in-context annotation capabilities of GPT-4 and GPT-3.5 for scRNA-seq data, without further investigation into the general cell-type annotation performance of current open-source LLMs or strategies for enhancing their reasoning abilities on cell-type annotation.

Consequently, due to the lack of throughout standardized genomics-related benchmarks, several key questions remain unresolved: **(1)** What is the ability of contemporary instruction-tuned LLMs to analyze genomic data without additional fine-tuning? **(2)** If these LLMs can analyze genomic data, how can their reasoning performance in genomics analysis be further enhanced? and **(3)** Can LLMs handle heterogeneous genomic data with uninterpretable features beyond text (e.g. epigenetic features)?

To address the aforementioned questions, we first introduce a single-cell cell-type annotation benchmark, ***SOAR***, which spans scRNA-seq and other omic data, 11 publicly available datasets, five widely adopted automatic quantitative metrics from the field of NLP, and eight instruction-tuned LLMs. This benchmark is designed to evaluate the instruction-following capabilities of LLMs on cell-type annotations. Based on our results, we found that many LLMs demonstrate a robust ability to interpret single-cell genomics data, particularly when leveraging the proposed domain-specific CoT prompting method. Furthermore, to extend the capability of LLMs to analyze genomic features that are not in the text domain, we propose a cross-modality instruction method using RNA-seq as a bridge, which is shown to be effective through extensive experiments.

Method

We begin by introducing the foundational concepts of cell-type annotation using scRNA-seq. We then propose two distinct

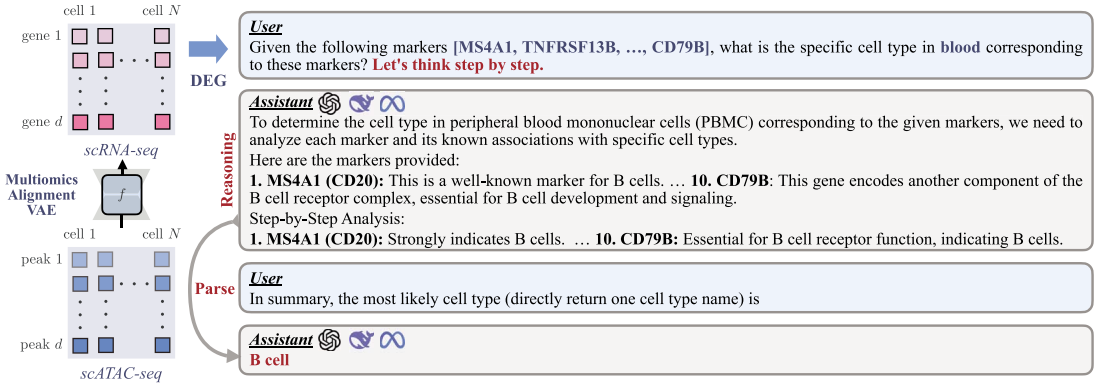

Figure 1. An illustration of leveraging LLMs for automated cell-type annotation by reasoning over marker genes.

strategies to guide LLMs in performing cell-type annotation for single-cell analysis through RNA-seq bridging. An overview of the proposed method is presented in Fig. 1.

Preliminary of cell-type annotation

scRNA-seq experiments provide a comprehensive view of cellular functionality by quantifying the number of transcribed messenger RNA molecules, corresponding to gene expression levels. Let a list of unique sequenced gene names be denoted as $\mathcal{G} = [g_1, \ldots, g_{d_g}]$. Consequently, a sequenced scRNA-seq sample can be represented as a continuous vector $\boldsymbol{x} \in \mathbb{R}^{d_g}$, where $\boldsymbol{x}_i$ denotes the observed expression value of the ith gene in the gene list $\mathcal{G}$. Based on the scRNA-seq data of a cell, denoted by $\boldsymbol{x}$, biologists can annotate its corresponding cell type, $\boldsymbol{y}$, by carefully analyzing gene expression patterns using domain expertise or by referencing established findings in the literature.

Cell-type annotation is a critical task in scRNA-seq analysis that typically requires significant effort and specialized domain knowledge. This raises a compelling question: "Can contemporary LLMs assist in analyzing single-cell genomics data?" However, directly applying LLMs to this task poses significant challenges for several reasons. First, most state-of-the-art LLMs are trained exclusively on natural language, which raises concerns about whether these models can achieve the same level of comprehension and reasoning when dealing with domain-specific content, such as gene names and biological terminology. Second, sequences generated by many single-cell sequencing technologies are not always easily represented in conventional text formats suitable for LLM input. This limitation constrains the models' ability to reason effectively about various biological signals.

Prompting strategies for LLM-driven cell-type annotation

To leverage LLMs for the task of cell-type annotation, we formulate it as a standard question-answering problem. Specifically, we define the question $\boldsymbol{q}$ using the following template: "Given the following markers {X}, what is the specific cell type in {C} corresponding to these markers?" Here, {·} represents a placeholder, with X denoting a gene expression profile description derived from an scRNA-seq vector $\boldsymbol{x}$, and C representing any associated metadata for the cell (e.g. tissue type). To guide the LLMs toward generating an appropriate response for annotating cell types, we append a trigger sentence $\boldsymbol{t}$ after the question $\boldsymbol{q}$. Consequently, the LLMs are tasked with generating a response sentence $\boldsymbol{r}$ as follows:

$$\boldsymbol{r} = \text{LLM}(\boldsymbol{q}, \boldsymbol{t}). \tag{1}$$

Gene expressions from scRNA-seq

As previously mentioned, the scRNA-seq vector $\boldsymbol{x}$ is structured data, where each dimension represents the expression value of a specific gene. To enable LLMs to process these structured data, a serialization method must be employed. In our experiments, we used differentially expressed gene (DEG) analysis to select the k most distinguishable genes, $\{\hat{g}_1, \ldots, \hat{g}_k\}$, and then listed the selected k gene names in decreasing order of their P-values, which can be expressed as follows:

$$\text{X} = [\hat{g}_1, \ldots, \hat{g}_k],\ \hat{g}_i \in \text{DEG}(\mathcal{G}, \boldsymbol{x}), \tag{2}$$

where $\hat{g}_i$ represents the gene name of the ith selected DEG, and X is the formulated gene expression profile in text format to be provided as input to the LLMs.

Zero-shot prompting for LLM-driven cell-type annotation

To guide LLMs in annotating cell types based on the provided gene expression profile, we use a zero-shot prompting trigger sentence, $\boldsymbol{t}_{zero}$, which is set as "The most likely cell type (directly return one cell type name) is." Given a cell $\boldsymbol{x}$ to be annotated and its gene expression profile X, the response $\boldsymbol{r}$ generated by the LLMs using (1) is parsed to extract the predicted cell type, denoted as $\hat{\boldsymbol{y}}$.

Domain-specific zero-shot chain-of-thought prompting for LLM-driven cell-type annotation

Cell-type annotation requires reasoning skills to analyze complex co-expression patterns of different genes, which presents a significant challenge, even for expert biologists. Inspired by the CoT method proposed by [22], we adopted a two-stage strategy to enhance the reasoning capabilities of LLMs for cell-type annotation. First, a CoT trigger sentence, $\boldsymbol{t}_{\text{cot}}$, is set as "Let us think step by step." This prompts the LLMs to sequentially reason through the genes listed in X one by one

$$\boldsymbol{z} = \text{LLM}(\boldsymbol{q}, \boldsymbol{t}_{\text{cot}}), \tag{3}$$

where $\boldsymbol{z}$ is the generated response from the first stage. In the second stage, the response $\boldsymbol{z}$, along with the initial prompt, is used

to prompt the LLMs to summarize the final annotation using the trigger sentence $\mathbf{t}_{sum}$ as follows:

$$\boldsymbol{r} = \text{LLM}(\boldsymbol{q}, \mathbf{t}_{cot}, \boldsymbol{z}, \mathbf{t}_{sum}), \tag{4}$$

where $\mathbf{t}_{sum}$ is set as "In summary, the most likely cell type (please return one cell type name) is." The final response is then parsed as the predicted cell-type annotation $\hat{\mathbf{y}}$.

Cell-type annotation on uninterpretable features from other modalities

Many recent sequencing technologies provide complementary information beyond gene expression, facilitating the profiling of molecular features within cells across various biological layers. However, the signals generated by these technologies are often not easily represented as text, limiting the capacity of LLMs to interpret such data in the same manner as they process scRNA-seq data. For instance, ATAC-seq is a widely used technique for measuring DNA accessibility [5] in individual cells. This method produces a sparse, binary matrix where a value of 1 indicates an accessible region and 0 indicates an inaccessible region, which were rarely captured in the text corpora used to train LLMs. In this study, we use ATAC-seq as an example to demonstrate how LLMs can be applied to address the challenge of multi-omics cell-type annotation.

Multiomics translation module

To address this challenge, we propose a multiomics translation method that enables LLMs to reason across different multiomics data. Specifically, given the strong NLP capability of LLMs, we select RNA-seq as the pivot modality for mapping various types of multiomics data. Consequently, the data from different modalities are aligned with RNA-seq using a multimodal translation module f to integrate data into a shared latent space. In our experiments, we employ the variational autoencoder-based modality integration model scACT proposed in [34] as the multimodal alignment module f to align RNA-seq and ATAC-seq data. In particular, utilizing the integration space learned by the two autoencoders for each modality, the translation $f : \mathbb{R}^{d'} \rightarrow \mathbb{R}^{d}$ is constructed by combining the encoder of ATAC-seq with the decoder of RNA-seq. This mapping translates the ATAC-seq modality $\boldsymbol{x}' \in \mathbb{R}^{d'}$ into the RNA-seq modality $\boldsymbol{x}$, where $\boldsymbol{x} = f(\boldsymbol{x}')$.

Evaluation benchmarks of SOAR-RNA and SOAR-MultiOmics

To fairly evaluate the ability of LLMs in cell-type annotation across both scRNA-seq and non-interpretable features from other modalities, in this section, we describe the proposed **SOAR-RNA** and ***SOAR-MultiOmics*** datasets, which are specifically designed to evaluate the cell-type annotation capabilities of LLMs across various modalities of single-cell data.

SOAR-RNA: cell-type annotation benchmark on scRNA-seq data

SOAR-RNA is a benchmark dataset designed to evaluate the reasoning capabilities of LLMs in the field of single-cell genomics. Following previous work [16], we curated cell-type annotation samples from the following scRNA-seq datasets: Azimuth [35], Human Cell Landscape (HCL) [36], Mouse Cell Atlas (MCA) [37], GTEx [7], B-cell lymphoma (BCL) [38], Literature [7], Colon Cancer [39], Lung Cancer [40], Tabula Sapiens (TS) [41], and Non-model Mammal [42]. A summary of all datasets utilized in this work is provided in Table A1.

Listing 1. An example of a cell-type annotation task in the proposed benchmark dataset. Each entry includes an identifier, the dataset subset (e.g. Azimuth), the tissue of origin (e.g. PBMC), a list of input marker genes, and a set of synonymous ground truth cell type labels.

```
{
  "id": 1,
  "subset": "Azimuth",
  "tissue": "PBMC",
  "gene list": [
        "MS4A1", "COCH", "AIM2",
        "BANK1", "SSPN", "CD79A",
        "TEX9", "RALGPS2",
        "TNFRSF13C", "LINC01781"
    ],
  "cell type": [
        "memory B cell",
        "memory B-lymphocyte",
        "memory B-cell",
        "memory B lymphocyte"
    ]
}
```

DEG Analysis

For each scRNA-seq, manually annotated cell types $\mathbf{y}$ and gene expression matrices $\boldsymbol{x}$ were obtained directly from the corresponding publications. For DEG analysis, raw gene expression counts were first log-transformed to the total sum of the maximum gene counts after adding a pseudocount of 1, using Scanpy [43]. Welch's t-test was then performed to identify DEGs by comparing each cell type against all others. Genes for each cell type were ranked in ascending order based on P-values, with ties in pP-values further ranked in descending order by t-statistics. The top 10 DEGs were selected to construct the gene list $\mathbf{x}$.

Cell-type normalization

To account for synonymy in cell-type descriptions, we normalized cell-type annotations to their unambiguous cell ontology (CL) names [44], using the API (https://bioportal.bioontology.org/) provided by [45]. All synonyms for a cell-type name returned by the API were included as ground truth annotation candidates $\mathbf{y}$.

After data preprocessing, SOAR-RNA contains **1191** cell types spanning **37** different tissues. Detailed statistics on tissue distribution, along with the complete tissue list, are illustrated in Figure A1. After preprocessing, all samples are represented in JSON format. Listing 1 provides an example of a cell-type annotation task in the proposed dataset. In addition, details about the preprocessing are provided in **Appendix: Benchmark Details**.

SOAR-MultiOmics: cell-type annotation benchmark on multiomics data

SOAR-MultiOmics is a benchmark dataset designed to evaluate the reasoning capabilities of LLMs on single-cell multiomics data. For this dataset, we included publicly available single-cell multiomics data, specifically the human peripheral blood mononuclear cells (PBMC) 10k dataset from 10X Genomics, as well as the prefrontal cortex (PFC) brain dataset [46], both of which feature parallel scRNA-seq and scATAC-seq sequencing. The preprocessing details are provided in Appendix A.3. The DEG analysis and CL name normalization were conducted following the same procedures as in SOAR-RNA.

After the data preprocessing stage, SOAR-MultiOmics contains 28 cell types for the human PBMC 10k dataset, and seven major cell types (excitatory neurons, inhibitory neurons, astrocytes, endothelial cells, microglia, oligodendrocytes, and OPCs) for the PFC brain dataset. In addition, details about the preprocessing are provided in **Appendix: Benchmark Details**.

Experiment results

Baseline methods

In this work, we conducted detailed evaluations of single-cell cell-type annotation ability across both open-source and close-source instruction-tuned LLMs, such as Qwen2 [10] Llama-3 [11], Mixtral [13], DeepSeek [12], and GPT-4o [9]. To further compare the reasoning ability of general instructed and domain-specific LLMs, we also include Cell2Sentence [8] in our benchmark. A brief summary of all LLMs utilized in this work is provided in Table 2. Examples of using LLMs for cell-type annotation can be found in Table 5. To better understand the performance gap between LLM-based methods and traditional non-LLM-based methods in cell-type annotation from scRNA-seq data, we incorporate traditional annotation methods, including SingleR [18], ScType [19], and CellMarker2.0 [17].

Implementation details for LLM-driven cell-type annotation

Our code and benchmark datasets are provided in the Supplementary Materials. All the experiments were conducted on NVIDIA A6000 GPUs and completed after ∼5 days. For our experiments, we used a consistent set of hyperparameters across all instruction-tuned models, adhering to commonly accepted values from previous studies [47]. Specifically, we set the temperature at 0.6, top_p at 0.9, and top_k at 50. The temperature parameter adjusts the randomness during sampling, top_p filters out tokens with lower probabilities, and top_k limits the sampling process to the k most probable tokens. For Cell2Sentence, we followed the instruction finetuning method proposed in [8] to train the pretrained Cell2Sentence model using the instruction-following dataset provided by the authors.

Notation and prompts for cell-type annotation

Table 1 summarizes the prompts and notations used in the benchmarking process for cell-type annotation. The table includes definitions of key variables and their corresponding templates or descriptions. Table 5 illustrate the cell-type annotation process of using LLMs by zero-shot CoT prompting. This comprehensive overview provides a clear framework for understanding the cell-type annotation workflow.

Answer cleansing for cell-type annotation

Since the cell-type annotation is in free format, we extract the final answer by matching any character that is not a newline (\n), comma (,), or period (.), zero or more times. The pseudo code to implement this is `pred = re.match(r"^[^\n,\.]*", pred)`. The first matched result is selected as the cell-type annotation. The annotation is further normalized to its singular form for comparison (https://github.com/jaraco/inflect).

Examples of LLM-driven cell-type annotation

The reasoning processes employed by LLMs for cell-type annotation, using Mixtral-8×22B as an example, are illustrated in Table 5 and A4. Leveraging the zero-shot CoT approach, LLMs can analyze each marker in the provided list, generate a reasoning trajectory, and summarize the final predicted cell type.

Evaluation metrics and protocols

Cell-type annotation is provided in a free format without a strict answer template, distinguishing it from binary classification tasks. Moreover, cell-type annotation often involves multiple synonyms, necessitating the evaluation of predicted answers against several candidate labels. To ensure a fair assessment of free-format responses in cell-type annotation, we propose the use of four widely adopted metrics for evaluating generation quality in NLP. These include ROUGE [48], which measures similarity based on unigram (ROUGE-1) and bigram (ROUGE-2) overlaps as well as the longest common subsequence (ROUGE-L); METEOR [49], which computes the harmonic mean of precision and recall for unigram matches based on surface forms, stemmed forms, and meanings; and the BLEU score [50, 51], which quantifies n-gram overlap between the predicted text $\hat{\mathbf{y}}$ and the ground truth annotation $\mathbf{y}$. Specifically, we report BLEU-1, BLEU-2, and the geometric average BLEU score, as many cell-type annotations contain fewer than three words. To further evaluate the precision of cell-type annotations generated by LLMs, we also employ the Exact Match (EM) and F1 score, which are commonly used to assess response quality in question-answering tasks [52].

LLM-based cell-type annotation: effectiveness and generalization on SOAR-RNA

LLMs excel in automatic cell-type annotation using the proposed zero-shot chain-of-thought approach

We first evaluated the cell-type annotation capabilities of LLMs on our proposed SOAR-RNA. The major quantitative scores (ROUGE-L, METEOR, and BLEU) for SOAR-RNA are reported in Table 3, while the detailed evaluation metrics (ROUGE-1, ROUGE-2, BLEU-1, BLEU-2) for the SOAR-RNA benchmark is shown in Table A2. As shown in the table, we observed that general open-source LLMs, such as Llama-3-70B and Mixtral-8×22B, using the zero-shot prompting strategy, achieved performance comparable with non-LLM-based approaches (e.g. CellMarker2.0). Specifically, the ROUGE-L and METEOR scores for Llama-3-70B were 28.78 and 27.35, respectively, which are close to those of the best-performing non-LLM-based method, CellMarker2.0 (31.76 and 23.83). Furthermore, we observed that general open-source LLMs, such as Mixtral-8×22B, using the zero-shot prompting strategy, outperformed the domain-specific pretrained model (i.e. Cell2Sentence). Specifically, Mixtral-8×22B achieved ROUGE-L and METEOR scores of 39.60 and 35.67, respectively, compared with 26.76 and 19.45 for Cell2Sentence. The average BLEU score for Mixtral-8×22B was 16.65, while Cell2Sentence achieved a slightly higher score of 17.25. After applying zero-shot CoT prompting, all open-source LLMs and GPT-4o mini showed significant improvement in BLEU scores. For instance, the average BLEU score of Mixtral-8×22B using zero-shot CoT increased to 28.19, representing a relative improvement of over 69% compared with the zero-shot prompting strategy. This result demonstrates the effectiveness of the zero-shot CoT prompting method in enhancing LLM performance for cell-type annotations.

LLMs with the zero-shot CoT perform better than domain-specific language models and traditionalnon-LLM-based methods

Furthermore, we observed that the zero-shot CoT strategy enabled open-source LLMs trained solely on natural language to significantly outperform the domain-specific model and the

Table 1. The prompts and notations used in the benchmarking framework. The upper section defines various types of prompts used to query LLMs, including standard prompts (***q***) and three zero-shot prompting strategies: ***t****zero*, ***t****cot*, and ***t****sum*. The lower section describes the key notations for input data (***x***, x, c), model responses (***r***), and reference cell types (***y***, $\hat{\boldsymbol{y}}$), which are used to evaluate model performance in automated cell-type annotation tasks

Notation	Type	Template
q	Prompt	Given the following markers {x}, what is the specific cell type in {c} corresponding to these markers?
t*zero*	Zero-shot prompt	The most likely cell type (directly return one cell type name) is
t*cot*	Zero-shot CoT prompt	Let us think step by step.
t*sum*	Zero-shot CoT prompt	In summary, the most likely cell type (please return one cell type name) is
Notation	**Type**	**Definition**
x	Single-cell data	Raw single-cell data (e.g. scRNA-seq, scATAC-seq)
x	Gene Profile	Genes selected by DEG analysis from raw data (e.g. ["MS4A1," "COCH," "AIM2," "BANK1," "SSPN," "CD79A," "TEX9," "RALGPS2," "TNFRSF13C," "LINC01781"])
c	Meta data	Tissue type (e.g. PBMC)
r	Response	Response from LLMs with the prompt above (e.g. examples in Appendix 4.2.3)
y	Cell type	Cell-type annotation by humans
$\hat{\boldsymbol{y}}$	Cell type	Cell-type annotated by answer cleansing of the response ***r***

Table 2. The LLMs evaluated in our experiments, along with their parameter sizes and instruction formats. All models, except Cell2Sentence, are general-purpose LLMs prompted with natural language instructions. Cell2Sentence is a domain-specific model fine-tuned for cell-type annotation using cell-specific sentence prompts

Language model	Params	Instruction format
DeepSeek-LLM	67B	General text
Qwen2	72B	General text
Llama-3	70B	General text
Mixtral-8×7B	56B	General text
Mixtral-8×22B	176B	General text
Cell2Sentence	160M	Cell Sentence
GPT-4o mini	∼8B	General text
GPT-4o	∼200B	General text

traditional non-LLM-based methods in cell-type annotation tasks. For example, Mixtral-8×22B achieved over 64% relative improvement compared with Cell2Sentence. As shown in Fig. 2, open-source LLMs outperformed the domain-specific model in EM and F1 scores with zero-shot CoT prompting (Section 3.2). Notably, Qwen2-72B achieved a 35% relative improvement in F1 score (46.38 versus 27.01) compared with the Cell2Sentence. Due to the free-form nature of cell-type annotation and variations in synonyms, the EM metric remains overly restrictive, even after normalizing annotations to unambiguous CL names. When comparing LLMs utilizing zero-shot CoT prompting with traditional non-LLM-based methods, open-source LLMs demonstrated substantial improvements. Specifically, Mixtral-8×22B achieved an F1 score of 51.40, representing a 63% relative improvement over the best-performing non-LLM-based method, which attained an F1 score of 31.49. These observations highlight the reasoning capabilities of LLMs in analyzing complex gene co-expression patterns and their broad acquisition of domain-specific knowledge. The reasoning processes on cell-type annotation performed by LLMs is detailed in Table 5. This underscores the potential of LLMs trained on text domains to outperform domain-specific models pretrained from scratch in specialized tasks. Consequently, it motivates the research community to focus more on leveraging the domain-specific knowledge and reasoning abilities inherent in existing LLMs, which may offer superior problem-solving capabilities in specialized domains.

Nevertheless, it is important to note that there remains a significant performance gap between open-source LLMs and closed-source models (i.e. GPT-4o mini and GPT-4o) across all metrics as shown in Table 3 and Fig. 2, including BLEU score and EM/F1 performance. Additionally, we found that applying zero-shot CoT prompting slightly degraded the cell-type annotation performance of GPT-4o in F1 score (58.29 versus 57.36) as shown in Fig. 2. A potential explanation for this is that GPT-4o is already fine-tuned for CoT reasoning, even when users do not explicitly prompt it to do so [53]. Based on our observations, using zero-shot CoT may cause GPT-4o to generate broader cell type names rather than more specific sub-cell type names, which negatively impacts precision and recall scores. To verify this, we examined representative outputs and found that CoT prompting indeed leads to a wider range of more generic predictions, such as omitting subpopulation qualifiers. This suggests that the observed F1 score decrease may stem from a loss of specificity in the annotated labels.

LLMs generalize well across various tissue types

To assess the generalization ability of LLMs in cell-type annotation across tissues, we analyzed per-tissue annotation accuracy (Fig. 3). The open-source Mixtral-8×22B performs similarly to GPT-4o series across tissues like breast, mammary, and pancreas, likely due to correlated language corpora during pretraining. For well-studied tissues with extensive literature (e.g. PBMC, duodenum, and mammary glands), LLMs show stronger reasoning and accuracy as shown in Fig. 3. In contrast, the domain-specific model (i.e. Cell2Sentence) performs less consistently, highlighting the superior generalization ability of LLMs to infer complex gene co-expression patterns and leverage broad knowledge from training. EM/F1 results (**Appendix: Additional Evaluation Results**) confirm these findings.

Cross-omics alignment and LLM annotation on SOAR-MultiOmics

Cross-omics translation is accurate

To assess whether LLMs can effectively analyze multiomics data, we further evaluated cell-type annotation performance using the proposed SOAR-MultiOmics benchmark. Following the methodology described earlier, for RNA-seq data, we directly formatted the gene expression vector ***x*** into a profile description x. For ATAC-seq

Table 3. Quantitative evaluation of cell-type annotation on the SOAR-RNA benchmark. The table compares traditional non-LLM-based methods (CellMarker2.0, SingleR, and ScType) with LLM-based methods under two prompting strategies: standard zero-shot and zero-shot CoT. Metrics include ROUGE-L, METEOR, and overall BLEU score

	W/o CoT			Zero-Shot CoT		
Model	**ROUGE-L**	**METEOR**	**BLEU**	**ROUGE-L**	**METEOR**	**BLEU**
CellMarker2.0	31.76	23.83	27.23	–	–	–
SingleR	16.49	2.96	0.00	–	–	–
ScType	12.24	20.18	10.77	–	–	–
DeepSeek-LLM-67B	32.74	24.27	16.87	40.47	31.13	21.02
Qwen2-72B	32.05	29.96	11.13	46.34	37.09	25.69
Llama-3-70B	29.83	27.35	14.33	42.02	34.09	17.38
Mixtral-8×7B	28.78	28.72	10.23	42.08	35.45	20.90
Mixtral-8×22B	39.60	35.67	16.65	51.26	41.97	28.19
Cell2Sentence	26.76	19.45	17.25	–	–	–
GPT-4o mini	52.26	41.08	32.64	51.17	40.84	37.45
GPT-4o	58.12	45.39	51.79	57.34	45.36	42.15

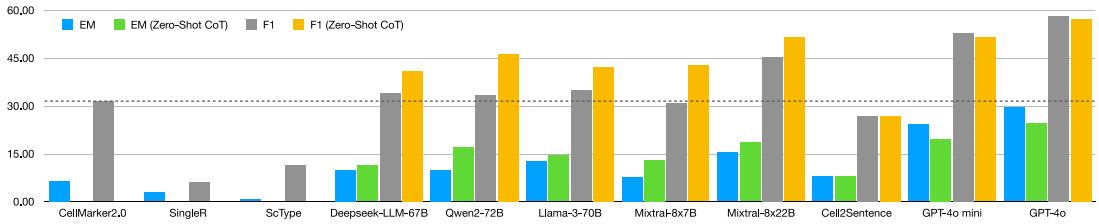

Figure 2. The EM and F1 evaluation results on the SOAR-RNA benchmark compare traditional non-LLM-based methods with LLM-based approaches, employing zero-shot and zero-shot CoT prompting strategies to guide the LLMs, respectively.

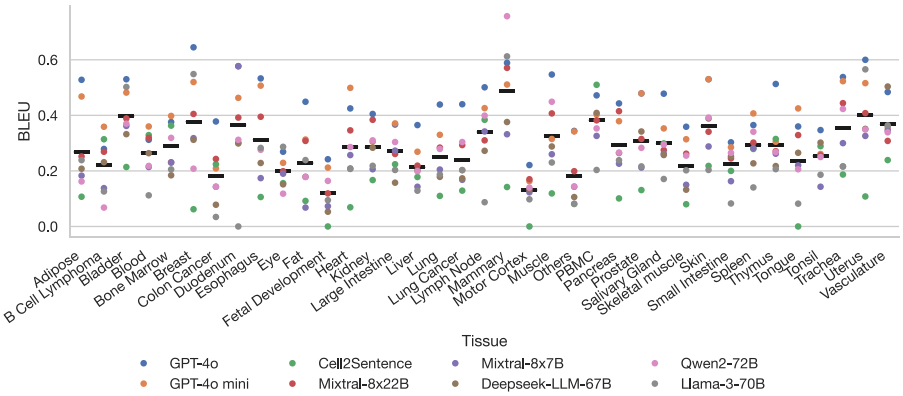

Figure 3. The BLEU evaluation results per tissue of the SOAR-RNA benchmark. Each dot represents the BLEU score of a specific LLM for a given tissue, with black bars indicating the mean BLEU score per tissue.

data, we propose to leverage a pretrained cross-modality alignment model, f, using a VAE architecture to align both RNA-seq and ATAC-seq data into a common semantic space. Consequently, the ATAC-seq vector $\mathbf{x}'$ was translated into the RNA-seq domain using $f(\mathbf{x}')$, and subsequently formatted into a profile description x. With this unified semantic space, we can bridge any omic data with RNA-seq, enabling multiomics data to be processed by LLMs through pure text inputs. In our experiments, we adopted the pretrained VAE from scACT [34] as the cross-modality alignment model f. To evaluate the translation accuracy of f, we assessed the ATAC-to-RNA translation performance on the human brain PFC dataset, as shown in Fig. A5. The marker gene, defined as the most expressively indicative gene for each cell type, is analogous to a category label in classification tasks. We found that f preserved expression patterns similar to the observed ground truth (mean $R^2 = 0.914$). We then focused on four key marker genes, SATB2 for excitatory cells, GAD2 for inhibitory cells, FLT1 for endothelial cells, and MOG for oligodendrocyte cells, as these are widely recognized canonical markers. The UMAP results for these genes were consistent with our previous findings, with the translated expression highlighting the corresponding cell types and showing high correlations with the observed expression ($R^2 = 0.633, 0.603, 0.460, 0.684$, Fig. A5B). These results demonstrate that the cross-modality alignment model f effectively translates ATAC-seq data into the RNA-seq domain.

Cross-omics translation effectively enables LLMs to analyze uninterpretable features

We summarize the multiomics cell-type annotation results in Table 4 and Fig. 4. As shown in the results, GPT-4o, GPT-4o mini, and Mixtral-8×22B achieve comparable annotation performance with the domain-specific model Cell2Sentence on both RNA-seq

Table 4. The quantitative evaluation results of cell-type annotation on the SOAR-MultiOmics benchmark using the zero-shot CoT prompting strategies to prompt LLMs

	RNA-seq			ATAC-seq		
Model	**ROUGE-L**	**METEOR**	**BLEU**	**ROUGE-L**	**METEOR**	**BLEU**
Qwen2-72B	20.57	11.54	10.51	18.43	11.80	7.80
Llama-3-70B	27.55	17.73	16.64	31.83	19.39	17.47
Mixtral-8×7B	33.65	26.11	24.71	29.36	24.21	15.02
Mixtral-8×22B	27.67	16.27	12.90	30.22	19.39	12.99
Cell2Sentence	28.42	20.63	38.20	26.45	17.29	28.58
GPT-4o mini	39.26	29.92	28.06	37.06	28.08	24.66
GPT-4o	41.37	30.20	33.93	38.31	25.23	28.68

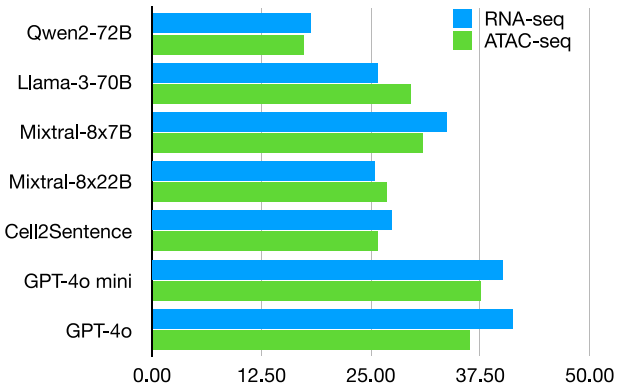

Figure 4. The F1 evaluation results on the SOAR-MultiOmics benchmark.

and ATAC-seq data. For example, GPT-4o achieves a METEOR score of 30.20, a BLEU score of 33.93, and an F1 score of 41.28, compared with Cell2Sentence's METEOR score of 20.63, BLEU score of 38.20, and F1 score of 27.41. This observation aligns with the tissue-level annotation results shown in Fig. 3, where the majority of samples in SOAR-MultiOmics come from PBMC tissue, which is well annotated by the Cell2Sentence model. As illustrated in Fig. 4, the comparable performance of cell-type annotation from ATAC-seq data demonstrates the ability of LLMs to analyze multiomics data using the multimodal alignment module f. Despite the lower BLEU scores observed for ATAC-seq compared with RNA-seq, we note that the ROUGE and METEOR scores remain relatively consistent across modalities. This can be attributed to the differences in how these metrics evaluate generated annotations. BLEU is precision-oriented and relies on exact n-gram overlaps, making it particularly sensitive to variations in surface form and phrasing. Since ATAC-seq profiles open chromatin regions that lack direct gene annotations, the resulting LLM outputs tend to be more diverse and less lexically aligned with the reference labels, thus leading to a drop in BLEU scores. In contrast, ROUGE and METEOR emphasize recall and semantic similarity, allowing them to capture the correctness of annotations even when the exact wording differs. The stability of these metrics indicates that LLMs can still produce semantically accurate cell-type labels from ATAC-seq inputs, further supporting the effectiveness of our cross-omics translation strategy in enabling LLMs to handle less interpretable input modalities. This finding highlights the potential for adapting LLMs to analyze a broader range of biological sequencing data.

Error analysis

We identified two major errors that prevent LLMs from accurately annotating cell types: (1) insufficient prior knowledge of existing literature, and (2) the inability of LLMs to predict exact subtypes, often resulting in the annotation of broader cell types. One major limitation of LLMs in cell-type annotation lies in their insufficient prior knowledge of existing literature, particularly for less-studied tissues. Our analysis of annotation accuracy across different tissues (Fig. 3) highlights this issue. While LLMs, such as the open-source Mixtral-8×22B and GPT-4o series, perform comparably well across tissues like breast, mammary, and pancreas—likely benefiting from correlated language corpora during pretraining—their performance is notably stronger in well-studied tissues with extensive literature, including PBMC, duodenum, breast, uterus, and mammary glands. Interestingly, the domain-specific model Cell2Sentence exhibited less consistent performance across tissues, reaffirming the potential of general-purpose LLMs to infer complex gene co-expression patterns and leverage broad knowledge acquired during training. These observations are further validated by detailed EM/F1 results presented in **Appendix: Additional Evaluation Results**, emphasizing the critical role of prior knowledge in enhancing LLM performance.

Another major error observed in LLM predictions is their tendency to assign broader cell-type annotations than those labeled by human experts. This discrepancy arises from the limitations in the representation of input data, and the free-form nature of cell-type annotation and variations in synonyms. These gene lists often fail to accurately capture the defining characteristics of specific cell types, leading to ambiguous or overly generalized predictions. For example, when the input gene set lacks sufficient specificity, the model may predict a higher level cell-type category rather than pinpointing a distinct subtype. This issue highlights a critical challenge in ensuring that input data effectively represent the unique molecular signatures required for precise annotations, which directly impacts the predictive accuracy of LLMs in cell-type identification tasks.

Cost analysis

We further evaluated the computational cost of LLM-based cell-type annotation by measuring both the average elapsed time per sample and the associated API costs (Table 6). The LLMs evaluated in this study include both locally deployed open-weight models and API-accessed proprietary models. For locally deployed models (DeepSeek-LLM, Qwen2, Llama-3, Mixtral-8×7B, Mixtral-8×22B, and Cell2Sentence), inference was performed on four NVIDIA A6000 GPUs (48 GB VRAM each) using FP16 precision and the HuggingFace Transformers library. These models incur no API cost, and their inference time is largely determined by model parameter size. In contrast, GPT-4o mini and GPT-4o were accessed via the official OpenAI SDK, and their API costs were calculated accordingly. Results show that while smaller proprietary models

Table 5. Example of an LLM annotating cell type using the proposed zero-shot CoT approach. The LLM successfully reasons based on the provided gene list to determine the most likely cell type

System: You are a biology expert who always responds the cell-type annotation result by carefully considering the markers provided by the user.

User: Given the following markers [*CLEC9A*, *PPY*, *AL118508.1*, *LINC02206*, *LINC01724*, *MIR1273H*, *CEACAM3*, *CLCN1*, *AC092809.2*, *CYP2E1*], what is the cell type in *Thymus* corresponding to these markers? ***Let'sthinkstepbystep.***

Assistant: To determine the cell type in the thymus corresponding to the given markers, we need to analyze each marker and its known associations.

Here are the steps to identify the cell type:
1. CLEC9A: this marker is typically associated with dendritic cells, particularly a subset known as CD141+ (BDCA-3+) dendritic cells. These cells are involved in antigen presentation and are known for their role in the immune response.
2. PPY (Pancreatic Polypeptide): this marker is generally associated with pancreatic cells, specifically PP cells in the pancreas. However, its presence in the thymus might be less common and could indicate a specific subset of cells or a particular state of differentiation.
3. AL118508.1, LINC02206, LINC01724, MIR1273H, AC092809.2: these are noncoding RNAs (long intergenic noncoding RNAs and microRNAs). The specific expression patterns of these noncoding RNAs can be cell-type-specific, but they are less commonly used as primary markers for identifying cell types without additional context.
4. CEACAM3: this marker is typically associated with granulocytes, particularly neutrophils. It is involved in the immune response and pathogen recognition.
5. CLCN1: this gene encodes a chloride channel protein, which is primarily associated with muscle cells. Its relevance in the thymus might be less direct.
6. CYP2E1: this gene encodes an enzyme involved in drug metabolism and is primarily expressed in the liver. Its expression in the thymus could be indicative of specific metabolic activities but is not a primary marker for thymic cell types.
Given the context of the thymus, the most relevant marker here is CLEC9A, which points toward dendritic cells. The presence of CEACAM3 might suggest some involvement of granulocytes or neutrophils, but this is less typical for the thymus.

Considering the primary role of the thymus in T-cell development and the presence of dendritic cells for antigen presentation, the most likely cell type corresponding to these markers in the thymus is a subset of dendritic cells.
Therefore, the cell type in the thymus corresponding to these markers is likely dendritic cells, specifically the CD141+ (BDCA-3+) subset.

User: In summary, the most likely cell type (directly return one cell type name) is

Assistant: Dendritic cells ✓ (Correct Answer: Myeloid dendritic cell)

Table 6. Inference time and API cost of evaluated LLMs for cell-type annotation

Language model	Elapsed time	API cost
DeepSeek-LLM	23.6	–
Qwen2	17.19	–
Llama-3	12.72	–
Mixtral-8×7B	16.34	–
Mixtral-8×22B	44.72	–
Cell2Sentence	0.11	–
GPT-4o mini	1.46	$0.26
GPT-4o	1.53	$7.71

such as GPT-4o mini offer low latency and cost, larger models like GPT-4o incur significantly higher API expenses despite comparable inference speed. For open-weight LLMs, inference latency can be further optimized using inference acceleration frameworks such as vLLM or alternative deployment strategies, which may help narrow the performance gap with API-based models.

Limitations

We acknowledge several additional limitations of the current approach. First, LLMs may exhibit biases derived from their pretraining corpora, which can influence performance across tissues with varying levels of representation in biomedical literature (Fig. 3). This may explain why LLMs perform better on well-studied tissues but less accurately on rare or less-characterized tissues. Second, while our experiments show that LLMs generalize well across tissue types, they tend to misclassify rare or novel cell types due to limited prior knowledge. Third, in the cross-omics translation module used to map ATAC-seq data into the RNA-seq domain, the transformation process may lose information critical for fine-grained cell identity, leading to reduced annotation accuracy (Fig. 4). Furthermore, the applicability of the proposed RNA-seq bridging technique is contingent on the availability of a robust pretrained VAE across relevant tissue types. In this study, we adopted the scACT module, which was pretrained on PBMC and PFC datasets, for cross-modality translation. Extending this method to other tissue types (e.g. tumors) or to cross-species datasets would require either leveraging a VAE pretrained on the target domain or conducting pretraining on a domain-specific custom dataset. Addressing these challenges offers a promising direction for future research to further improve the robustness and generalizability of LLMs in biological applications.

Although extensive evaluations have been conducted on the proposed benchmarks, SOAR-RNA and SOAR-MultiOmics, further evaluation is needed to fully understand the ability of LLMs to analyze genomics data. One area for improvement involves a thorough investigation into retrieval-augmented generation during the reasoning stage of cell-type annotation. This approach aims to enhance LLMs' reasoning capabilities by leveraging existing biological knowledge, thereby improving the precision of annotating novel and complex gene co-expression patterns. Additionally, due to data constraints, the majority of multiomics benchmarks in this study are limited to RNA-seq and ATAC-seq data. While the proposed benchmark has significant influence within the LLM for scientific data community, comprehensive validation across a broader range of omics data is essential to fully assess LLMs' capabilities in analyzing biological data. These improvements can be facilitated through collaborative efforts within the research community to collect more publicly accessible multiomics data.

Conclusion

This study introduces SOAR, a pioneering benchmarking effort that evaluates the capabilities of instruction-tuned LLMs for cell-type annotation across single-cell genomics data from various modalities. By curating a diverse dataset encompassing multiple species and cell types, we systematically assessed the ability of LLMs to process and analyze complex biological data. Our findings demonstrate that LLMs exhibit strong interpretive capabilities in genomics analysis via RNA-seq bridging, even without extensive fine-tuning, and can effectively generate reasoning processes that support biological insights through the proposed domain-specific CoT prompting techniques. Additionally, we explored the application of these models to multiomics data, highlighting their potential for cross-modality analysis and providing a foundation for future advancements in automated single-cell annotation. This work underscores the promise of LLMs in transforming cell-type annotation workflows, while also emphasizing the need for ongoing innovation to fully exploit their potential across diverse molecular modalities.

Key Points

- This study establishes a systematic benchmarking framework by evaluating eight instruction-tuned LLMs across 1226 cell-type annotation-related tasks using 11 publicly available datasets, providing a robust foundation for automated and interpretable cell-type annotations.
- By incorporating the proposed domain-specific CoT techniques, this work significantly improves the accuracy of LLMs in cell-type classification while enhancing their reasoning ability to extract and utilize relevant biological insights.
- To extend LLM-based annotation beyond scRNA-seq, this study introduces a cross-modality translation module that converts uninterpretable features, such as epigenetic marks and chromatin structures, into interpretable ones, enabling seamless application of LLMs to a broader range of single-cell sequencing technologies.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions.

Author contributions

Junhao Liu (Conceptualization, Formal analysis, Investigation, Writing—original draft, Writing—review & editing), Siwei Xu (Formal analysis), Yongxian Wu (Writing—original draft, Writing—review & editing), and Jing Zhang (Conceptualization, Funding acquisition, Writing—original draft, Writing—review & editing)

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Competing interests: No competing interest is declared.

Funding

This work was supported by the National Institutes of Health (R01HG012572, R01DA063316).

Data availability

The source code and benchmark datasets of SOAR are publicly available at https://github.com/aicb-ZhangLabs/SOAR.

References

1. Vandereyken K, Sifrim A, Thienpont B. *et al*. Methods and applications for single-cell and spatial multi-omics. *Nat Rev Genet* 2023;**24**:494–515. https://doi.org/10.1038/s41576-023-00580-2
2. Baysoy A, Bai Z, Satija R. *et al*. The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 2023;**24**:695–713. https://doi.org/10.1038/s41580-023-00615-w
3. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet* 2019;**20**:257–72. https://doi.org/10.1038/s41576-019-0093-7
4. Ma A, McDermaid A, Jennifer X. *et al*. Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol* 2020;**38**:1007–22. https://doi.org/10.1016/j.tibtech.2020.02.013
5. Wu KE, Yost KE, Chang HY. *et al*. Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc Natl Acad Sci* 2021;**118**:e2023070118. https://doi.org/10.1073/pnas.2023070118
6. Jagadeesh KA, Dey KK, Montoro DT. *et al*. Identifying disease-critical cell types and cellular processes by integrating single-cell rna-sequencing and human genetics. *Nat Genet* 2022;**54**:1479–92. https://doi.org/10.1038/s41588-022-01187-9
7. Eraslan G, Drokhlyansky E, Anand S. *et al*. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* 2022;**376**:eabl4290. https://doi.org/10.1126/science.abl4290
8. Levine D, Rizvi SA, Lévy S. *et al*. Cell2Sentence: teaching large language models the language of biology. In: *Proceedings of the 41st International Conference on Machine Learning*. 2024;**235**:27299–325.
9. Achiam J, Adler S, Agarwal S. *et al*. GPT-4 technical report. arXiv 2023;arXiv:2303.08774.
10. Yang A, Yang B, Hui B. *et al*. Qwen2 technical report. arXiv 2024;arXiv:2407.10671.
11. Grattafiori A, Dubey A, Jauhri A. *et al*. The Llama 3 herd of models. arXiv 2024;arXiv:2407.21783.
12. Bi X, Chen D, Chen G. *et al*. DeepSeek LLM: scaling open-source language models with Longtermism. arXiv 2024; arXiv:2401.02954.
13. Jiang AQ, Sablayrolles A, Roux A. *et al*. Mixtral of experts. arXiv 2024; arXiv:2401.04088.
14. Theodoris CV, Xiao L, Chopra A. *et al*. Transfer learning enables predictions in network biology. *Nature* 2023;**618**:616–24. https://doi.org/10.1038/s41586-023-06139-9
15. Cui H, Wang C, Maan H. *et al*. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:1470–80.
16. Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods* 2024;**21**:1462–5.
17. Congxue H, Li T, Yingqi X. *et al*. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* 2023;**51**:D870–6.
18. 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. https://doi.org/10.1038/s41590-018-0276-y
19. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* 2022;**13**:1246. https://doi.org/10.1038/s41467-022-28803-w

20. Hendrycks D, Burns C, Basart S. *et al*. Measuring massive multitask language understanding. In: *International Conference on Learning Representations*. 2021.
21. Cobbe K, Kosaraju V, Bavarian M. *et al*. Training verifiers to solve math word problems. arXiv 2021; arXiv:2110.14168.
22. Kojima T, Shixiang Shane G, Reid M. *et al*. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 2022;**35**:22199–213.
23. Mimitou EP, Lareau CA, Chen KY. *et al*. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat Biotechnol* 2021;**39**(10):1246–58.
24. Lieberman-Aiden E, Van Berkum, Williams L. *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**:289–93. https://doi.org/10.1126/science.1181369
25. Brown TB, Mann B, Ryder N. *et al*. Language models are few-shot learners. *Advances in neural information processing systems* 2020;**33**:1877–901.
26. Bai J, Bai S, Chu Y. *et al*. Qwen technical report. arXiv 2023; arXiv:2309.16609.
27. Touvron H, Martin L, Stone K. *et al*. Llama 2: open foundation and fine-tuned chat models. arXiv 2023; arXiv:2307.09288.
28. Touvron H, Lavril T, Izacard G. *et al*. Llama: open and efficient foundation language models. arXiv 2023; arXiv:2302.13971.
29. Taori R, Gulrajani I, Zhang T. *et al*. *Stanford Alpaca: An Instruction-Following Llama Model* https://github.com/tatsu-lab/stanford_alpaca2023.
30. Liu H, Li C, Wu Q. *et al*. Visual instruction tuning. *Advances in neural information processing systems* 2023;**36**:34892–916.
31. Liu Y, Shi K, He K. *et al*. On learning to summarize with large language models as references. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024:8647–66.
32. Ouyang L, Jeffrey W, Jiang X. *et al*. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 2022;**35**:27730–44.
33. Devlin J, Chang MW, Lee K. *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT* 2019:4171–86.
34. Xu S, Liu J, Zhang J. scACT: accurate cross-modality translation via cycle-consistent training from unpaired single-cell data. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2024;2722–31.
35. consortium HuBMAP. The human body at cellular resolution: the NIH human biomolecular atlas program. *Nature* 2019;**574**:187–92. https://doi.org/10.1038/s41586-019-1629-x
36. Han X, Zhou Z, Fei L. *et al*. Construction of a human cell landscape at single-cell level. *Nature* 2020;**581**:303–9. https://doi.org/10.1038/s41586-020-2157-4
37. Han X, Wang R, Zhou Y. *et al*. Mapping the mouse cell atlas by Microwell-seq. *Cell* 2018;**172**:1091–1107.e17. https://doi.org/10.1016/j.cell.2018.02.001
38. Liu N, Jiang C, Yao X. *et al*. Single-cell landscape of primary central nervous system diffuse large B-cell lymphoma. *Cell Discovery* 2023;**9**:55. https://doi.org/10.1038/s41421-023-00559-7
39. Lee H-O, Hong Y, Etlioglu HE. *et al*. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat Genet* 2020;**52**:594–603. https://doi.org/10.1038/s41588-020-0636-z
40. Kim N, Kim HK, Lee K. *et al*. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* 2020;**11**:2285. https://doi.org/10.1038/s41467-020-16164-1
41. Consortium The Tabula Sapiens, Jones RC, Karkanias J. *et al*. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* 2022;**376**:eabl4896.
42. Chen D, Sun J, Zhu J. *et al*. Single cell atlas for 11 non-model mammals, reptiles and birds. *Nat Commun* 2021;**12**:7083. https://doi.org/10.1038/s41467-021-27162-2
43. Wolf FA, Angerer P, Theis FJ. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:1–5.
44. Jupp S, Burdett T, Leroy C. *et al*. A new ontology lookup service at EMBL-EBI. *SWAT4LS* 2015;**2**:118–9.
45. Noy NF, Shah NH, Whetzel PL. *et al*. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;**37**:W170–3.
46. Emani PS, Liu JJ, Clarke D. *et al*. Single-cell genomics and regulatory networks for 388 human brains. *Science* 2024;**384**:eadi5199.
47. Wolf T, Debut L, Sanh V. *et al*. Transformers: state-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020;38–45.
48. Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74–81.
49. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor (MI): Association for Computational Linguistics; 2005. p. 65–72.
50. Papineni K, Roukos S, Ward T. *et al*. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002:311–8.
51. Lin C-Y, Och FJ. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland: COLING; 2004. p. 501–7.
52. Rajpurkar P, Zhang J, Lopyrev K. *et al*. SQuAD: 100, 000+ questions for machine comprehension of text. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016:2383–92.
53. Chen J, Chen L, Huang H. *et al*. When do you need chain-of-thought prompting for chatGPT? arXiv 2023;arXiv:2304.03262.

Briefings in Bioinformatics, 2025, **26(6)**, bbaf622
https://doi.org/10.1093/bib/bbaf622
Problem Solving Protocol