ENABLING LARGE LANGUAGE MODELS TO COMPREHEND THE LANGUAGE OF SINGLE-CELL PAIRED MULTI-OMICS DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have recently shown great potential for single-cell analysis, offering powerful performance, natural-language interpretability, and flexible usability. However, existing LLM-based single-cell methods remain largely centered on single-omics (single-modal) data, and their extension to paired multi-omics (multi-modal) data remains challenging. An empirical study reveals a counterintuitive "*more modalities, worse performance*" phenomenon: directly providing paired multi-omics data to LLMs fails to improve performance and can even lead to worse performance than single-omics inputs. We argue that this limitation stems from the prevailing *text-level concatenation paradigm* in this field, which serializes heterogeneous multi-omics data into discrete text and simply concatenates them as inputs to an LLM. This paradigm has fundamental limitations and cannot effectively integrate complex and heterogeneous multi-omics information. To address this challenge, we propose **LLM M**ulti-**O**mics **Reader** (**LLM-MOReader**), a framework that shifts LLM-based paired single-cell multi-omics analysis from the text-level concatenation paradigm to *explicit multi-omics fusion and projection* paradigm. LLM-MOReader first uses variational inference to jointly model paired multi-omics data and learn a *fused multi-omics representation*, which is then mapped into LLM token space through *continuous multi-omics prompt*, enabling LLM-based reasoning over already-integrated multi-omics information. Experiments on diverse datasets across multiple species, modality combinations, and downstream tasks, including cell type annotation and multi-omics integration, show that LLM-MOReader matches or outperforms existing methods, including domain-specific LLMs, frontier large-scale LLMs and specialized non-LLM models. Further analyses show that our approach resolves the "more modalities, worse performance" problem and exhibits model-agnostic generality. LLM-MOReader enables LLMs to read the language of multi-omics, making LLM-based paired single-cell multi-omics analysis practically feasible.

1 INTRODUCTION

Single-cell omics technologies (Lee et al., 2020) measure molecular profiles at single-cell resolution, capturing complementary biological modality from different views, such as gene expression (Tang et al., 2009), chromatin accessibility (Buenrostro et al., 2015), and surface protein abundance (Stoeckius et al., 2017). **From a deep learning perspective, each omics type can be viewed as a biological modality. Accordingly, single-cell paired multi-omics data are analogous to multi-modal data, where multiple heterogeneous but complementary views are jointly observed for the same single-cell.** Recently, large language models (LLMs) have emerged as a promising paradigm for single-cell analysis. A growing body of work (Hou & Ji, 2024; Zhang et al., 2026; Zhou et al., 2026) has shown that LLMs achieve strong performance and that directly applying LLMs can even outperform domain-specific models. Compared with conventional domain-specific models, LLMs offer several attractive advantages: **1) powerful performance**, as LLMs can achieve competitive or even superior performance when directly applied. **2) natural-language interpretability**, since LLMs can generate human-readable reasoning explanations rather than only returning predictions. **3) flexible usability**: as LLM-based models can leverage open-source LLMs, thereby benefiting from

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

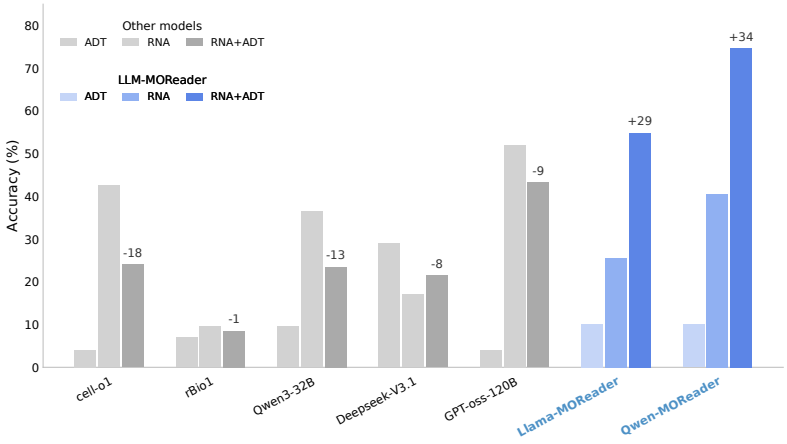


Figure 1: The previous LLM-based method (gray color) had counterintuitive "more modalities, worse performance" phenomenon. We proposed LLM-MOReader (blue color) model to solve this problem.

rapid advancements in LLMs. These attractive properties make LLM-based methods appealing for single-cell analysis.

The successful application of current LLM-based model remains largely limited to single-omics (i.e., single-modal) inputs, particularly scRNA-seq (single-cell RNA sequencing) (Jovic et al., 2022), despite the rapid development of single-cell multi-omics technology (Lee et al., 2020). This focus is understandable: scRNA-seq is the most widely available modality and can be easily converted into a text list that are naturally connected to LLMs. However, scRNA-seq reflects only one modality of a cell, while modern single-cell technologies increasingly provide paired multi-omics (i.e., multi-modal) data that characterize the same cells from multiple biological perspectives (Lee et al., 2020). **Analogous to the shift from text-only LLMs toward multi-modal LLMs, LLM-based single-cell analysis is also expected to move from single-omics analysis toward paired multi-omics analysis.** However, previous study (Zhou et al., 2026) and empirical study in Figure 1 reveal a counterintuitive phenomenon called *"more modalities, worse performance"*: directly feeding paired multi-omics inputs into LLMs often fails to improve and even leads to performance degradation.

We argue that this limitation stems from the dominant text-level concatenation paradigm used in current LLM-based single-cell methods. A representative early work, Cell2Sentence (Levine et al., 2024), converts scRNA-seq into lists of genes and feeds them to an LLM as natural language. A series of subsequent studies (Fang et al., 2026; Istrate et al., 2025; Zhou et al., 2026) follow this paradigm. However, extending this strategy to paired multi-omics data is fundamentally limited: under this paradigm, different omics modalities are separately converted into text fragments and then simply concatenated into the input sequence, which cannot faithfully integrate multiple omics modalities since different omics modalities often have distinct data distributions, sparsity patterns, and biological semantics. In other words, LLMs need an effective way to read the language of multi-omics data.

To address this problem, we propose **LLM M**ulti-**O**mics **Reader (LLM-MOReader)** as illustrated in Figure 2 and 3, a framework that enables LLMs to effectively analyze paired single-cell multi-omics data. LLM-MOReader first performs **explicit multi-omics fusion** in the omics feature space. Specifically, it uses a variational inference-based multi-omics encoder-decoder framework to jointly model paired multi-omics data and learns a **fused multi-omics representation** that integrates complementary information from multiple omics. LLM-MOReader then introduces a Projection Module that maps this representation into continuous prompt embeddings called the **joint multi-omics prompt**, bridging the omics space and the LLM embedding space. In this way, the LLM receives already-integrated information rather than loosely concatenated text fragments. Through a paradigm shift, LLM-MOReader avoids the limitations of previous text-level concatenation paradigm while preserving the reasoning and natural-language interaction abilities of LLMs. Conceptually, LLM-MOReader allows LLMs to better read the language of paired single-cell multi-omics data.

Extensive experiments across multiple species, multi-omics combinations, and downstream tasks demonstrate the effectiveness of LLM-MOReader. For the cell type annotation task, LLM-MOReader

substantially outperforms existing baselines, including frontier large-scale LLMs. More importantly, it reverses the "more modalities, worse performance" phenomenon as shown in Figure 1. For the multi-omics integration task, LLM-MOReader matches or outperforms state-of-the-art performance. In addition, LLM-MOReader is model-agnostic and performs well across different LLM backbones. Taken together, these compelling results establish LLM-MOReader as a practical route for extending LLM-based models from single-omics analysis to paired single-cell multi-omics analysis.

Our contributions are summarized as follows.

- **We address the key bottleneck of LLM-based single-cell analysis.** The previous LLM-based method faced "more modalities, worse performance" problem, while LLM-MOReader resolves it and enables LLMs to benefit from paired single-cell multi-omics data.
- **We propose LLM-MOReader.** LLM-MOReader performs explicit multi-omics fusion and projection, so that a general-purpose LLM can reason over already-integrated multi-omics evidence and read the language of paired single-cell multi-omics data.
- **We achieve strong performance across diverse settings.** Experiments on cell type annotation and multi-omics integration across multiple species and modality combinations show that LLM-MOReader matches or surpasses state-of-the-art baselines.

2 Related Work

LLM-based Models for Single-Cell Analysis. Large language models (LLMs) have recently emerged as a promising route for single-cell analysis. Several domain-specific LLM-based models (Fang et al., 2026; Istrate et al., 2025; Xiao et al., 2024) have been proposed for cell analysis. More recent benchmarks (Zhang et al., 2026; Zhou et al., 2026), further report that strong general LLMs often outperform domain-specific models. A complementary line of work (Li et al., 2024; Fang et al., 2025; Schaefer et al., 2025) couples omics measurements with the LLMs. These studies show that LLM reasoning can be transferred to single-cell analysis. Despite such advances, however, **existing LLM-based models are all formulated over a single omics modality, typically scRNA-seq, and provide limited evidence that they benefit from paired single-cell multi-omics data.**

Non-LLM Models for Paired Multi-Omics Analysis. In parallel with LLM-based single-cell analysis, a large body of work has focused on non-LLM models. Classical and statistical integration methods (Cheng et al., 2022; Stuart et al., 2019; Kim et al., 2020; Wang et al., 2023; Jin et al., 2020; Lynch et al., 2022; Singh et al., 2021; Ma et al., 2023) model cross-omics relationships using machine learning methods, while deep generative models (Gayoso et al., 2021; Ashuach et al., 2023; Du et al., 2022; Lotfollahi et al., 2022b;a) are based on deep learning. However, non-LLM multi-omics analysis models are usually based on unsupervised learning and learn embeddings for further analysis, but they do not directly support downstream tasks or natural-language interaction.

3 Methods

We propose LLM-MOReader, a two-stage framework for LLM-based paired single-cell multi-omics analysis, as illustrated in Figure 2 (stage 1) and Figure 3 (stage 2). This section is organized as follows: Section 3.1 defines the data format and the downstream tasks. Section 3.2 introduces the first stage of LLM-MOReader, in which a variational inference-based multi-omics encoder-decoder framework jointly models paired modalities and learns a **fused multi-omics representation**. Section 3.3 presents the second stage of LLM-MOReader, in which the fused representation is mapped into a continuous **joint multi-omics prompt**, which bridges the omics space and the LLM embedding space.

3.1 Input and Task Definition

In this work, we focus mainly on two common settings of paired single-cell multi-omics data: scRNA-seq + scATAC-seq (**RNA+ATAC**) (Chen et al., 2019) and scRNA-seq + surface protein abundance (**RNA+ADT**) (Stoeckius et al., 2017).

For the RNA modality, each cell c is represented by a gene expression vector $G_c = \{g_{c,1}, g_{c,2}, \ldots, g_{c,N}\}, \quad N \in \mathbb{Z}_+$, where $g_{c,i}$ denotes the gene expression level of the i-th gene

in cell c. In practice, N is typically on the order of 10^4. The feature vector G_c is often sparse and noisy (Imoto et al., 2022). **For the ATAC modality**, each cell c is represented by a chromatin accessibility vector $A_c = \{a_{c,1}, a_{c,2}, \ldots, a_{c,M}\}, \quad M \in \mathbb{Z}_+$, where $a_{c,i}$ denotes the accessibility signal of the i-th genomic region in cell c. In practice, M is typically on the order of 10^5. Compared with the RNA modality, ATAC data are even sparser, more high-dimensional and noisier due to the nature of chromatin accessibility data (Li et al., 2021). **For the ADT modality**, each cell c is represented by a surface protein abundance vector $P_c = \{p_{c,1}, p_{c,2}, \ldots, p_{c,K}\}, \quad K \in \mathbb{Z}_+$, where $p_{c,i}$ denotes the abundance of the i-th surface protein in cell c. ADT modality is highly informative and dense (Zheng et al., 2025), has a much smaller dimensionality K, often on the order of 10^2.

Therefore, RNA, ATAC, and ADT differ substantially in dimensionality, sparsity, noise characteristics, and biological semantics. This heterogeneity and complexity of different modalities make naive text-level concatenation suboptimal, which motivates our explicit multi-omics fusion paradigm. We further discuss the limitations of text-level concatenation from a formal point of view in Appendix D.

Given a paired single-cell multi-omics data sample from cell c, we formally represent it as (X_c^1, X_c^2), where X_c^1 and X_c^2 correspond to two different omics views of the same single-cell c. In the RNA+ATAC setting, $\left(X_c^1, X_c^2\right) = (G_c, A_c)$, while in the RNA+ADT setting, $\left(X_c^1, X_c^2\right) = (G_c, P_c)$.

In the Cell Type Annotation (CTA) task, the goal is to predict the ground-truth cell type label y_c of cell c from its input (X_c^1, X_c^2), that is $f_{\text{CTA}} : (X_c^1, X_c^2) \mapsto \hat{y}_c$, where $\hat{y}_c$ denotes the predicted cell type. **In the Multi-Omics Integration (MOI) task**, the goal is to extract a high-dimensional cell-level embedding h_c for downstream integration analysis, that is $f_{\text{MOI}} : (X_c^1, X_c^2) \mapsto h_c$. The objective of the MOI task is to learn a cell-embedding h_c with strong biological conservation.

3.2 Variational Inference-based Framework for Explicit Multi-Omics Fusion

The first stage of LLM-MOReader aims to transform heterogeneous paired multi-omics inputs into a unified and biologically meaningful representation, thereby achieving explicit multi-omics fusion in the omics space. As illustrated in Figure 2, LLM-MOReader uses a variational inference-based multi-omics encoder-decoder framework to jointly model the paired multi-omics data of each cell.

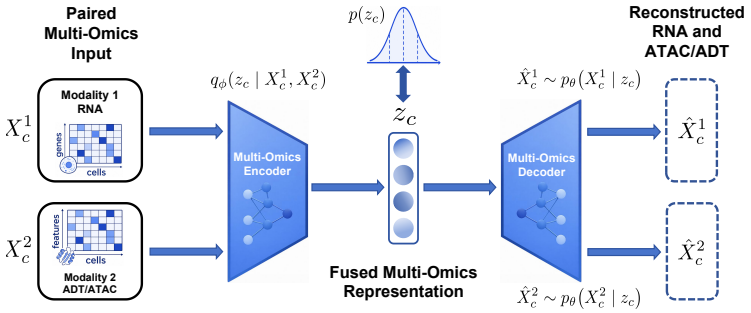


Figure 2: **Variational inference-based multi-omics encoder-decoder framework proposed in LLM-MOReader.** The Multi-Omics Encoder infers the fused multi-omics representation from the input to achieve explicit multi-omics fusion, while the Multi-Omics Decoder reconstruct the inputs.

Given a cell c with two paired omics modalities (X_c^1, X_c^2), we assume that both modalities are generated from a shared latent variable $z_c \in \mathbb{R}^{d_z}$ called the **fused multi-omics representation**. The key idea is that, although the two modalities are heterogeneous, they describe the same underlying cellular state of cell c. Both modalities are then assumed to be generated from this representation as

$$z_c \sim p(z_c), \qquad X_c^1 \sim p_\theta\left(X_c^1 \mid z_c\right), \qquad X_c^2 \sim p_\theta\left(X_c^2 \mid z_c\right), \tag{1}$$

where $p(z_c) = \mathcal{N}(0, I)$ is the prior distribution, and $p_\theta(X_c^1 \mid z_c)$ and $p_\theta(X_c^2 \mid z_c)$ approximate modality-specific likelihood functions. The likelihood function is chosen according to statistical properties, allowing to respect the heterogeneity of RNA, ATAC, and ADT rather than forcing all modalities into a unified textual format. See Appendix E for more details about likelihood functions.

Since posterior $p(z_c|X_c^1, X_c^2)$ is intractable, Multi-Omics Encoder is introduced to approximate it:

$$q_\phi\left(z_c \mid X_c^1, X_c^2\right). \tag{2}$$

To ensure that z_c preserves information from both modalities, we use Multi-Omics Decoder to reconstruct the original inputs and optimize the evidence lower bound (ELBO) (Blei et al., 2017):

$$\hat{X}_c^1 \sim p_\theta\left(X_c^1 \mid z_c\right), \qquad \hat{X}_c^2 \sim p_\theta\left(X_c^2 \mid z_c\right). \tag{3}$$

$$\mathcal{L}_{\mathrm{ELBO}}^c = \mathbb{E}_{q_\phi(z_c \mid X_c^1, X_c^2)}\left[\log p_\theta\left(X_c^1 \mid z_c\right) + \log p_\theta\left(X_c^2 \mid z_c\right)\right] - \mathrm{KL}\left(q_\phi\left(z_c \mid X_c^1, X_c^2\right) \,\|\, p(z_c)\right). \tag{4}$$

The first reconstruction term encourages z_c to preserve modality-specific and cross-modal biological signals, while the second KL divergence term regularizes the posterior toward a smooth latent space.

The proposed variational inference-based encoder-decoder framework provides an explicit multi-omics fusion paradigm: 1) the Multi-Omics Encoder integrates paired multi-omics inputs into unified representation through $q_\phi(z_c \mid X_c^1, X_c^2)$, rather than treating each modality as independent textual inputs; 2) modality-specific likelihoods allow the model to handle heterogeneous multi-omics data; 3) the shared fused multi-omics representation z_c acts as a fusion bottleneck that must explain both modalities to reconstruct the original inputs, thereby encouraging it to capture cross-omics complementarity. Consequently, the first stage produces an integrated **fused multi-omics representation** z_c, which is further transformed into the **joint multi-omics prompt** in the next stage.

3.3 Bridging the Omics Space and the LLM Embedding Space

The second stage of LLM-MOReader aims to bridge the gap between the fused multi-omics representation z_c space and the LLM token embedding space, as illustrated in Figure 3. Formally, we use a trainable Projection Module f_{proj} to map the omics space into the input embedding space of LLM:

$$P_c = f_{\mathrm{proj}}(z_c), \qquad P_c \in \mathbb{R}^{L \times d}, \tag{5}$$

where L is the length of the continuous prompt and d is the hidden dimension of the LLM, respectively.

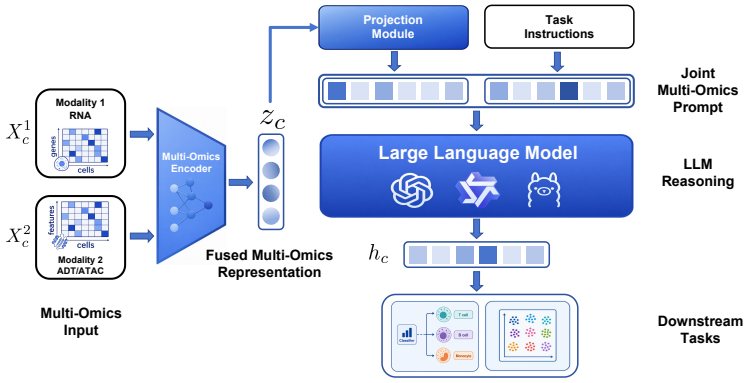


Figure 3: **Joint multi-omics prompt for LLM reasoning.** The fused multi-omics representation z_c is mapped into the LLM token embedding space by the Projection Module and combined with task instructions to form the LLM input. The resulting representation h_c is used for downstream tasks.

The resulting P_c is referred to as the **joint multi-omics prompt**. We further provide a textual instruction T_{task} that describe the task details in natural language. Let $E(T_{\mathrm{task}}) \in \mathbb{R}^{l \times d}$ denote the token embedding sequence of the instruction, where l is the instruction length. The final input to the LLM is constructed by combining joint multi-omics prompt and task instruction $\widetilde{P}_c = [P_c, E(T_{\mathrm{task}})]$. The LLM then processes this combined embedding sequence to infer $h_c = f_{\mathrm{LLM}}(\widetilde{P}_c)$, where h_c denotes the hidden states conditioned on both fused multi-omics information and task instruction.

During training, h_c is further transformed by the LLM's original output head into textual response, and the Projection Module is optimized using cross-entropy loss: $\mathcal{L}_{\mathrm{proj}} = -\sum_c \log p_{\mathrm{LLM}}(y_c \mid h_c)$. During inference, how h_c is used depends on the downstream task. For cell type annotation, h_c is decoded by the LLM's output head into a textual cell type prediction. For multi-omics integration task, h_c is further extracted as the cell-level embedding of cell c and used for downstream analysis.

The proposed joint multi-omics prompt serves as a carrier of integrated biological multi-omics information from single-cell and is aligned to the embedding space of the LLM.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Baselines. For cell type annotation, we include the following baselines: 1) domain-specific LLMs, including `rBio1` and `cell-o1`, which are specifically developed and trained for single-cell analysis; 2) general-purpose LLMs, including `DeepSeek-V3.1`, the `GPT-oss` series and the `Qwen3` series; 3) frontier large-scale LLMs, including `Gemini` series, `Qwen 3.6 Plus`, `GLM-5.1`, `Kimi k2.6`, `DeepSeek V4 Flash`, and `GPT-5`; 4) three specialized non-LLM models including `scArches`, `Seurat`, and `totalVI`, are used with MLPs trained on their output features to perform the CTA task. For multi-omics integration, we include `Seurat`, `MOJITOO`, `totalVI`, `Multigrate`, `SCOIT`, `CiteFuse`, `DeepMAPS`, `scArches`, `scVAEIT`, `scAI`, `MultiVI`, `MIRA`, `Schema`, `totalANVI`, and `UnitedNet`. See Appendix F for more baseline details.

Instantiation. We instantiate LLM-MOReader as Llama-MOReader (based on Llama-3.1-8B and Qwen-MOReader (based on Qwen3-32B). See Appendix G for more model training details. For LLM-MOReader in the main experiments, the LLM backbones use their officially released weights without fine-tuning. This controlled setting prevents backbone adaptation from confounding the comparison and ensures that the gains of LLM-MOReader reflect its framework design.

Datasets and Metrics. For the cell type annotation task, we use the **Human Bone Marrow Mononuclear Cells dataset** (Luecken et al., 2021). We report Accuracy, F1, Precision and Recall, with **Accuracy** and **F1** as the primary metric. For the multi-omics integration task, we use four public datasets from Hu et al. (2024) referred as **Mouse Brain (RNA+ADT)**, **Human PBMC (RNA+ADT)**, **Human Brain (RNA+ATAC)**, and **Mouse Retina (RNA+ATAC)** for readability. We report NMI, ARI, cASW, cLISI, and BVC, with **BVC** as the primary metric. See Appendices B and C.

4.2 MAIN RESULTS

4.2.1 CELL TYPE ANNOTATION TASK

LLM-MOReader reverses the "more modalities, worse performance" phenomenon. Table 1 contrasts the ADT-only, RNA-only, and paired RNA+ADT settings for LLM-based models. Although paired multi-omics data carry more biological evidence, every general-purpose LLM baseline degrades when moving from its best single-omics setting to the multi-omics setting. LLM-MOReader reverses this trend: Llama-MOReader and Qwen-MOReader gains 29.29 and 34.16 Accuracy points, respectively. This contrast indicates that the degradation is not an intrinsic property of multi-omics inputs, but rather a consequence of how these inputs are presented to the LLM.

Table 1: Effect of "more modalities, worse performance" phenomenon on LLM-based cell type annotation. **F1 and Accuracy (Acc)** are the primary metrics. ΔAccuracy denotes the Accuracy change from the better single-omics setting to the paired multi-omics setting; The best result of each metric under the multi-omics setting is shown in **bold** and the second-best is underlined.

Modality	**Single-Omics (ADT)**				**Single-Omics (RNA)**				**Multi-Omics (RNA+ADT)**				
Metric	Prec	Recall	**F1**	**Acc**	Prec	Recall	**F1**	**Acc**	Prec	Recall	**F1**	**Acc**	**ΔAcc**
Domain-specific LLMs													
cell-o1	1.81	1.01	0.70	4.03	14.27	18.50	14.81	42.51	7.97	7.95	7.32	24.09	−18.42
rBio1	2.23	3.87	2.11	7.00	6.77	6.04	3.72	9.50	3.41	4.42	3.06	8.50	−1.00
General LLMs													
Qwen3-4B	3.44	2.69	2.24	7.01	4.50	9.81	5.15	24.00	2.19	1.58	1.41	3.52	−20.48
Qwen3-8B	2.24	2.46	1.56	2.51	9.94	7.72	7.31	25.44	9.04	3.54	3.90	9.00	−16.44
Qwen3-14B	9.51	9.51	6.50	12.09	10.83	14.06	8.92	30.52	15.39	7.25	7.82	14.49	−16.03
Qwen3-30B	5.72	7.70	4.73	10.00	9.53	10.89	9.05	37.50	9.77	7.94	7.22	26.00	−11.50
Qwen3-32B	9.26	10.88	5.91	9.50	12.84	12.79	10.80	36.50	12.52	12.60	8.56	23.44	−13.06
Deepseek-V3.1	17.78	19.01	15.20	29.01	17.08	6.81	8.77	17.00	18.90	12.64	13.05	21.48	−7.53
GPT-oss-20B	10.49	2.68	3.46	3.00	21.86	15.74	16.37	42.05	20.73	12.82	12.98	30.41	−11.64
GPT-oss-120B	3.83	4.71	3.08	4.00	22.18	24.68	22.83	52.04	32.93	30.81	27.63	43.30	−8.74
LLM-MOReader													
Llama-MOReader	5.60	9.27	4.58	10.12	7.86	10.66	7.83	25.56	44.97	42.10	37.60	54.85	+29.29
Qwen-MOReader	8.54	11.04	6.39	10.00	10.21	15.94	11.04	40.51	**55.01**	**50.95**	**49.48**	**74.67**	+34.16

LLM-MOReader achieves state-of-the-art performance on paired multi-omics cell type annotation. Table 2 compares all evaluated models under the paired multi-omics (RNA+ADT) setting. Qwen-MOReader attains the best result on all four metrics. Notably, the strongest baselines are the specialized non-LLM models rather than the LLM-based methods, which is consistent with our observation that current LLM-based methods cannot effectively exploit paired multi-omics inputs. Qwen-MOReader improves upon Seurat, the strongest baseline, by 12.75 Accuracy points and 11.04 Macro F1 points. Llama-MOReader, built on a frozen 8B backbone, outperforms every LLM-based baseline by a large margin and is comparable to the specialized non-LLM models. These results demonstrate that LLM-MOReader is effective for LLM-based single-cell multi-omics annotation.

LLM-MOReader surpasses frontier large-scale LLMs, showing that explicit multi-omics fusion and projection is much more important than merely scaling the LLM. We further compare LLM-MOReader with frontier large-scale LLMs in Figure 4. This comparison highlights a scale gap: Llama-MOReader and Qwen-MOReader are built on only 8B Llama-3.1 and 32B Qwen3, respectively. By contrast, many of the frontier LLM baselines are much larger, such as the 1T Kimi-k2.6 and 754B GLM-5.1. Nevertheless, Llama-MOReader already outperforms 6 out of the 8 frontier large-scale LLMs and Qwen-MOReader surpasses all of them. This result reveals that simply increasing the scale of the LLM is not sufficient. In contrast, once equipped with explicit multi-omics fusion and projection, LLM-MOReader can outperform frontier LLMs with substantially larger scale.

Table 2: Main results on the paired multi-omics (**RNA+ADT**) cell type annotation task.

	Domain-Specific LLMs		**Specialized non-LLM Models**			**General LLMs**				**LLM-MOReader**	
Metric	cell-o1	rBio1	scArches	Seurat	totalVI	Qwen3-32B	Deepseek-V3.1	GPT-oss-20B	GPT-oss-120B	**Llama-MOReader**	**Qwen-MOReader**
Prec	7.97	3.41	47.31	38.50	42.25	12.52	18.90	20.73	32.93	44.97	**55.01**
Recall	7.95	4.42	40.13	44.16	37.76	12.60	12.64	12.82	30.81	42.10	**50.95**
F1	7.32	3.06	36.87	38.44	34.51	8.56	13.05	12.98	27.63	37.60	**49.48**
Acc	24.09	8.50	60.57	61.92	58.60	23.44	21.48	30.41	43.30	54.85	**74.67**

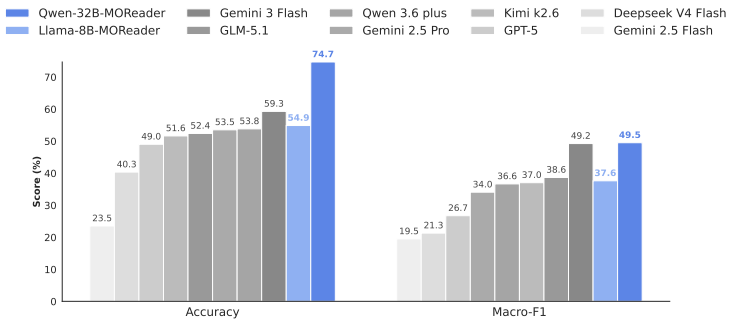


Figure 4: Performance comparison with frontier large-scale LLMs (gray color). LLM-MOReader (blue color) achieves stronger overall performance than frontier LLMs with substantially larger scale.

General-purpose LLMs outperform domain-specific LLMs. Consistent with previous studies (Zhang et al., 2026; Zhou et al., 2026), the general-purpose GPT-oss-120B reaches 52.04 Accuracy under single-omics inputs (Table 1) and 43.30 Accuracy under paired multi-omics inputs (Table 2), outperforming the domain-specific LLMs cell-o1 and rBio1 in both settings.

We further evaluate LLM-MOReader on the RNA+ATAC cell type annotation task in Appendix A.1.

4.2.2 Multi-Omics Integration Task

LLM-MOReader matches or surpasses state-of-the-art baselines on multi-omics integration, demonstrating its effectiveness in fusing multi-omics data. We further evaluate LLM-MOReader on multi-omics integration, a task that has traditionally been dominated by specialized non-LLM methods, as shown in Tables 3a and 3b. **On RNA+ADT integration, LLM-MOReader surpasses all baselines**: Qwen-MOReader obtains the highest NMI, ARI, cLISI, and BVC on both datasets,

(a) Main results on the multi-omics integration (**RNA+ADT**) task. **BVC** is the primary metric in this task.

Algorithm	Mouse Brain (RNA+ADT)					Human PBMC (RNA+ADT)				
	NMI	ARI	cASW	cLISI	**BVC**	NMI	ARI	cASW	cLISI	**BVC**
Seurat	0.62	0.69	**0.69**	1.00	0.75	0.63	0.62	**0.58**	0.98	0.70
MOJITOO	0.69	0.87	0.64	1.00	0.80	0.68	0.69	0.53	0.97	0.72
totalVI	0.56	0.36	0.57	1.00	0.62	0.68	0.71	0.53	0.97	0.72
totalANVI	0.62	0.86	0.54	1.00	0.76	0.69	0.70	0.55	1.00	0.73
Multigrate	0.53	0.35	0.66	1.00	0.63	0.66	0.66	0.56	0.98	0.72
SCOIT	0.62	0.70	0.55	1.00	0.72	0.65	0.69	0.52	0.95	0.70
CiteFuse	0.62	0.43	0.58	1.00	0.66	0.63	0.65	0.53	0.96	0.70
DeepMAPS	0.59	0.42	0.55	1.00	0.64	0.65	0.67	0.52	0.96	0.70
scArches	0.54	0.35	0.56	1.00	0.61	0.67	0.70	0.53	0.97	0.72
scVAEIT	0.50	0.36	0.61	1.00	0.62	0.64	0.66	0.54	0.97	0.70
Llama-MOReader	0.65	0.88	0.53	1.00	0.77	0.61	0.71	0.54	1.00	0.73
Qwen-MOReader	**0.73**	**0.92**	0.59	**1.00**	**0.81**	**0.70**	**0.75**	0.56	**1.00**	**0.75**

(b) Main results on the multi-omics integration (**RNA+ATAC**) task. **BVC** is the primary metric in this task.

Algorithm	Human Brain (RNA+ATAC)					Mouse Retina (RNA+ATAC)				
	NMI	ARI	cASW	cLISI	**BVC**	NMI	ARI	cASW	cLISI	**BVC**
scAI	**0.90**	0.95	0.73	1.00	**0.89**	0.66	0.32	0.70	1.00	0.67
MOJITOO	0.85	0.89	0.64	1.00	0.84	**0.84**	0.90	0.69	1.00	**0.86**
MultiVI	0.79	0.65	0.80	1.00	0.81	0.74	0.64	0.79	1.00	0.79
Seurat	0.78	0.57	**0.82**	1.00	0.79	0.72	0.45	0.80	1.00	0.74
Multigrate	0.81	0.94	0.71	1.00	0.87	0.75	0.72	0.66	1.00	0.78
MIRA	0.75	0.88	0.69	1.00	0.83	0.78	0.86	0.70	1.00	0.84
DeepMAPS	0.82	0.93	0.66	1.00	0.85	0.83	0.90	0.59	1.00	0.83
Schema	0.67	0.53	0.58	0.99	0.69	0.76	0.88	0.75	1.00	0.85
UnitedNet	0.77	0.40	0.69	1.00	0.72	0.67	0.31	**0.91**	1.00	0.72
Llama-MOReader	0.84	0.90	0.67	1.00	0.85	0.78	**0.92**	0.66	1.00	0.84
Qwen-MOReader	0.87	**0.96**	0.65	**1.00**	0.87	0.80	0.89	0.69	**1.00**	0.85

with BVC scores of 0.81 and 0.75. **On RNA+ATAC integration, LLM-MOReader is on par with the strongest baselines**. On Human Brain, Qwen-MOReader achieves the best ARI of 0.96 and a BVC of 0.87, within 0.02 of the best baseline. On Mouse Retina, Llama-MOReader achieves the best ARI of 0.92, while Qwen-MOReader obtains a BVC of 0.85, within 0.01 of the best baseline.

LLM-MOReader demonstrates strong generalization. LLM-MOReader demonstrates consistently strong performance and generalization across human and mouse datasets, RNA+ADT and RNA+ATAC combinations, and both annotation and integration tasks. The robust performance of LLM-MOReader provides a practical route for extending LLMs to paired multi-omics analysis.

4.3 Key Observations from Ablation Study

Observation 1: Explicit multi-omics fusion and projection is essential for single-cell multi-omics analysis. To further verify the essential of framework design, we introduce a **Trainable Soft-Prompt Model**, which is trained on exactly the same data, keep the same official Qwen3-32B backbone frozen, and use a comparable scale of trainable parameters in LLM-MOReader. Thus, their principal difference is how paired multi-omics information is represented: text-level concatenation for the former versus explicit multi-omics fusion and projection for the latter. As shown in Table 4, Trainable Soft-Prompt improves slightly but significantly inferior to LLM-MOReader, suggesting that the main improvement of LLM-MOReader does not merely come from using more trainable modules or performing more training on specific tasks. Instead, the key factor of multi-omics analysis is the explicit multi-omics fusion and projection paradigm of LLM-MOReader.

Observation 2: Explicit multi-omics fusion and projection is crucial and generalizes well across different LLM backbones. We conduct an ablation study by removing the explicit multi-omics

Table 4: Performance comparison with Trainable Soft-Prompt Model. Values in brackets denote performance improvements computed relative to the Text-level Concatenation baseline model.

Paradigm	Precision	Recall	F1	Accuracy
Text-level Concatenation Model	12.52	12.60	8.56	23.44
Trainable Soft-Prompt Model	15.51 (+2.99)	16.85 (+4.25)	16.15 (+7.59)	37.30 (+13.86)
LLM-MOReader Model	**55.01** (+42.49)	**50.95** (+38.35)	**49.48** (+40.92)	**74.67** (+51.23)

fusion and projection from LLM-MOReader, denoted as **LLM-MOReader w/o Omics Fusion**. As shown in Figure 5, removing omics fusion leads to a clear performance drop, which demonstrates it is crucial for LLM-MOReader. This experiments also demonstrates the LLM backbone-level generality of LLM-MOReader. LLM-MOReader brings improvements on three open-source LLMs with different architectures, model scales, and pretraining characteristics.

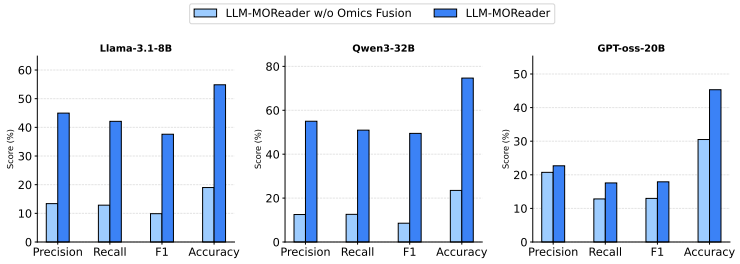


Figure 5: Ablation study and generalization experiments of LLM-MOReader.

Observation 3: The LLM backbone contributes substantially for downstream tasks. To isolate the contribution of the LLM, we replace the projection module and Qwen3-32B backbone in Qwen-MOReader with an MLP while retaining the fusion stage, yielding an **MLP-MOReader** ablation, and evaluate it on both the CTA task and the Mouse Brain (RNA+ADT) MOI task. As shown in Table 5, this replacement reduces Macro F1 and Accuracy on CTA by 17.36 and 19.45, respectively, and consistently degrades MOI performance. Together, these results reveal complementary roles underlying LLM-MOReader: explicit fusion and projection resolves cross-omics heterogeneity before prompting, while the LLM yielding a more biologically informative post-fusion embedding.

Table 5: Ablation of the LLM backbone on both multi-omics cell type annotation and integration.

Model	CTA		MOI				
	Macro F1	Accuracy	NMI	ARI	cASW	cLISI	BVC
MLP-MOReader	32.12 ± 0.17	55.22 ± 0.07	0.62	0.86	0.53	**1.00**	0.75
Qwen-MOReader	**49.48**	**74.67**	**0.73**	**0.92**	**0.59**	**1.00**	**0.81**

5 Conclusion

We introduce LLM-MOReader, a framework that enables LLMs to effectively analyze paired single-cell multi-omics data. An empirical study reveals a counterintuitive "more modalities, worse performance" phenomenon: directly providing paired multi-omics data to LLMs fails to improve performance and can even lead to worse performance than single-omics inputs. To address this challenge, LLM-MOReader shifts from the text-level concatenation paradigm to an explicit multi-omics fusion and projection paradigm. Experiments across multiple species, multi-omics combinations, and tasks show that LLM-MOReader matches or surpasses state-of-the-art baselines. Further analyses show that our approach resolves the "more modalities, worse performance" problem and exhibits model-agnostic generality. LLM-MOReader enables LLMs to read the language of multi-omics, making LLM-based paired single-cell multi-omics analysis practically feasible.

Reproducibility statement

Code for reproducing the model is provided in the supplementary materials, and the datasets used in the experiments—along with details regarding their processing—are presented in the experimental section and the appendix. Further details concerning the model's design and implementation can be found in the appendix.

AI use statement

Generative AI tools assisted in the drafting, revision, and LaTeX formatting of this manuscript. The authors are responsible for reviewing the generated content and verifying the final text, references, and empirical claims.

References

Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. Multivi: deep generative model for the integration of multimodal data. *Nature methods*, 20(8):1222–1231, 2023.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457, 2019.

Mingbo Cheng, Zhijian Li, and Ivan G Costa. Mojitoo: a fast and universal method for integration of multimodal single-cell data. *Bioinformatics*, 38(Supplement_1):i282–i289, 2022.

Jinzhuang Dou, Shaoheng Liang, Vakul Mohanty, Qi Miao, Yuefan Huang, Qingnan Liang, Xuesen Cheng, Sangbae Kim, Jongsu Choi, Yumei Li, et al. Bi-order multimodal integration of single-cell data. *Genome biology*, 23(1):112, 2022.

Jin-Hong Du, Zhanrui Cai, and Kathryn Roeder. Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scvaeit. *Proceedings of the National Academy of Sciences*, 119(49):e2214414119, 2022.

Yin Fang, Xinle Deng, Kangwei Liu, Ningyu Zhang, Jingyang Qian, Penghui Yang, Xiaohui Fan, and Huajun Chen. A multi-modal ai copilot for single-cell analysis with instruction following. *arXiv preprint arXiv:2501.08187*, 2025.

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

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature methods*, 18 (3):272–282, 2021.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, et al. A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology*, 40(2):163–166, 2022.

Samantha M Golomb, Ian H Guldner, Anqi Zhao, Qingfei Wang, Bhavana Palakurthi, Emilija A Aleksandrovic, Jacqueline A Lopez, Shaun W Lee, Kai Yang, and Siyuan Zhang. Multi-modal single-cell analysis reveals brain immune landscape plasticity during aging and gut microbiota dysbiosis. *Cell reports*, 33(9), 2020.

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

Yinlei Hu, Siyuan Wan, Yuanhanyu Luo, Yuanzhe Li, Tong Wu, Wentao Deng, Chen Jiang, Shan Jiang, Yueping Zhang, Nianping Liu, et al. Benchmarking algorithms for single-cell multi-omics prediction and integration. *Nature Methods*, 21(11):2182–2194, 2024.

Yusuke Imoto, Tomonori Nakamura, Emerson G Escolar, Michio Yoshiwaki, Yoji Kojima, Yukihiro Yabuta, Yoshitaka Katou, Takuya Yamamoto, Yasuaki Hiraoka, and Mitinori Saitou. Resolution of the curse of dimensionality in single-cell rna sequencing data analysis. *Life Science Alliance*, 5 (12):e202201591, 2022.

Ana-Maria Istrate, Fausto Milletari, Fabrizio Castrotorres, Jakub M Tomczak, Michaela Torkar, Donghui Li, and Theofanis Karaletsos. rbio1-training scientific reasoning llms with biological world models as soft verifiers. *bioRxiv*, pp. 2025–08, 2025.

Suoqin Jin, Lihua Zhang, and Qing Nie. scai: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome biology*, 21(1):25, 2020.

Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell rna sequencing technologies and applications: a brief overview. *Clinical and translational medicine*, 12(3):e694, 2022.

Hani Jieun Kim, Yingxin Lin, Thomas A Geddes, Jean Yee Hwa Yang, and Pengyi Yang. Citefuse enables multi-modal analysis of cite-seq data. *Bioinformatics*, 36(14):4137–4143, 2020.

Jeongwoo Lee, Do Young Hyeon, and Daehee Hwang. Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine*, 52(9):1428–1442, 2020.

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

Cong Li, Qingqing Long, Yuanchun Zhou, and Meng Xiao. Screader: Prompting large language models to interpret scrna-seq data. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 665–672. IEEE, 2024.

Zhijian Li, Christoph Kuppe, Susanne Ziegler, Mingbo Cheng, Nazanin Kabgani, Sylvia Menzel, Martin Zenke, Rafael Kramann, and Ivan G Costa. Chromatin-accessibility estimation from single-cell atac-seq data with scopen. *Nature communications*, 12(1):6386, 2021.

Mohammad Lotfollahi, Anastasia Litinetskaya, and Fabian J Theis. Multigrate: single-cell multi-omic data integration. *BioRxiv*, 21(1):111, 2022a.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature biotechnology*, 40(1):121–130, 2022b.

Malte D Luecken, Daniel Bernard Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann T Chen, Louise Deconinck, Angela M Detweiler, Alejandro A Granados, et al. A sandbox for prediction and integration of dna, rna, and proteins in single cells. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*, 2021.

Allen W Lynch, Christina V Theodoris, Henry W Long, Myles Brown, X Shirley Liu, and Clifford A Meyer. Mira: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nature methods*, 19(9):1097–1108, 2022.

Anjun Ma, Xiaoying Wang, Jingxian Li, Cankun Wang, Tong Xiao, Yuntao Liu, Hao Cheng, Juexin Wang, Yang Li, Yuzhou Chang, et al. Single-cell biological network inference using a heterogeneous graph transformer. *Nature Communications*, 14(1):964, 2023.

Eleni P Mimitou, Caleb A Lareau, Kelvin Y Chen, Andre L Zorzetto-Fernandes, Yuhan Hao, Yusuke Takeshima, Wendy Luo, Tse-Shun Huang, Bertrand Z Yeung, Efthymia Papalexi, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature biotechnology*, 39(10):1246–1258, 2021.

Moritz Schaefer, Peter Peneder, Daniel Malzl, Salvo Danilo Lombardo, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Celine Sin, et al. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, pp. 1–11, 2025.

Rohit Singh, Brian L Hie, Ashwin Narayan, and Bonnie Berger. Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities. *Genome biology*, 22(1):131, 2021.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

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

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

Xin Tang, Jiawei Zhang, Yichun He, Xinhe Zhang, Zuwan Lin, Sebastian Partarrieu, Emma Bou Hanna, Zhaolin Ren, Hao Shen, Yuhong Yang, et al. Explainable multi-task learning for multi-modality biological data analysis. *Nature communications*, 14(1):2546, 2023.

Ruo Han Wang, Jianping Wang, and Shuai Cheng Li. Probabilistic tensor decomposition extracts better latent embeddings from single-cell multiomic data. *Nucleic acids research*, 51(15):e81–e81, 2023.

Yihang Xiao, Jinyi Liu, Yan Zheng, Shaoqing Jiao, Jianye Hao, Xiaohan Xie, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, et al. Cellagent: Llm-driven multi-agent framework for natural language-based single-cell analysis. *BioRxiv*, pp. 2024–05, 2024.

Fan Zhang, Tianyu Liu, Zhihong Zhu, Hao Wu, Haixin Wang, Donghao Zhou, Yefeng Zheng, Kun Wang, Xian Wu, and Pheng-Ann Heng. Cellverse: Do large language models really understand cell biology? *Advances in Neural Information Processing Systems*, 38, 2026.

Ye Zheng, Daniel P Caron, Ju Yeong Kim, Seong-Hwan Jun, Yuan Tian, Florian Mair, Kenneth D Stuart, Peter A Sims, and Raphael Gottardo. Adtnorm: robust integration of single-cell protein measurement across cite-seq datasets. *Nature communications*, 16(1):5852, 2025.

Jingbo Zhou, Zhongqi Wang, Mutian Hong, Zhuoli Ouyang, Yixuan Du, Siyuan Li, Siqi Ma, Bolin Yang, Changxi Chi, Yunfan Liu, et al. CeLLM: Can large language models achieve the AI virtual cell? Submitted to The Fourteenth International Conference on Learning Representations (ICLR 2026), 2026. URL `https://openreview.net/forum?id=qyS3gtL2Fx`.

A More Experimental Results

A.1 Additional Results on RNA+ATAC Cell Type Annotation

We did not include this part of the experiments in the main text due to the lack of an LLM-based baseline for comparison. Unlike RNA and ADT modalities, which can be naturally described by gene names or protein markers that carry biomedical semantics in natural language, ATAC peak coordinates are not directly language-readable and are difficult for standard LLMs to interpret through text serialization. Therefore, it is difficult to construct strong LLM-based baselines for RNA+ATAC annotation, since naively converting ATAC peaks into coordinate strings does not provide meaningful linguistic tokens for LLM reasoning.

We further evaluate whether LLM-MOReader can be extended to cell type annotation on paired RNA+ATAC data, and additionally compare it with `MultiVI` (Ashuach et al., 2023), a deep generative model designed specifically for paired RNA+ATAC data. As shown in Table 6, Qwen-MOReader obtains 90.86% Accuracy and 59.42% Macro F1 on Human Brain, and 91.63% Accuracy and 43.85% Macro F1 on Mouse Retina, achieving higher Accuracy than MultiVI on both datasets

as well as higher Macro F1 on Mouse Retina. On Human Brain, MultiVI attains higher macro-averaged scores, which is expected given that its architecture is tailored to this specific modality pair. Considering that ATAC peaks provide no directly language-readable tokens, these results indicate that LLM-MOReader remains competitive with a specialized model even on a modality combination that is intrinsically unfavorable to LLM-based methods, and demonstrate that LLM-MOReader can effectively handle RNA+ATAC cell type annotation through explicit multi-omics fusion. This further supports the central motivation of our work: paired multi-omics analysis should not rely on text-level concatenation alone, especially for modalities such as ATAC, but instead requires explicit multi-omics fusion in the omics feature space before LLM reasoning.

Table 6: Additional results on RNA+ATAC cell type annotation. The best result of each metric on each dataset is shown in **bold**.

Dataset	Model	**Accuracy**	**Macro F1**
Human Brain (RNA+ATAC)	MultiVI	88.79	**64.26**
	Qwen-MOReader	**90.86**	59.42
Mouse Retina (RNA+ATAC)	MultiVI	89.45	40.10
	Qwen-MOReader	**91.63**	**43.85**

A.2 Qualitative Dialogue Case of LLM for Paired Multi-Omics Cell Type Annotation.

We randomly selected a multi-omics (RNA+ADT) cell for qualitative dialogue analysis by performing cell type annotation while simultaneously outputting the analysis process; the results are shown in the Figure 6.

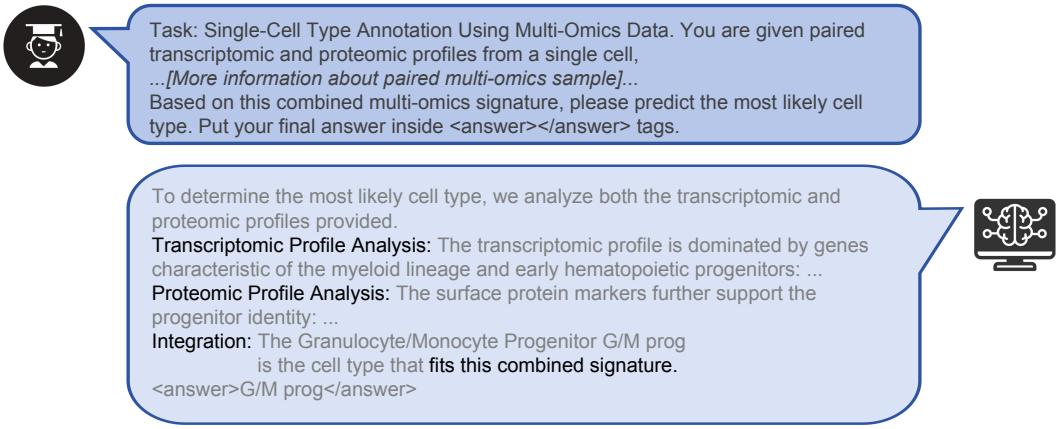


Figure 6: **Qualitative dialogue case for paired multi-omics cell type annotation.** The model integrates multi-omics data to produce an interpretable prediction in natural-language dialog.

A.3 Standard Deviation of Main Experiment

To report error bars and verify the stability of the model, we repeat each experiment **three times** with different random seeds and compute the standard deviation across the three runs. The results are shown in Table 7, with the standard deviation given in brackets.

B Dataset

Overview. We evaluate LLM-MOReader on two representative paired single-cell multi-omics tasks: cell type annotation and multi-omics integration. For cell type annotation, we use the Human BMMC dataset from the Open Problems (Luecken et al., 2021) benchmark. For multi-omics integration, we

Table 7: Standard deviation of the cell type annotation experiment over three runs with different random seeds.

Modality	**Multi-Omics (RNA+ADT)**			
Metric	**Precision**	**Recall**	**Macro-F1**	**Accuracy**
LLM-MOReader				
Llama-MOReader	44.97($\pm$2.12)	42.10($\pm$2.09)	37.60($\pm$1.78)	54.85($\pm$2.55)
Qwen-MOReader	**55.01**($\pm$2.13)	**50.95**($\pm$2.87)	**49.48**($\pm$1.99)	**74.67**($\pm$3.11)

use four public paired multi-omics datasets from Hu et al. (2024), covering both RNA+ADT and RNA+ATAC settings. These datasets span different species, tissues, modality combinations, and downstream tasks, providing a diverse benchmark for evaluating the generalization ability of our method.

Table 8 summarizes the main statistics of all datasets used in this work. For readability, we use short dataset names such as **Human BMMC, Mouse Brain, Human PBMC, Human Brain, and Mouse Retina** in the main text, while retaining the original dataset identifiers and accession numbers in the appendix.

Table 8: Summary of datasets used in this work.

Source	Dataset Name	Accession	Species / Tissue	Modality	Statistics
Luecken et al. (2021)	Human BMMC	GSE194122	Human bone marrow mononuclear cells	RNA+ADT	90,261 cells; 13,953 RNA features; 134 proteins
Golomb et al. (2020)	Mouse Brain	GSE148127	Mouse naive brain	RNA+ADT	13,052 cells; 19,848 RNA features; 34 proteins
Mimitou et al. (2021)	Human PBMC	GSE156478	Human peripheral blood mononuclear cells	RNA+ADT	7,624 cells; 28,310 RNA features; 210 proteins
`https://www.10xgenomics.com/`	Human Brain	-	Human adult brain	RNA+ATAC	2,855 cells; 16,910 RNA features; 132,465 ATAC peaks
Dou et al. (2022)	Mouse Retina	GSE201402	Mouse retina	RNA+ATAC	9,383 cells; 6,275 RNA features; 59,353 ATAC peaks

Dataset details. For the cell type annotation task, we use the **Human BMMC** dataset from the Open Challenge benchmark, corresponding to GSE194122. This dataset is derived from human bone marrow mononuclear cells and contains **90,261 cells** from **12 batches**, with **13,953** transcript features and **134** protein features. The sparsity of the mRNA modality is **0.90**, while the sparsity of the protein modality is **0.19**. We follow the original split protocol provided by the Open Challenge benchmark.

For the multi-omics integration task, we use four public datasets. **Mouse Brain (RNA+ADT)** corresponds to **GSE148127**, derived from mouse naive brain. It contains **13,052 cells**, **19,848** transcript features, and **34** protein features, with mRNA sparsity **0.88** and protein sparsity **0.31**. **Human PBMC (RNA+ADT)** corresponds to **GSE156478**, derived from human peripheral blood mononuclear cells. It contains **7,624 cells**, **28,310** transcript features, and **210** protein features. In

our integration experiments, we use the paired RNA and ADT modalities from this dataset. The mRNA and protein sparsities are **0.93** and **0.48**, respectively. **Human Brain (RNA+ATAC)** derived from human adult brain. It contains **2,855 cells**, **16,910** transcript features, and **132,465** ATAC peaks, with mRNA sparsity **0.91** and ATAC sparsity **0.91**. **Mouse Retina (RNA+ATAC)** corresponds to **GSE201402**, derived from mouse retina. It contains **9,383 cells**, **6,275** transcript features, and **59,353** ATAC peaks, with mRNA sparsity **0.94** and ATAC sparsity **0.95**.

Preprocessing and data splits. We apply modality-specific preprocessing and quality control to each dataset, following the original protocols whenever available. For the **Human BMMC** dataset used in the cell type annotation task, RNA profiles are filtered according to mitochondrial content, UMI counts per cell, and the number of genes detected per cell. For the ADT modality, protein measurements are further quality controlled based on the total ADT counts per cell, the number of proteins detected in each cell, and the summed ADT counts of six isotype controls. Specifically, the total ADT count thresholds vary across samples, the lower bound for the number of detected proteins is set to 80, and the accepted range of summed isotype-control counts is 1–100. For this dataset, we follow the original split protocol provided by the Open Challenge benchmark.

For the **Mouse Brain** RNA+ADT dataset, preprocessing follows the original CITE-seq analysis pipeline implemented with Seurat. The RNA and protein measurements are processed through cell- and gene-level filtering, normalization, demultiplexing, scaling, dimensionality reduction, clustering, and differential expression analysis. This procedure removes low-quality cells and genes and retains high-quality paired RNA and ADT measurements for downstream integration evaluation.

For the **Human PBMC** DOGMA-seq dataset, we perform quality control for each modality and require retained cells to have high-quality measurements across all available modalities. The RNA and ADT modalities are filtered to remove low-quality measurements before constructing paired RNA+ADT inputs. For the ATAC modality in the original DOGMA-seq preprocessing, the top 1% of barcodes from the Cell Ranger ATAC knee call are removed to reduce potential artifacts such as doublets. In this work, we use the paired RNA and ADT modalities from this dataset for the RNA+ADT integration task.

For the **Human Brain** and **Mouse Retina** RNA+ATAC datasets, we filter both transcriptomic and chromatin-accessibility features according to their prevalence across cells. Transcripts expressed in fewer than 10% of cells are removed from the RNA modality, and ATAC peaks detected in fewer than 10% of cells are removed from the ATAC modality. This filtering step reduces extremely sparse and uninformative features, which is particularly important for ATAC data due to its high dimensionality and strong sparsity.

For data splitting, the **Human BMMC** cell type annotation dataset follows the original split protocol of the Open Challenge benchmark. For the remaining four datasets used in the multi-omics integration task, namely **Mouse Brain**, **Human PBMC**, **Human Brain**, and **Mouse Retina**, we split cells into training and testing sets with an 8:2 ratio. This setting ensures that model evaluation is conducted on held-out cells while maintaining a consistent protocol across different species, tissues, and modality pairs.

C Detailed Metric Definitions

In this section, we provide detailed definitions of all evaluation metrics used in our experiments.

Cell type annotation metrics. For the cell type annotation task, we report **Precision**, **Recall**, **F1**, and **Accuracy**. Let y_i and $\hat{y}_i$ denote the ground-truth and predicted cell type labels for cell i, respectively.

Accuracy is defined as the proportion of correctly classified cells:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\hat{y}_i = y_i),$$

where N is the total number of cells and $\mathbf{1}(\cdot)$ is the indicator function.

For each class, **Precision** is defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where TP and FP denote the numbers of true positives and false positives, respectively.

Recall is defined as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where FN denotes the number of false negatives.

F1 score is the harmonic mean of Precision and Recall:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

In our experiments, these class-wise scores are computed following the standard multi-class setting and then aggregated over all cell types. Accuracy and F1 are treated as the primary metric because it provides the most direct measure of overall classification correctness, while Precision, Recall, and F1 offer complementary views on prediction quality under class imbalance.

Multi-omics integration metrics. For the multi-omics integration task, we report **Normalized Mutual Information (NMI)**, **Adjusted Rand Index (ARI)**, **cell-type Average Silhouette Width (cASW)**, **cell-type Local Inverse Simpson's Index (cLISI)**, and **biological variation conservation (BVC)**.

NMI measures the agreement between the clustering assignments obtained from the integrated embedding and the ground-truth cell type labels. Let U denote the predicted clustering and V denote the true cell type partition. NMI is defined as

$$\text{NMI}(U, V) = \frac{2I(U; V)}{H(U) + H(V)},$$

where $I(U; V)$ is the mutual information between U and V, and $H(\cdot)$ denotes entropy. A higher NMI indicates better preservation of biological cell type structure.

ARI also measures the agreement between predicted clusters and ground-truth labels, while correcting for chance. Given two partitions of the same set of cells, ARI is defined as

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]},$$

where RI is the Rand Index. Higher ARI indicates better alignment between the integrated representation and biological cell identities.

cASW evaluates how well cells of the same biological type are grouped together in the integrated embedding. For a given cell i, the silhouette width is

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where $a(i)$ is the average distance between cell i and other cells of the same cell type, and $b(i)$ is the minimum average distance between cell i and cells from another cell type. The cell-type ASW is then obtained by averaging $s(i)$ over all cells. A higher cASW indicates better separation between different biological cell types and stronger within-type compactness.

cLISI measures the local purity of cell types in the integrated embedding using the Local Inverse Simpson's Index. For the neighborhood of each cell, LISI is computed from the local cell-type distribution. Lower diversity of cell types within local neighborhoods corresponds to stronger biological conservation. Following standard practice, we use the cell-type version of LISI, denoted as cLISI, so that a better integrated embedding yields a more favorable cLISI score according to the evaluation protocol.

Finally, **BVC** serves as our primary metric for multi-omics integration. It is defined as the average of the first four metrics:

$$\text{BVC} = \frac{1}{4}\left(\text{NMI} + \text{ARI} + \text{cASW} + \text{cLISI}\right).$$

BVC provides a single summary score for variation conservation by jointly considering cluster-label agreement and cell-type structure preservation in the integrated embedding space.

Interpretation. Overall, NMI and ARI mainly evaluate whether the integrated representation preserves global biological clustering structure, while cASW and cLISI characterize local compactness and neighborhood-level cell-type consistency. Their combination through BVC provides a more comprehensive assessment of biological conservation quality in multi-omics integration.

D Limitations Text-level Concatenation Paradigms

Although paired modalities describe the same underlying cell state, they are highly heterogeneous in both statistical properties and biological semantics. RNA modality (Imoto et al., 2022) are high-dimensional, sparse, and affected by dropout and sequencing-depth variation; ATAC modality (Li et al., 2021) are typically even higher-dimensional and sparser, reflecting chromatin accessibility and regulatory potential; ADT modality (Zheng et al., 2025) are much lower-dimensional, denser, and more directly associated with cell surface phenotypes.

Previous studies typically adopt a *cell-to-sentence* (Levine et al., 2024) strategy, in which paired omics measurements are converted into textual descriptions and then fed into the LLM as hard prompts. Formally, given the paired multi-omics input $(X_c^{(1)}, X_c^{(2)})$, a hand-crafted serialization function

$$\phi_{\text{hard}} : \left(X_c^{(1)}, X_c^{(2)}\right) \mapsto S_c$$

is constructed to transform the structured omics measurements into a discrete text sequence $S_c = (w_{c,1}, w_{c,2}, \ldots, w_{c,T_c})$, where each $w_{c,t}$ is a token from the vocabulary of the LLM and T_c denotes the token length of the resulting hard prompt. The final input to the LLM is then formed as

$$\widetilde{S}_c = \text{Concat}(T_{\text{task}}, S_c),$$

where T_{task} denotes the textual task instruction. For example, S_c may consist of the names of highly expressed genes, accessible peaks, or non-zero proteins associated with cell c.

While this strategy enables the use of off-the-shelf LLMs, it has important limitations, especially for paired multi-omics inputs. First, text serialization inevitably discards fine-grained quantitative information. The original omics measurements $X_c^{(1)} \in \mathbb{R}^{d_1}$ and $X_c^{(2)} \in \mathbb{R}^{d_2}$ are continuous, structured, and high-dimensional, whereas the hard prompt S_c is a discrete token sequence. Therefore, the mapping ϕ_{hard} is generally many-to-one, i.e.,

$$\exists \left(X_c^{(1)}, X_c^{(2)}\right) \neq \left(X_{c'}^{(1)}, X_{c'}^{(2)}\right) \quad \text{s.t.} \quad \phi_{\text{hard}}\left(X_c^{(1)}, X_c^{(2)}\right) = \phi_{\text{hard}}\left(X_{c'}^{(1)}, X_{c'}^{(2)}\right),$$

which implies that distinct quantitative omics states may be mapped to the same textual description. As a result, magnitude relationships among features may be weakened or lost, and the original biological signal may be distorted during serialization.

Second, hard prompt construction treats different modalities as independent text fragments rather than jointly modeled biological observations. In the cell-to-sentence paradigm, the paired input is typically serialized as

$$S_c = \text{Concat}\big(\phi_1(X_c^{(1)}), \phi_2(X_c^{(2)})\big),$$

where ϕ_1 and ϕ_2 are modality-specific text serialization rules. Under this formulation, the two modalities are only concatenated in the token space, without explicit fusion in the representation space. That is, there is no joint latent variable z_c such that

$$z_c = f_{\text{joint}}(X_c^{(1)}, X_c^{(2)})$$

is learned before the information is passed to the LLM. Consequently, complementary information across modalities is not explicitly integrated prior to downstream reasoning, which is particularly problematic for paired multi-omics analysis, where the key value lies precisely in modeling different views of the same cell jointly.

Third, LLMs are pretrained on natural language rather than on noisy serialized multi-omics descriptions. Under the hard prompt formulation, the LLM is required to directly learn

$$p(y_c \mid \widetilde{S}_c) = p\big(y_c \mid \text{Concat}(T_{\text{task}}, S_c)\big),$$

or, for representation learning,

$$h_c = f_{\mathrm{LLM}}(\widetilde{S}_c),$$

where h_c denotes the resulting cell embedding. When S_c is formed by naively concatenating heterogeneous modalities, the LLM must implicitly resolve distribution mismatch, modality conflict, and technical noise purely in the language space. In other words, the burden of multi-omics fusion is shifted to the natural language interface:

$$(X_c^{(1)}, X_c^{(2)}) \xrightarrow{\phi_{\mathrm{hard}}} S_c \xrightarrow{\mathrm{LLM}} \hat{y}_c \text{ or } h_c.$$

In practice, this often leads to degraded performance, and in some cases paired multi-omics hard prompts may even underperform single-omics inputs.

To address these limitations, we depart from the cell-to-sentence paradigm. Instead of converting raw multi-omics matrices into hard prompts, we preserve the original omics information in matrix form and first perform joint modeling in the omics space. Specifically, we learn a **joint multi-omics representation**

$$z_c = f_{\mathrm{enc}}(X_c^{(1)}, X_c^{(2)}),$$

where $z_c \in \mathbb{R}^{d_z}$ captures the fused biological information of cell c. This representation is then transformed into a **joint multi-omics prompt**

$$P_c = f_{\mathrm{proj}}(z_c), \qquad P_c \in \mathbb{R}^{L \times d},$$

which serves as a soft prompt compatible with the LLM input space. In this way, the burden of multi-omics fusion is handled in the omics representation space, where it is more natural and effective, rather than in the natural language space.

E Rationale for Modality-specific Likelihoods

In our variational multi-omics encoder-decoder framework, different omics modalities are modeled with different likelihood functions, inspired by previous work (Gayoso et al., 2021; Ashuach et al., 2023). This design is motivated by the fact that RNA, ATAC, and ADT measurements have substantially different statistical properties. Instead of forcing all modalities to share the same observation model, we specify a modality-specific likelihood for each data type, allowing the decoder to better capture the data-generating characteristics of each modality.

RNA likelihood. For the RNA modality, each cell is represented by a vector of gene expression counts. These counts are discrete, non-negative, high-dimensional, and often highly sparse. Moreover, scRNA-seq data typically exhibit overdispersion, meaning that the variance of gene expression counts is larger than the mean. This overdispersion arises from both biological variability and technical factors, such as sequencing depth variation, capture efficiency, and dropout events. A simple Poisson likelihood assumes equal mean and variance and is therefore often too restrictive for scRNA-seq data. By contrast, the Negative Binomial distribution introduces an additional dispersion parameter, making it more suitable for modeling overdispersed count data. Therefore, we model RNA counts as

$$X_c^{\mathrm{RNA}} \mid z_c \sim \mathrm{NB}\big(\mu_c^{\mathrm{RNA}}, \theta^{\mathrm{RNA}}\big),$$

where μ_c^{RNA} is predicted from the latent representation z_c, and θ^{RNA} controls the gene-wise dispersion. This likelihood encourages the latent variable to preserve the transcriptional signal required to reconstruct RNA expression while accounting for count variability and technical noise.

ATAC likelihood. For the ATAC modality, each feature corresponds to the accessibility status of a genomic region or chromatin peak. In many single-cell ATAC-seq preprocessing pipelines, the accessibility matrix is represented as a binary or near-binary matrix, where each entry indicates whether a genomic region is accessible in a given cell. Compared with RNA, ATAC data are usually even sparser and have a much larger feature dimension, with most genomic regions being inaccessible in most cells. Therefore, the key observation for each peak is naturally interpreted as a binary accessibility event.

Based on this property, we model ATAC measurements using a Bernoulli likelihood:

$$X_c^{\mathrm{ATAC}} \mid z_c \sim \mathrm{Bernoulli}\big(\pi_c^{\mathrm{ATAC}}\big),$$

where π_c^{ATAC} denotes the probability that each genomic region is accessible, as predicted from the latent representation z_c. This formulation matches the binary nature of chromatin accessibility data and allows the decoder to reconstruct regulatory activity as peak-wise accessibility probabilities. Compared with treating ATAC features as continuous values, the Bernoulli likelihood better reflects the sparse open-or-closed structure of chromatin accessibility measurements.

ADT likelihood. For the ADT modality, each feature represents the abundance of an antibody-derived tag corresponding to a surface protein. Although ADT data are usually much lower-dimensional and denser than RNA or ATAC data, they are still generated as sequencing-based count measurements. Similar to RNA counts, ADT counts are discrete, non-negative, and can exhibit overdispersion due to biological variation, antibody capture efficiency, sequencing depth, and background noise. Therefore, a Gaussian likelihood is not ideal because it treats the observations as continuous and does not naturally account for count overdispersion.

We thus model ADT abundance using a Negative Binomial likelihood:

$$X_c^{\mathrm{ADT}} \mid z_c \sim \mathrm{NB}\big(\mu_c^{\mathrm{ADT}}, \theta^{\mathrm{ADT}}\big)\,,$$

where μ_c^{ADT} is the protein-specific mean predicted from z_c, and θ^{ADT} denotes the corresponding dispersion parameter. This likelihood allows the model to capture the count-based and overdispersed nature of ADT measurements while preserving protein-level phenotypic information.

Why modality-specific likelihoods are important. The use of modality-specific likelihoods is central to our explicit multi-omics fusion design. RNA, ATAC, and ADT differ not only in dimensionality and sparsity, but also in the type of biological signal they encode: RNA reflects transcriptional output, ATAC captures chromatin accessibility and regulatory potential, and ADT directly measures surface protein abundance. Using a single generic reconstruction loss for all modalities would ignore these differences and may cause the latent representation to be dominated by the modality with the largest dimension, strongest signal, or lowest noise. In contrast, modality-specific likelihoods allow each modality to be reconstructed according to its own statistical structure.

Under this formulation, the shared latent variable z_c must explain all observed modalities, while each decoder models the observation process of its corresponding modality. This creates a principled fusion mechanism: cross-modal biological information is integrated through the shared latent representation, whereas modality-specific noise is handled by the corresponding likelihood model. As a result, the learned z_c provides a denoised and biologically meaningful fused multi-omics representation, which is better suited for constructing the joint multi-omics prompt used by the LLM.

F Baseline Model Sources

We provide the release dates and official sources of the evaluated LLM baselines in Table 9. We organize the baselines into four categories according to their experimental roles. First, we include **frontier large-scale LLMs**, such as **GPT-5**, **Gemini 3 Flash**, **Gemini 2.5 Pro**, **Gemini 2.5 Flash**, **Qwen 3.6 Plus**, **GLM-5.1**, **Kimi k2.6**, and **DeepSeek V4 Flash**. This group is used to probe the empirical performance upper bound of current LLMs on paired multi-omics analysis. Second, we include **open-source small-scale general LLMs**, including the **Qwen3** series, **GPT-oss-20B**, **GPT-oss-120B**, and **DeepSeek-V3.1**. These models evaluate whether deployable open-source LLM backbones can effectively exploit paired omics information and whether LLM-MOReader can generalize across different model families and scales. Finally, we include **domain-specific LLMs**, such as **cell-o1** and **rBio1**, to examine whether existing biological or single-cell language models can be directly transferred to paired single-cell multi-omics tasks. For all LLM baselines, we use the same task instruction template and the same multi-omics input format whenever applicable, ensuring that performance differences mainly reflect model capability.

We further compare LLM-MOReader with a broad set of classical single-cell multi-omics methods. The source and implementation of all classical multi-omics models follow Hu et al. (2024). These baselines are designed for paired multi-omics integration, modality alignment, representation learning, imputation, or regulatory analysis. We briefly summarize the technical characteristics and typical application scenarios of each method.

Table 9: The release dates and sources of the evaluated models.

Model Name	Release Date	Official Source
Frontier Large-scale LLMs		
Deepseek-V4-Flash	2026-04	https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash
Kimi-K2.6	2026-04	https://huggingface.co/moonshotai/Kimi-K2.6
GLM-5.1	2026-04	https://z.ai/blog/glm-5.1
Qwen-3.6-Plus	2026-03	https://qwen.ai/blog?id=qwen3.6
Gemini-3-Flash	2025-12	https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview
GPT-5	2025-08	https://platform.openai.com/docs/models/gpt-5
Gemini-2.5-pro	2025-06	https://ai.google.dev/gemini-api/docs/models#gemini-2.5-pro
Gemini-2.5-flash	2025-06	https://ai.google.dev/gemini-api/docs/models#gemini-2.5-flash
Domain-specific LLMs		
rBio1	2025-08	https://virtualcellmodels.cziscience.com/model/rbio
Cell-o1	2025-06	https://huggingface.co/ncbi/Cell-o1
Open-source Small-scale LLMs		
Deepseek-V3.1	2025-09	https://huggingface.co/deepseek-ai/DeepSeek-V3.1
GPT-oss-120B	2025-08	https://huggingface.co/openai/gpt-oss-120b
GPT-oss-20B	2025-08	https://huggingface.co/openai/gpt-oss-20b
Qwen3-4B	2025-08	https://huggingface.co/Qwen/Qwen3-4B
Qwen3-8B	2025-08	https://huggingface.co/Qwen/Qwen3-8B
Qwen3-14B	2025-08	https://huggingface.co/Qwen/Qwen3-14B
Qwen3-30B-A3B	2025-08	https://huggingface.co/Qwen/Qwen3-30B-A3B
Qwen3-32B	2025-08	https://huggingface.co/Qwen/Qwen3-32B

- **Seurat** (Stuart et al., 2019). Seurat is a widely used single-cell analysis framework that supports multimodal data integration through weighted nearest neighbor (WNN) analysis. WNN learns the relative contribution of each modality for each cell and constructs a joint neighborhood graph for downstream analysis.
- **MOJITOO** (Cheng et al., 2022). MOJITOO is a fast and broadly applicable multimodal integration method based on canonical correlation analysis. It identifies a shared low-dimensional representation across modalities and provides interpretable canonical components that can be associated with modality-specific molecular features.
- **totalVI** (Gayoso et al., 2021). totalVI is a deep generative probabilistic model designed for CITE-seq data. It jointly models RNA counts and protein abundance while accounting for technical factors such as sequencing depth, protein background, and batch effects.
- **Multigrate** (Lotfollahi et al., 2022a). Multigrate is a generative multi-view neural network for single-cell multi-omic data integration. It is designed to construct multimodal reference atlases.
- **SCOIT** (Wang et al., 2023). SCOIT is a probabilistic tensor decomposition framework. It constructs a omic tensor from multiple modalities and decomposes it under modality-appropriate probabilistic assumptions, allowing it to handle sparse, noisy, and heterogeneous single-cell measurements.
- **CiteFuse** (Kim et al., 2020). CiteFuse is a computational framework specifically designed for CITE-seq data, integrating RNA and ADT modalities for joint single-cell analysis.
- **DeepMAPS** (Ma et al., 2023). DeepMAPS is a deep learning-based multi-omics analysis platform that models single-cell multi-omics data using a heterogeneous graph transformer. It represents cells and genes as nodes in a heterogeneous graph and learns their relations through graph neural networks and attention mechanisms.
- **scArches** (Lotfollahi et al., 2022b). scArches is a transfer-learning framework for mapping query single-cell datasets onto reference atlases. It performs architectural surgery on trained neural-network models, enabling efficient reference updating and query-to-reference mapping without retraining the entire model from scratch.
- **scVAEIT** (Du et al., 2022). scVAEIT is a probabilistic framework for single-cell multimodal mosaic integration and imputation. It is designed to model incomplete multimodal measurements and can impute missing modalities or missing features while preserving biological variation.

- **scAI** (Jin et al., 2020). scAI is an unsupervised method for integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. It learns low-dimensional factors by aggregating sparse epigenomic signals from similar cells and combining them with transcriptomic measurements, thereby addressing the sparsity and near-binary nature of single-cell epigenomic data.
- **MultiVI** (Ashuach et al., 2023). MultiVI is a deep generative model for multimodal single-cell integration. It jointly models gene expression and chromatin accessibility, supports both paired and unpaired observations, and learns a shared latent representation while accounting for modality-specific statistical properties.
- **MIRA** (Lynch et al., 2022). MIRA is a probabilistic multimodal framework for joint regulatory modeling of gene expression and chromatin accessibility in single cells. It combines cell-level topic modeling with gene-level regulatory potential modeling to infer interpretable cell-state representations and regulatory mechanisms.
- **totalANVI** (Gayoso et al., 2022). totalANVI is the annotation-aware extension of totalVI implemented in `scvi-tools`. It augments the totalVI generative model with a cell-type latent variable, so that available cell-type labels can be used in a semi-supervised manner to shape the joint RNA–protein latent space.
- **UnitedNet** (Tang et al., 2023). UnitedNet is an explainable multi-task deep neural network for single-cell multi-modality data. It jointly optimizes cross-modal prediction and shared-representation learning within one network, and uses feature-attribution analysis to identify the modality-specific features that drive each predicted cell state.
- **Schema** (Singh et al., 2021). Schema is a metric-learning framework for interpretable synthesis of heterogeneous single-cell modalities. It identifies informative features in one modality to guide the integration and interpretation of another modality, producing a coherent representation that preserves cross-modal biological signals.

G Implementation Details

We implement LLM-MOReader based on open-source LLMs such as Qwen3-32B. The fused multi-omics representation dimension is treated as a hyperparameter and searched within the range of 8 to 512. The projection module is implemented by multi-layer MLP, and the activation function uses GELU function. In the MOI task, the cell embedding is derived from the feature pooling of the penultimate layer of the LLM backbone.

We train LLM-MOReader using DeepSpeed ZeRO for memory-efficient distributed training. All main experiments are conducted on 8 NVIDIA A100 GPUs with 80GB memory. The experiment uses AdamW optimizer for optimization. Unless otherwise specified, the model is trained for 30 epochs with a learning rate of 2×10^{-3}. The training batch size of each GPU is 2 (for qwen3-32b backbone) or 8 (for llama-3.1-8b and GPT OSS 20b). The training time is up to 87 hours in cell type annotation and nearly 25 hours in integration. The training time increases linearly with the amount of training data. The multi omics encoder decoder will stop early if there is no decrease in loss during continuous training for 5 epochs. We use a cosine learning-rate scheduler, bfloat16 mixed precision, TF32 acceleration, and gradient checkpointing to improve memory efficiency. The effective batch size is controlled by the per-device batch size and gradient accumulation steps. During inference, we run LLM-MOReader on a single NVIDIA A100 GPU with 80GB memory.

The Trainable Soft-Prompt Model employs a projection module and training/testing datasets identical to those of the LLM-MOReader. The distinction lies in the use of text-level concatenation in the former, whereas the latter utilizes explicit multi-omics fusion. This variant is designed to investigate whether additional training alone can achieve results comparable to those obtained through explicit multi-omics fusion and projection.

G.1 Runtime Analysis

Table 10 reports wall-clock runtime over the same 16,750-sample workload under the model execution setting above. Llama-MOReader completes it in 2 h 21 min 43 s (0.51 s per sample), whereas Qwen-MOReader requires 7 h 50 min 20 s (1.68 s per sample).

Table 10: Runtime comparison of LLM-MOReader instantiations over the same 16,750-sample workload; lower is better.

Model	Samples	Total Runtime	Runtime per Sample
Llama-MOReader	16,750	2 h 21 min 43 s	0.51 s/sample
Qwen-MOReader	16,750	7 h 50 min 20 s	1.68 s/sample

Qwen-MOReader therefore uses approximately $3.32\times$ as much wall-clock time as Llama-MOReader. This cost accompanies its stronger predictive results, revealing a practical trade-off: Qwen-MOReader prioritizes performance, whereas Llama-MOReader offers lower runtime. Because both instantiate the same explicit multi-omics fusion paradigm, the framework's core contribution is not tied to one backbone scale; practitioners can select a model for their accuracy–compute budget without redesigning the fusion module.

MLP-MOReader: In terms of model architecture, this variant replaces the projection head and LLM backbone of LLM-MOReader with an MLP, while using the exact same training data. This variant aims to investigate whether the LLM backbone plays a critical role in the task.

H Societal Impacts

LLM-MOReader may have positive societal impacts by making paired single-cell multi-omics analysis more effective and accessible. By combining explicit multi-omics fusion with LLM reasoning, it can help researchers better analyze heterogeneous cellular information from RNA, ATAC, and protein modalities, potentially accelerating studies in immunology, cancer biology, developmental biology, and precision medicine. The negative impacts of this research may include malicious use—for instance, analyzing single-cell data or conducting studies that could be harmful to society. Potential negative impacts can be mitigated through the gated release of models.