

# CELLM: CAN LARGE LANGUAGE MODELS ACHIEVE THE AI VIRTUAL CELL?

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

High-throughput single-cell sequencing has enabled large-scale cellular profiling and spurred the development of single-cell foundation models. These models, typically pretrained on transcriptomic data, learn general-purpose cellular representations but remain limited in modality coverage, causal reasoning, and interpretability, thus falling short of the vision of an Artificial Intelligence Virtual Cell (AIVC). In parallel, large language models (LLMs) have demonstrated strong potential for unifying heterogeneous modalities, adapting to diverse tasks, and generating interpretable reasoning chains in natural language, making them promising candidates toward AIVC. Recent progress in applying LLMs to tasks such as cell annotation and perturbation prediction highlights this potential, yet key challenges persist, including insufficient task coverage, narrow evaluation metrics, and limited robustness to input and prompting factors. To address these gaps, we introduce **CeLLM**, a comprehensive benchmarking framework for evaluating LLMs in the **CeLLular** domain. CeLLM covers a broad spectrum of tasks spanning gene, cell, and omics-level analyses, systematically assesses 15 open-source, proprietary, and biology-specialized models, and incorporates diverse evaluation criteria under multiple task settings. As a cross-scale, reproducible, and dynamic benchmark, CeLLM provides a sustainable platform to track progress, foster methodological innovation, and accelerate the development of LLMs toward virtual cell modeling.

## 1 INTRODUCTION

High-throughput single-cell sequencing technologies have profoundly advanced our ability to dissect cellular systems (Stuart & Satija, 2019; Lähnemann, 2020). They enable researchers to construct cell atlases at tissue and disease levels, supporting a wide range of downstream analyses such as cell type annotation (Stuart, 2019; Klein, 2015), trajectory inference (Trapnell, 2013; Saelens, 2019), perturbation screening (Chen, 2015), and cross-study integration (Butler, 2018; Luecken, 2022). In this context, single-cell foundation models (Cui, 2024; Zeng, 2025; Hao, 2024) have recently emerged. Typically pretrained on large-scale single-cell transcriptomic data, these models treat genes as “tokens” and cells as “sentences”, thereby learning general-purpose cell representations that demonstrate potential across diverse tasks. Their advent marks a paradigm shift in single-cell analysis, from task-specific algorithms toward generalizable representation learning.

Meanwhile, the concept of the Artificial Intelligence Virtual Cell (AIVC) (Bunne, 2024) has gained increasing attention. The vision is to construct a multi-scale digital surrogate that can simulate and predict cellular behaviors in silico. Unlike models restricted to a single modality, AIVC emphasizes cross-modal integration, simultaneously absorbing information from transcriptomics, epigenomics, proteomics, metabolomics, spatial omics, and imaging data (Qian, 2025; Yang, 2025). Such integration enables modeling of molecular, cellular, and microenvironmental dynamics in a unified framework. Crucially, the core objective of AIVC is not merely to fit existing data, but to enable causal reasoning and counterfactual prediction under perturbations—for example, forecasting how cellular states evolve under drug treatment or gene editing. In addition, AIVC must be interpretable and interactive: beyond delivering predictions, it should expose its reasoning chains in a human-comprehensible manner and allow researchers to query it using natural language, thereby closing the loop between computation and experimentation. As such, AIVC is increasingly regarded as a key direction for transforming single-cell data-driven research into knowledge generation and experimental design support.

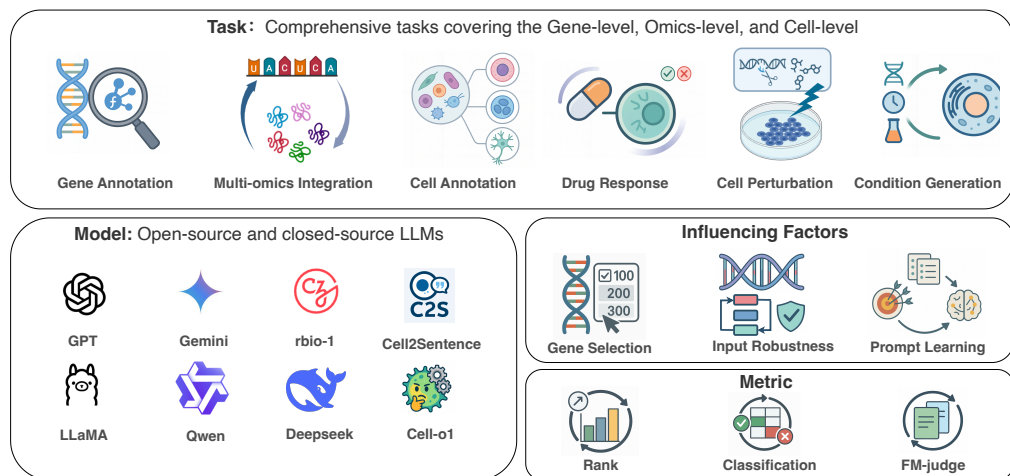


Figure 1: The overview of the CeLLM benchmark. The benchmark is organized incrementally from datasets to models, to influencing factors, to metrics, covering six major tasks and 15 open-source and closed-source models.

Despite recent progress, current single-cell foundation models remain far from fulfilling this vision. Their input layer typically relies on a single modality (e.g., scRNA-seq (Kedzierska, 2025)), limiting their ability to capture multi-omics signals; their discrete representations often lose information on expression intensity and co-expression patterns. At the task level, these models are usually optimized for a single objective, with insufficient zero-shot or few-shot generalization, requiring fine-tuning to adapt to new tasks and unable to simultaneously support multiple downstream applications within a unified system. From a modeling perspective, they largely learn correlational representations, lacking causal reasoning capabilities and thus unable to predict counterfactual outcomes under perturbations. In terms of output, most models only provide embeddings or labels, offering limited interpretability, no reasoning traces, and no natural language interface for direct scientific interaction. These limitations constrain their ability to serve as the foundation for virtual cells.

In contrast, large language models (LLMs) exhibit unique advantages that make them promising candidates toward realizing AIVC. By leveraging language as a universal representation, LLMs can map heterogeneous modalities into a shared semantic space, supporting cross-modal integration. They also demonstrate strong few-shot and zero-shot learning capabilities, enabling rapid adaptation to new tasks without extensive fine-tuning, and can handle multiple tasks under a single framework. Furthermore, LLMs internalize prior knowledge from vast literature and databases, allowing them to generate causal hypotheses about perturbations when combined with single-cell data, while producing dynamic reasoning chains in natural language. This not only enhances interpretability and usability but also facilitates interdisciplinary communication. Importantly, the knowledge coverage of LLMs extends beyond any single training dataset, enabling natural fusion of experimental data with external knowledge to support cross-domain reasoning. Notably, recent advances have shown the power of LLMs in molecular and protein sciences—for example, in small-molecule property prediction, drug generation, and protein structure (Ma, 2025; Xiao, 2025b; Liu, 2024; Xiao, 2025a) and function modeling—providing further evidence of their potential for tackling biological problems and laying the groundwork for extending their application to cellular-scale virtual modeling.

Recently, large language models (LLMs) have achieved remarkable progress in downstream tasks such as cell annotation and perturbation prediction in the field of cell biology. However, several key challenges remain, which severely limit further advances in this area:

- **Comprehensive task coverage for evaluation.** Current applications of LLMs in the cellular domain are mainly limited to tasks such as cell annotation and perturbation prediction. There is a lack of exploration for tasks such as multi-omics integration analysis, gene annotation, and both gene-level and omics-level tasks. Moreover, models applied to different

tasks have not yet been compared under a unified framework. For instance, `cell-o1` (Fang, 2025a) has been evaluated on batch-level cell annotation, but its performance on single-cell annotation remains unclear.

- **Comprehensive evaluation metrics.** While existing studies provide metrics for cell type annotation and perturbation prediction, current evaluation standards are often limited to accuracy, F1-score, and related measures. For tasks such as conditional cell generation and perturbation gene signature prediction, there is a lack of holistic evaluation metrics, making it difficult to comprehensively assess the biological relevance of model outputs.
- **Robustness to key influencing factors.** At the input level, transforming cellular expression values into textual representations suitable for LLMs can be affected by multiple factors, such as the length of selected genes and noise-induced gene ordering changes. At the downstream task level, it remains unclear whether different prompting frameworks or few-shot prompting strategies influence LLM performance across tasks, as this has not yet been systematically explored.

To address these challenges, we propose a new comprehensive LLM benchmarking framework for the cellular domain, named **CeLLM**. This benchmark unifies the evaluation of different LLMs on a wide spectrum of cellular tasks, ranging from gene annotation and cell type annotation to drug response prediction and cell perturbation prediction. CeLLM systematically evaluates 15 models, including both open-source and proprietary LLMs as well as biology-specialized foundation models, covering diverse indicators from simple classification to foundation model assessments. Furthermore, CeLLM characterizes performance under multiple task settings influenced by key factors. Designed as a cross-scale, reproducible, and future-oriented dynamic benchmark, CeLLM will continuously evolve with the emergence of new models, tasks, and datasets, serving as a long-term sustainable evaluation platform that promotes the development of LLMs toward virtual cells and lays the foundation for subsequent methodological innovations.

## 2 RELATED WORK

### 2.1 LARGE LANGUAGE MODELS FOR SCIENTIFIC PROBLEMS

In recent years, large language models (LLMs) have demonstrated broad potential in solving scientific problems across diverse domains such as mathematics, chemistry, and biomedicine. Research in this direction has primarily focused on scientific reasoning and domain adaptation. For example, OmniScience (Prabhakar, 2025) leverages domain-adaptive pretraining and instruction tuning to improve performance on scientific discovery and cross-disciplinary reasoning tasks; MedBioLM (Kim, 2025) integrates fine-tuning with retrieval-augmented generation (RAG) to enhance factual consistency and contextual understanding in medical and biomedical question answering. These studies highlight that LLMs are not only capable of processing scientific text but also hold promise for empowering complex research tasks. To systematically evaluate LLMs in more specialized scientific domains, several new benchmarks have recently been introduced. CURIE (Cui, 2025) proposes a multi-task benchmark for scientific long-context understanding and reasoning, spanning materials science, quantum computing, and protein studies, in order to assess model performance within realistic scientific workflows. PhySense (Xu, 2025) focuses on principle-based reasoning in physics, examining whether LLMs can adhere to scientific laws while generating explanatory answers. QCBench (Xie, 2025) introduces a quantitative chemistry benchmark designed to evaluate LLMs’ reasoning capabilities and computational accuracy in quantitative chemical problems. In particular, at the single-cell level, how to effectively harness LLMs for data interpretation and knowledge discovery remains an open question. This study specifically targets this gap, aiming to systematically evaluate and advance the application of LLMs in cellular tasks.

### 2.2 LARGE LANGUAGE MODELS FOR CELL BIOLOGY TASKS

With the rapid progress of LLMs in natural language processing, researchers have begun to explore their potential in cell biology, particularly for addressing challenges in single-cell data analysis. For instance, Cell2Sentence (C2S) (Biderman, 2025) transforms single-cell transcriptomic data into “cell sentences”, enabling LLMs to interpret and process gene expression profiles for tasks such as cell type prediction and data generation. `rbio1` (Ye, 2024) leverages reinforcement learning from a

pretrained LLM, using biological world models (e.g., Gene Ontology databases) as soft verifiers to incorporate biological knowledge, thereby demonstrating the feasibility of training strong biological reasoning models with simulation-based rather than purely experimental data. In the direction of virtual cell modeling, CellForge (Chen, 2024) introduces a multi-agent framework that automates the design and optimization of virtual cell models, improving predictions of cellular responses to perturbations such as gene editing and drug treatment. cell-o1 (Fang, 2025a) reformulates cell type annotation of single-cell RNA sequencing data into a batch-level reasoning problem. PerturbQA (Wang, 2025) proposes a benchmark dedicated to perturbation experiments, covering tasks such as predicting gene expression changes, performing gene set enrichment analysis, and identifying expression directionality shifts. CellVerse (Fang, 2025b) introduces a language-driven question answering benchmark that integrates multiple single-cell datasets, aiming to evaluate LLMs’ ability to operate across cell biology tasks. This benchmark highlights persistent performance gaps in handling complex perturbation and drug response prediction, underscoring both the potential and limitations of current LLMs in cellular applications. While these efforts have made notable progress in applying LLMs to cell biology, they generally suffer from limited task coverage, insufficient baseline models, incomplete evaluation metrics, and a lack of systematic exploration of key influencing factors. These challenges motivate our work, which introduces a new framework to comprehensively evaluate LLMs in cellular tasks and to advance their applications in this domain.

### 3 MODEL

In CeLLM, we evaluate three categories of models: closed-source LLMs, open-source LLMs, and domain-specific LLMs to comprehensively assess the performance of large language models.

#### 3.1 CLOSED-SOURCE GENERAL LLMS

Closed-source general models, including OpenAI GPT-5 and Google Gemini 2.5, demonstrate strong generalization, long-context handling, and tool-use abilities. GPT-5 has 1.8 trillion parameters with sparse activation, using a dual-core architecture and dynamic routing to select sub-models per task. Gemini 2.5 utilizes a sparse Mixture-of-Experts (MoE) Transformer with dynamic routing to assign tokens to expert subsets, decoupling model capacity from per-token cost. Both are trained on large-scale corpora with supervised fine-tuning and RLHF, and offer reasoning modes for multi-step inference (OpenAI, 2025; Shazeer, 2025).

#### 3.2 DOMAIN-SPECIFIC LLMS

Domain-specific models, including cell-o1, rBio1, and C2S, target single-cell and biological reasoning tasks. Cell-o1, built on a 7B-parameter Qwen2.5-7B-Instruct, performs batch-level scRNA-seq cell type annotation (CellPuzzles) via multi-stage training: reasoning distillation, supervised fine-tuning, and reinforcement learning, achieving joint batch reasoning and unique label assignment (Fang, 2025a). rBio1, based on a lightweight 3B-parameter Qwen2.5, combines multiple soft verifiers with RL to distill biological knowledge, enabling gene perturbation prediction and cross-cell-line generalization (Ye, 2024). C2S uses decoder-only Transformers (GPT-2 / Pythia), fine-tuned on concatenated cell sentences and natural language prompts (Biderman, 2025; Rizvi et al., 2025).

#### 3.3 OPEN-SOURCE GENERAL LLMS

Open-source general-purpose models in our benchmark include DeepSeek-V3.1, the gpt-oss series, Llama-3.1-8B, and the Qwen3 series. DeepSeek-V3.1 adopts a mixture-of-experts (MoE) architecture combined with multi-level attention (MLA), low-precision computation, and hybrid inference mechanisms, leading to significant improvements over V3 in long-context modeling, tool use, and reasoning ability (Abdin, 2024). The gpt-oss series (120B/20B) employs autoregressive MoE Transformers with alternating banded and global attention, and follows a training pipeline consisting of large-scale pretraining and post-training with instruction tuning and preference optimization (OpenAI, 2025). Llama-3.1-8B, a decoder-only Transformer, extends context length through RoPE scaling and is trained in multiple stages, including pretraining, continual domain-adaptive pretraining, instruction fine-tuning, and preference alignment, thereby supporting long-text comprehension and generation (Dubey, 2025). The Qwen3 series covers a wide range of model scales, from 4B to

216 32B, while providing both thinking and non-thinking model variants to explicitly assess the role of  
217 reasoning in downstream tasks (Team, 2025).

## 218 219 4 CELLM TASK

220 In this work, we select six representative cell-level tasks to comprehensively evaluate the potential  
221 of large language models (LLMs) in realizing the vision of AI for Cell Biology (AIVC). These tasks  
222 span the genetic, multi-omics, and cellular levels.

223 **Gene Annotation.** The gene annotation task requires the model to describe the major functions and  
224 biological roles of a given gene. We curated a benchmark set of approximately 50 genes based on  
225 the Gene Ontology (GO) database. Since the GO database may already be present in the pretraining  
226 corpus of LLMs, this task has a certain “recall” nature. It allows us to probe whether the model has  
227 acquired fundamental molecular biology knowledge, serving as a prerequisite reference for more  
228 complex downstream tasks.

229 **Multi-omics Integration.** The multi-omics integration task requires the model to generate com-  
230 prehensive biological interpretations or predictions by leveraging different data modalities, such as  
231 genomics, transcriptomics, epigenomics, and proteomics. This task simulates real-world research  
232 scenarios where multidimensional data must be jointly analyzed. It evaluates the ability of LLMs to  
233 perform cross-modal knowledge fusion and reasoning, thereby assessing whether they can extract  
234 patterns from heterogeneous sources and produce coherent biological insights.

235 **Cell Annotation.** Cell annotation is the most fundamental cell-level task. We construct datasets  
236 across three dimensions: single-cell type annotation, batch annotation, and spatial transcriptomic  
237 cell annotation. These correspond to phenotype recognition at the single-cell, batch, and spatial  
238 levels, respectively. This task systematically evaluates whether the model can correctly identify cell  
239 identity—a prerequisite for all downstream cell biology analyses.

240 **Drug Response Prediction.** The drug response prediction task requires the model to determine  
241 whether a cell is sensitive or resistant to a given drug. This task not only reflects whether the model  
242 can capture the relationship between drug mechanisms and cellular states but also directly maps  
243 to critical applications in precision medicine. It tests whether LLMs can transfer molecular-level  
244 knowledge into clinically relevant functional predictions.

245 **Cell Perturbation.** The cell perturbation task focuses on the ability of the model to reason about  
246 cellular state changes under external perturbations, such as gene knockouts. Specifically, the model  
247 must predict the impact of a perturbation on the transcriptome of a cell. The significance of this  
248 task lies in measuring whether LLMs can engage in causal reasoning about the relationships among  
249 genes, cells, and functions, rather than remaining at the level of static description.

250 **Conditional Cell Generation.** The conditional cell generation task requires the model to generate  
251 biologically plausible expression profiles under specified conditions, such as cell type or microen-  
252 vironment. This generative and creative task evaluates whether LLMs possess the potential for  
253 design-oriented applications, such as drug development or synthetic biology. It reflects the bound-  
254 ary between “understanding” and “generation” in LLM-based cell biology research.

## 255 256 257 5 INFLUENCING FACTOR

258 In this section, we systematically investigate multiple factors that influence the performance of large  
259 language models (LLMs) on cell-related tasks, to assess their robustness in downstream applications.  
260 Specifically, we focus on three aspects: gene length of input, robustness to input, and prompting  
261 learning.

262 **Gene Length Selection.** When constructing inputs, genes are sorted by expression values in de-  
263 scending order, and the top  $k$  gene names are sequentially provided to the model. This truncation  
264 strategy partially mimics the commonly used “feature selection” process, since highly expressed  
265 genes are often more representative and carry stronger biological signals. However, discarding low-  
266 expression genes may result in the loss of critical information. By varying the gene input length, we  
267 evaluate the sensitivity and dependency of LLMs on the amount of input information.

**Input Robustness.** Typically, LLMs take a gene list sorted from high to low expression as input. In this study, we perturb the order of genes to examine robustness: on one hand, this simulates potential noise in sequencing data; on the other, it tests whether the model has truly captured the ordering relationships among genes.

**Prompting Learning.** Prompting is a key component in the application of LLMs. We focus on the impact of few-shot prompting and prompt template design on model performance. Few-shot prompting reflects the adaptability and generalization of LLMs in low-resource scenarios, while different prompt templates help evaluate their stability under complex contexts. These factors are critical for advancing the practical deployment of LLMs in biomedical applications.

## 6 METRIC

In this section, we introduce the metrics used for CeLLM evaluation, which fall into three categories: classification metrics, ranking metrics, and foundation model-based evaluation. These three categories correspond to different task types and together provide a comprehensive view of model performance on cell-related tasks.

**Classification.** For tasks that are inherently classification problems, such as cell type annotation and drug response prediction, we adopt commonly used metrics including accuracy, recall, precision, and F1-score. In addition, we employ the Top@k overlap and ROUGE metric to evaluate the degree of overlap between predicted results and ground truth, which is particularly relevant in candidate set prediction or multi-label scenarios. These metrics directly reflect the reliability and correctness of the model in standard classification tasks.

**Rank.** For tasks that require predicting ranked outputs, such as cell perturbation prediction or cell condition generation, we employ ranking-based consistency metrics. Specifically, Spearman’s rank correlation coefficient measures the overall correlation between predicted and true expression ranks, while NDCG@k emphasizes the prediction quality of the top-expressed genes. These ranking metrics assess not only whether the model captures the statistical patterns of real expression profiles, but also whether it produces biologically reasonable preferences in gene selection.

**FM-as-judge.** Relying solely on statistical measures is insufficient to capture the full quality of model outputs. Therefore, we further propose to use domain-specific foundation models (FM) as judges. [In particular, we leverage a single-cell foundation model geneformer to compute the distance between predicted and true cells in biologically meaningful embedding spaces. Traditional metrics measure whether “gene orderings are similar,” whereas the geneformer cosine measures whether “the generated cells are in the correct biological state.” Together, they form a complementary evaluation framework that spans low-dimensional and high-dimensional spaces, combining explicit structure with implicit semantics.](#) Moreover, for descriptive text outputs (gene annotation), we evaluate them with biomedical language models such as BioBERT (Lee, 2020) to verify their semantic plausibility and domain alignment. This class of metrics enables a holistic evaluation of LLMs, covering both biological relevance and semantic coherence.

## 7 RESULTS AND ANALYSIS

### 7.1 EXPERIMENT SETTINGS

In this section, we conduct a comprehensive experimental evaluation of large language models on cell-related tasks. The benchmark covers three levels and six categories of tasks: gene annotation, multi-omics integration, cell annotation, drug response, cell perturbation, and cell condition generation. We evaluate 15 models: GPT-5, Gemini-2.5-pro, Gemini-2.5-flash, cell-o1, rBio1, C2S, Deepseek-V3.1, gpt-oss-120B, gpt-oss-20B, Llama-3.1-8B, Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-32B, Qwen3-30B-A3B, including open-source, closed-source, and domain-specific models, spanning a variety of scales, reasoning and non-reasoning types, as well as dense and MoE architectures. The domain-specific models are deployed locally on H800 GPUs using vLLM, while the other models are accessed via their official APIs for inference. The prompt templates and dataset detail information for each task are provided in the Appendix. Due to space limitations, only the key results are presented in the main text, while the full results can be found in the Appendix.

324 7.2 MAIN RESULTS

325 7.2.1 GENE ANNOTATION

326 **Setup.** In the gene annotation task, we require the model to recall the biological function of a gene,  
 327 using GO and NCBI as reference answers, and we evaluate performance by comparing the semantic  
 328 similarity between the model outputs and the reference biological texts using **BioBert** and **ROUGE**.  
 329  
 330

331 **Results.** From the Table 18, we observe that in the simple gene annotation task, all models perform  
 332 similarly. Neither model size nor the use of reasoning appears to be a determining factor. All models  
 333 demonstrate a basic level of gene comprehension.  
 334

335 7.2.2 MULTI-OMICS INTEGRATION

336 **Setup.** In the multi-omics task, we used CITE-seq and RNA-seq data as two different omics sources  
 337 and set up three experimental conditions: CITE-seq, RNA-seq, and Integrated (using both). This  
 338 design aims to investigate the performance differences of LLMs in handling single-omics versus in-  
 339 tegrated multi-omics data, thereby evaluating their cross-omics generalization ability and assessing  
 340 whether they can leverage complementary information to improve the accuracy and robustness of  
 341 downstream tasks.  
 342

343 **Results.** As summarized in Table 11, additional omics information does not lead to consistent  
 344 performance improvements. In most models, the results of the Integrated and RNA settings are  
 345 similar, while in some cases the Integrated setting performs much worse than RNA. The trend shows  
 346 that larger models are more capable of leveraging multi-omics information, with GPT-5 and Gemini-  
 347 2.5 Pro achieving better performance in the Integrated setting compared to RNA. The think mode  
 348 does not provide stable improvements.

349 7.2.3 CELL ANNOTATION

350 **Setup.** We evaluated three levels of cell annotation: single-cell annotation, batch-level annotation,  
 351 and spatial annotation. For batch-level annotation, we adopted the same metrics as cell-o1. To ex-  
 352 amine whether large language models can leverage spatial transcriptomics information, we designed  
 353 comparative experiments where only single-cell information from spatial transcriptomics was pro-  
 354 vided, along with the five or ten nearest neighbors, as well as the five next-nearest neighbors, as  
 355 prompt information.  
 356

357 **Results.** The results summarized in Table 1, Table 2, and Table 9 demonstrate that open-source  
 358 models and domain-specific models exhibit capabilities comparable to closed-source models, and  
 359 in some datasets even surpass them. Models trained at the batch level, such as cell-o1, show trans-  
 360 ferable performance on single-cell annotation tasks, highlighting the potential of “all-in-one” large  
 361 language models for cross-level generalization. Moreover, both model size and explicit reasoning  
 362 modes consistently contribute to performance improvements in cell annotation. In contrast, the C2S  
 363 model, with its limited capacity, lacks the ability to effectively follow instructions. Regarding spatial  
 364 information, most models tend to treat it as noise and fail to utilize it efficiently, although a subset  
 365 of models can benefit from spatial transcriptomics data to achieve performance gains.  
 366

367 Table 1: Results of the single-cell annotation task.

368

Setting	Brain				Myeloid				Organoid			
Metric	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
<i>Closed-source LLMs</i>												
GPT-5	81.06	65.17	67.10	67.50	12.37	18.43	12.01	20.00	<b>65.26</b>	<b>65.20</b>	<b>63.58</b>	<b>71.50</b>
<i>Domain-specific LLMs</i>												
cell-o1	54.31	47.90	44.00	47.00	13.83	22.20	14.67	21.00	22.31	20.40	17.29	23.50
C2S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>												
gpt-oss-120B	<b>81.27</b>	<b>75.31</b>	<b>76.08</b>	<b>75.50</b>	<b>19.91</b>	<b>22.25</b>	<b>18.09</b>	<b>26.00</b>	62.14	60.53	58.02	65.00
Qwen3-8B	37.11	28.15	23.45	29.50	6.05	8.69	5.12	9.50	21.74	22.57	19.74	27.00
Qwen3-8B-think	52.65	34.75	30.45	37.00	10.39	13.80	8.44	12.00	37.21	30.61	24.24	33.00
Qwen3-14B	68.61	48.08	46.88	49.00	11.37	16.46	9.68	15.50	28.76	28.85	23.34	35.00

371  
372  
373  
374  
375  
376  
377

Table 2: Results of the spatial cell annotation task.

Setting Metric	non				top5				top10				sample5			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
<i>Closed-source LLMs</i>																
GPT-5	24.00	27.37	21.29	21.02	27.00	33.72	<b>29.00</b>	<b>25.99</b>	20.00	30.68	23.27	21.18	19.00	25.79	22.82	19.19
Gemini-2.5-pro	<b>32.00</b>	26.76	<b>30.00</b>	25.54	23.00	25.96	23.20	21.17	23.00	22.78	<b>27.19</b>	21.58	29.00	27.50	<b>28.75</b>	<b>25.47</b>
<i>Domain-specific LLMs</i>																
rBio1	16.00	8.21	7.19	5.47	19.00	7.21	6.98	4.31	15.00	1.55	4.93	2.35	18.00	3.91	6.63	4.05
<i>Open-source LLMs</i>																
gpt-oss-120B	<b>32.00</b>	<b>37.17</b>	24.76	<b>25.58</b>	<b>30.00</b>	<b>36.15</b>	19.48	21.84	<b>35.00</b>	<b>42.62</b>	25.25	<b>27.63</b>	<b>33.00</b>	<b>32.42</b>	22.71	22.74
Qwen3-8B	23.00	16.27	13.79	12.10	16.00	6.94	8.99	6.46	19.00	11.74	9.53	7.98	18.00	12.66	10.21	8.56
Qwen3-8B-think	23.00	10.52	14.20	11.07	22.00	11.17	12.26	9.83	24.00	13.25	13.23	11.56	24.00	15.78	12.99	11.89

## 7.2.4 CELL PERTURBATION

**Setup.** We selected the high-quality gene perturbation VCC dataset as our benchmark and designed four tasks under perturbation conditions. Specifically, (i) given background cells and expression profiles, we determine whether a specific gene is differentially expressed under a perturbation condition; (ii) if a gene is differentially expressed, we classify whether it is increase or decrease; (iii) given the control group and the perturbed highly variable gene profile, we determine whether the perturbation is caused by a specific gene. These three tasks are formulated as binary question-answering problems. (iv) Finally, given the differentially expressed gene profile of the control group, we generate the perturbed gene profile, which is formulated as a gene re-ranking task.

**Results.** From the Table 12 and Table 14, we observe that when determining differentially expressed genes or identifying perturbation genes, models tend to conservatively answer No in the absence of absolute evidence. This bias persists even when prompts explicitly indicate that 50% of the samples are positive, suggesting that the models remain overly cautious. Nevertheless, as shown in Table 13, binary classification for direct differential expression (DIR) achieves performance above random. In the perturbation cell generation task shown in Table 15, models exhibit a clear and stable scale effect, highlighting that generative tasks demand strong generalization ability and substantial knowledge capacity, while smaller models are prone to severe hallucinations. Moreover, we find that the think mode does not always lead to improved performance and can sometimes even degrade it. We hypothesize that this discrepancy arises because annotation tasks, which rely on sequential and topological reasoning over gene lists, align well with conventional chain-of-thought training, whereas differential expression tasks involve complex information analysis and querying—integrating perturbation conditions, background expression profiles, and target genes—which do not follow a purely sequential topological structure, making standard think less effective.

## 7.2.5 DRUG RESPONSE

**Setup.** We evaluated the ability of LLMs on drug response prediction, formulated as a binary classification problem to determine whether a given cell is sensitive or resistant to a drug.

**Results.** As shown in Table 16, similar to the prediction of differentially expressed genes, the models tend to predict Sensitive, and this tendency is even more pronounced under the think mode.

## 7.2.6 CELL CONDITION GENERATION

**Setup.** We evaluated the ability of LLMs to generate biologically plausible gene expression profiles conditioned on specified cell types. Models were instructed to output the top 200 most highly expressed genes in descending order, given inputs of organ origin and cell type.

**Results.** As shown in Table 17, model performance varies significantly across categories. Top-tier closed-source models (GPT-5, Gemini-2.5-pro) consistently outperform others, particularly in Cosine Similarity (Geneformer embedding) and nDCG@100, indicating superior capture of global expression patterns and ranking structure. Performance generally improves with model scale, though thinking models did not demonstrate superior performance on this task.

### 7.3 INFLUENCING FACTOR

In this section, we investigate several key factors influencing LLM performance on cellular tasks. Due to space constraints, the results are provided in Table 19, Table 20 and Table 21 of the Appendix.

**Input length selection.** We conducted an ablation study to investigate whether the length of the input gene list affects model performance. While longer lists potentially provide richer information, models may not be able to fully exploit this advantage. We evaluated gene list lengths of 20, 50, 100, 200, and 500. As shown in the Table 21, the range of 100 to 200 emerges as the optimal window for nearly all models: this length is sufficient to retain critical marker genes while avoiding excessive context length that may dilute the model’s ability to focus on informative signals.

**Input robustness.** We performed random perturbations at levels of 0, 20, 40, 60, 80, and 100 with 0 corresponding to the standard brain single-cell annotation group. Here,  $p\%$  denotes randomly selecting  $p\%$  of the sequence and then shuffling it. We did not adopt reversal-based metrics, as perturbation scenarios in biology are more often characterized by local rank disruptions. As shown in Table 20, we found that model performance was largely consistent across all groups. Interestingly, most other models exhibited improved performance after perturbation. We hypothesize that this is due to the presence of highly expressed but non-specific genes within the same tissue in the brain dataset. These genes, appearing at the top of the gene list, may mislead models or obscure true marker genes. Our analysis confirmed this: without perturbation, models tended to focus on these non-specific genes, whereas perturbation shifted true marker genes upward in rank, allowing models to attend to them more effectively. While this phenomenon may be dataset-specific, it raises an important question: how should cell sentences be constructed to better capture cell specificity? In particular, for tissue or subtype classification, excluding highly expressed but non-marker genes may yield more representative inputs—an aspect largely overlooked by prior LLM-based cell models.

**Few-shot prompt learning.** In the single-cell annotation task on the brain dataset, we conducted a few-shot ablation experiment by placing  $k = 3$  samples from the training set along with their answers into the system prompt. We designed two settings: leak (the few-shot prompt contains the cell type of the current query) and mask (the few-shot prompt does not contain the cell type of the current query). The results shown in Table 19 show that leak consistently improves performance, while mask often leads to worse results than not using few-shot at all. This may be because the model tends to rely on the provided prompt information and often treats the gene list as a set, inferring cell types by checking intersections with few-shot samples rather than performing genuine biological reasoning. When the few-shot samples include some commonly highly expressed genes, the model is prone to misclassification, and this issue cannot be effectively mitigated by prompt design.

**Prompt framing.** In the perturbed differential expression gene prediction task, we adopted two prompting frameworks: (i) asking the model whether a given gene is differentially expressed; and (ii) providing a choice space of  $k = 5$  genes and asking the model to select the one most likely to be differentially expressed. We also assigned a rank to each gene based on its true differential expression level  $p$ , and computed the average rank of the model’s predictions. Both tasks shared the same set of positive samples and were of equal difficulty. As shown in Table 12, in the QA setting, model responses tended to be overly conservative. Upon examining the reasoning traces, we observed that most models preferred to answer No (i.e., not differentially expressed) in the absence of decisive evidence. As a result, the recall for negative samples approached 1 across nearly all models, while the accuracy was close to random. In contrast, under the CLS setting, models demonstrated genuine reasoning ability: their accuracy substantially exceeded the random baseline of 20%, and clear scaling effects emerged, with larger models consistently performing better. The rank metric, which reflects how close the predicted gene is to the ground truth, exhibited a strong negative correlation with accuracy; predictions closer to the correct answer yielded higher accuracy.

## 8 CONCLUSION

We present **CeLLM**, a unified benchmark for evaluating large language models in cellular biology. By covering diverse tasks, models, and metrics, CeLLM provides a systematic framework to assess strengths and limitations of current approaches. Serving as a sustainable and evolving platform, CeLLM aims to accelerate the development of LLMs toward the vision of virtual cells and enable more generalizable, interpretable, and biologically grounded modeling.

486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539

## ETHICS STATEMENT

We adhere to the ICLR Code of Ethics in all aspects of this research. All experimental data and results presented are authentic, with full transparency in methodology and analysis. We ensure that no form of plagiarism or falsification has occurred, and all sources are properly cited. In our work, we have taken steps to ensure fairness, equity, and respect for all contributors, avoiding any form of bias or discrimination. All contributors to this research are duly acknowledged for their roles, and any conflicts of interest have been disclosed. We have also respected the confidentiality and privacy of any sensitive data, ensuring compliance with relevant ethical guidelines. The peer review process has been conducted with the utmost respect for fairness and objectivity. Finally, we are committed to promoting an inclusive and open academic community and fostering a culture of collaboration and mutual respect.

## REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. The code and data used in this study will be made publicly available upon publication, and we provide clear instructions for their usage to enable replication of our experiments. All hyperparameters, datasets, and experimental setups are described in detail in the paper to allow others to reproduce our findings.

## REFERENCES

- Cellxgene portal. <https://cellxgene.cziscience.com/>, 2025. Accessed: 2025-09-25.
- 10x Genomics. 10x genomics pbmc 68k dataset. <https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc68k>, 2017. Accessed: 2025-09-25.
- A. et al. Abdin. Deepseek-v3 technical report. 2024. URL <https://arxiv.org/pdf/2412.19437>.
- Nadim et al. Aizarani. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature*, 572:199–204, 2019. doi: 10.1038/s41586-019-1373-2.
- S. et al. Biderman. Cell2sentence: Teaching large language models the language of biology. 2025. URL <https://openreview.net/pdf?id=EWt5wsEdvc>.
- Charlotte et al. Bunne. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015. URL <https://doi.org/10.1016/j.cell.2024.11.015>.
- Andrew et al. Butler. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018. doi: 10.1038/nbt.4096. URL <https://www.nature.com/articles/nbt.4096>.
- Seth Carbon et al. Amigo: online access to ontology and annotation. *Nucleic Acids Research*, 37 (suppl.2):W514–W520, 2009. doi: 10.1093/nar/gkp088.
- Sidi et al. Chen. Genome-wide crispr screen in a mouse model of tumor growth and metastasis. *Cell*, 160(6):1246–1260, 2015. doi: 10.1016/j.cell.2015.02.038. URL <https://www.sciencedirect.com/science/article/pii/S0092867415002044>.
- Yifan et al. Chen. Cellforge: Multi-agent virtual cell modeling for in-silico perturbation prediction. In *NeurIPS Workshop on Computational Biology*, 2024.
- Emily Clough et al. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Research*, 52(D1):D138–D144, 2024. doi: 10.1093/nar/gkad965.
- Hao et al. Cui. Curie: Evaluating llms on multitask scientific long context understanding and reasoning. In *ICLR*, 2025.

- 540 Haotian et al. Cui. scgpt: toward building a foundation model for single-cell multi-omics using  
541 generative ai. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.  
542 URL <https://doi.org/10.1038/s41592-024-02201-0>.
- 543
- 544 Rahul M Dhodapkar. Representing cells as sentences enables natural-language processing for single-  
545 cell transcriptomics. *bioRxiv*, pp. 2022–09, 2022.
- 546 S. et al. Dubey. Llama-3.1 foundationai security llm 8b instruct technical report. 2025. URL  
547 <https://arxiv.org/pdf/2508.01059>.
- 548
- 549 Yin et al. Fang. Cell-o1: Training llms to solve single-cell reasoning puzzles with reinforcement  
550 learning. 2025a. URL <https://arxiv.org/abs/2506.02911>.
- 551 Yin et al. Fang. Cellverse: A benchmark suite for language-based cell and drug tasks. 2025b.
- 552
- 553 Adam Frankish, Fiona Cunningham, et al. GENCODE 2025—advances in building a complete  
554 reference for human and mouse genomes. *Nucleic Acids Research*, 53(D1):DXXX–DXXX, 2025.  
555 URL <https://www.gencodegenes.org/>.
- 556
- 557 Minsheng et al. Hao. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*,  
558 21(8):1481–1491, 2024. doi: 10.1038/s41592-024-02305-7. URL <https://doi.org/10.1038/s41592-024-02305-7>.
- 559
- 560 Zhipeng et al. He. An atlas of human brain organoids. *Cell*, 187(22):5701–5721.e12, 2024. doi:  
561 10.1016/j.cell.2024.09.036.
- 562
- 563 Kasia Z. et al. Kedzierska. Zero-shot evaluation reveals limitations of single-cell foundation models.  
564 *Genome Biology*, 26(1):73, 2025. doi: 10.1186/s13059-025-03574-x. URL <https://doi.org/10.1186/s13059-025-03574-x>.
- 565
- 566 Seonok et al. Kim. Medbiolm: Optimizing medical and biological qa with fine-tuned large language  
567 models and retrieval-augmented generation. *arXiv preprint arXiv:2502.03004*, 2025.
- 568
- 569 Allon M. et al. Klein. Droplet barcoding for single-cell transcriptomics applied to embryonic stem  
570 cells. *Cell*, 161(5):1187–1201, 2015. doi: 10.1016/j.cell.2015.04.044. URL <https://www.sciencedirect.com/science/article/pii/S0092867415005000>.
- 571
- 572 Jinhyuk et al. Lee. Biobert: a pre-trained biomedical language representation model for biomedical  
573 text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682. URL  
574 <https://doi.org/10.1093/bioinformatics/btz682>.
- 575
- 576 Yuxuan et al. Liu. Evollama: Enhancing llms’ understanding of proteins via multimodal structure  
577 and sequence representations. *arXiv preprint arXiv:2412.11618*, 2024. URL <https://arxiv.org/abs/2412.11618>.
- 578
- 579 Malte D. et al. Luecken. A sandbox for prediction and integration of DNA, RNA, and proteins  
580 in single cells. In *NeurIPS Datasets and Benchmarks*, 2021. URL <https://openreview.net/forum?id=gN35BGa1Rt>.
- 581
- 582 Malte D. et al. Luecken. Benchmarking atlas-level data integration in single-cell genomics. *Nature*  
583 *Methods*, 19(1):41–50, 2022. doi: <https://doi.org/10.1038/s41592-021-01336-8>. URL <https://www.nature.com/articles/s41592-021-01336-8>.
- 584
- 585
- 586 Malte D. et al. Luecken. Defining and benchmarking open problems in single-cell analysis. *Nature*  
587 *Biotechnology*, 2025. doi: 10.1038/s41587-025-02694-w.
- 588
- 589 David et al. Lähnemann. Eleven grand challenges in single-cell data science. *Genome Biology*,  
590 21(1):31, 2020. doi: 10.1186/s13059-020-1926-6. URL <https://doi.org/10.1186/s13059-020-1926-6>.
- 591
- 592 Hengrui et al. Ma. Protex: Structure-in-context reasoning and editing of proteins with large language  
593 models. *arXiv preprint arXiv:2503.08179*, 2025. URL <https://arxiv.org/abs/2503.08179>.

- 594 OpenAI. Gpt-5 technical report. <https://openai.com/index/introducing-gpt-5/>,  
595 2025. Accessed: 2025-09-25.
- 596
- 597 OpenAI. gpt-oss-120b & gpt-oss-20b model card. 2025. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2508.10925)  
598 2508.10925.
- 599 S. et al. Peidli. scperturb: harmonized single-cell perturbation data. *G3:*  
600 *Genes—Genomes—Genetics*, 2024. doi: 10.1093/g3journal/jkae168.
- 601
- 602 Vignesh et al. Prabhakar. Omniscience: A domain-specialized llm for scientific reasoning and dis-  
603 covery, 2025. URL <https://arxiv.org/abs/2503.17604>.
- 604 Liujia et al. Qian. Grow ai virtual cells: three data pillars and closed-loop learning. *Cell Re-*  
605 *search*, 35(5), 2025. doi: 10.1038/s41422-025-01101-y. URL [https://doi.org/10.](https://doi.org/10.1038/s41422-025-01101-y)  
606 1038/s41422-025-01101-y.
- 607
- 608 Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Sizhuang He, David  
609 Zhang, Cerise Tang, Zhuoyang Lyu, Rayyan Darji, Chang Li, Emily Sun, David Jeong, Lawrence  
610 Zhao, Jennifer Kwan, David Braun, Brian Hafner, Jeffrey Ishizuka, Rahul M. Dhodapkar, Hattie  
611 Chung, Shekoofeh Azizi, Bryan Perozzi, and David van Dijk. Scaling large language models for  
612 next-generation single-cell analysis. 2025. URL [https://www.biorxiv.org/content/](https://www.biorxiv.org/content/10.1101/2025.04.14.648850v1)  
613 10.1101/2025.04.14.648850v1.
- 614 Yusuf H Roohani, Tony J Hua, Po-Yuan Tung, Lexi R Bounds, Feiqiao B Yu, Alexander Dobin,  
615 Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S Plosky, et al. Virtual cell challenge:  
616 Toward a turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025.
- 617 Wouter et al. Saelens. A comparison of single-cell trajectory inference methods. *Nature Biotechnol-*  
618 *ogy*, 37(5):547–554, 2019. doi: 10.1038/s41587-019-0071-9. URL [https://doi.org/10.](https://doi.org/10.1038/s41587-019-0071-9)  
619 1038/s41587-019-0071-9.
- 620
- 621 Eric W. Sayers et al. Database resources of the national center for biotechnology information.  
622 *Nucleic Acids Research*, 53(D1):D20–D26, 2024. doi: 10.1093/nar/gkad1067.
- 623
- 624 Noam et al. Shazeer. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long  
625 context, and next generation agentic capabilities. 2025. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2507.06261)  
626 2507.06261.
- 627
- 628 Line et al. Sikkema. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, 29:  
2412–2425, 2023. doi: 10.1038/s41591-023-02327-2.
- 629
- 630 Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):  
257–272, 2019. doi: 10.1038/s41576-019-0093-7. URL [https://doi.org/10.1038/](https://doi.org/10.1038/s41576-019-0093-7)  
631 s41576-019-0093-7.
- 632
- 633 Tim et al. Stuart. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.  
634 doi: 10.1016/j.cell.2019.05.031. URL [https://doi.org/10.1016/j.cell.2019.05.](https://doi.org/10.1016/j.cell.2019.05.031)  
635 031.
- 636
- 637 Aleksandra et al. Szałata. A benchmark for prediction of transcriptomic responses to perturbations  
(OP3). In *NeurIPS Datasets and Benchmarks Track*, 2024.
- 638
- 639 Q. et al. Team. Qwen3 technical report. 2025. URL [https://arxiv.org/pdf/2505.](https://arxiv.org/pdf/2505.09388)  
640 09388.
- 641
- 642 The Gene Ontology Consortium. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):  
iyad031, 2023. doi: 10.1093/genetics/iyad031.
- 643
- 644 Cole et al. Trapnell. Differential analysis of gene regulation at transcript resolution with rna-seq.  
645 *Nature Biotechnology*, 31(1):46–53, 2013. doi: 10.1038/nbt.2450. URL [https://doi.org/](https://doi.org/10.1038/nbt.2450)  
646 10.1038/nbt.2450.
- 647
- Susan Tweedie, Beatrice Braschi, Kieron Gray, et al. Genenames.org: the HGNC and VGNC re-  
sources in 2021. *Nucleic Acids Research*, 49(D1):D939–D946, 2021. doi: 10.1093/nar/gkaa1073.

- 648 Shinya et al. Velasco. Individual brain organoids reproducibly form cell diversity of the human  
649 cerebral cortex. *Nature*, 570:523–527, 2019. doi: 10.1038/s41586-019-1289-x.  
650
- 651 Jie et al. Wang. Perturbqa: A benchmark for perturbation experiments in single-cell analysis. *arXiv*  
652 *preprint arXiv:2503.12345*, 2025.
- 653 Zeyu et al. Xiao. Protein large language models: A comprehensive survey. *arXiv preprint*  
654 *arXiv:2502.17504*, 2025a. URL <https://arxiv.org/abs/2502.17504>.  
655
- 656 Zeyu et al. Xiao. Stella: Towards protein function prediction with multimodal llms integrating  
657 sequence-structure representations. *arXiv preprint arXiv:2506.03800*, 2025b. URL <https://arxiv.org/abs/2506.03800>.  
658
- 659 Jiaqing et al. Xie. Qcbench: Evaluating large language models on domain-specific quantitative  
660 chemistry. *arXiv preprint arXiv:2508.01670*, 2025.  
661
- 662 Yinggan et al. Xu. Physense: Principle-based physics reasoning benchmarking for large language  
663 models. *arXiv preprint arXiv:2505.24823*, 2025.
- 664 Rui Yang and Paula Gomez. cell-o1: Batch-level reasoning for single-cell rna-seq cell type annota-  
665 tion. *bioRxiv*, 2024.  
666
- 667 Tao et al. Yang. Build the virtual cell with artificial intelligence: a perspective for cancer research.  
668 *Military Medical Research*, 12(1):4, 2025. doi: 10.1186/s40779-025-00591-6. URL <https://doi.org/10.1186/s40779-025-00591-6>.  
669
- 670 Zizhen et al. Yao. A high-resolution transcriptomic and spatial cell-type atlas of the whole adult  
671 mouse brain. *Nature*, 624:121–134, 2023. doi: 10.1038/s41586-023-06812-z.  
672
- 673 Rong et al. Ye. rbio1: Reinforcement-based biological reasoning with world-model verifiers. *arXiv*  
674 *preprint arXiv:2410.04567*, 2024.
- 675 Yuansong et al. Zeng. Cellfm: a large-scale foundation model pre-trained on transcriptomics  
676 of 100 million human cells. *Nature Communications*, 16(1):4679, 2025. doi: 10.1038/  
677 s41467-025-59926-5. URL <https://doi.org/10.1038/s41467-025-59926-5>.  
678
- 679 Meng et al. Zhang. Molecularly defined and spatially resolved cell atlas of the whole mouse brain.  
680 *Nature*, 624:111–120, 2023. doi: 10.1038/s41586-023-06808-9.
- 681 Suyuan Zhao, Yizhen Luo, Ganbo Yang, Yan Zhong, Hao Zhou, Zaiqing Nie, et al. Stofm: A multi-  
682 scale foundation model for spatial transcriptomics. 2025. URL <https://arxiv.org/abs/2507.11588>.  
683
- 684 Grace X. Y. et al. Zheng. Massively parallel digital transcriptional profiling of single cells. *Nature*  
685 *Communications*, 8:14049, 2017. doi: 10.1038/ncomms14049.  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701

## A APPENDIX

### USE OF LARGE LANGUAGE MODELS

In accordance with ICLR’s disclosure requirements regarding the use of large language models (LLMs), this section specifies the role and scope of LLM usage in this work:

- As evaluation subjects (core of the study):** This work evaluates LLMs as the primary objects of study. We design a benchmark framework for AIVC (Artificial Intelligent Virtual Cells) to assess model performance across a wide range of tasks, including multi-omics integration analysis, gene annotation, cell type annotation, conditional cell generation, and perturbation signature prediction.
- As a writing refinement tool (limited assistance):** During manuscript preparation, LLMs were used to provide limited assistance with grammar correction, word choice, sentence restructuring, and improving overall fluency. Importantly, LLMs were not used to draft or rewrite any core scientific content, including conceptual arguments, experimental design, data analysis, results interpretation, or figure preparation. The specific prompts and representative examples of edits have been retained by the authors and can be made available for review upon request.
- Accountability and verification:** All text generated or modified with LLM assistance was carefully reviewed by the authors to ensure the absence of plagiarism, fabricated information, or misleading statements. The authors take full responsibility for all content in the manuscript. LLMs are not listed as authors or contributors to this work.

## REBUTTAL

### R.1 COMPARED WITH OTHER BENCHMARKS

Table 3: Task inclusion across different biological benchmarks.

Task	Single-Cell Omics Arena	CellVerse	CeLLM
Single-Cell Annotation	✓	✓	✓
Batch-Level Cell Annotation			✓
Spatial Cell Annotation			✓
Drug Response Prediction		✓	✓
Perturbation Prediction (DE & Dir)		✓(Only QA)	✓(QA & CLS)
Perturbation Prediction (HVG)			✓(QA & CLS)
Perturbed Cell Generation			✓
Gene Annotation			✓
Multi-Omics Integration			✓
Condition Cell Generation			✓
<b>Total Datasets</b>	<b>2</b>	<b>3</b>	<b>10</b>

Table 4: Model and metric inclusion across different biological benchmarks.

Benchmark	Single-Cell Omics Arena	CellVerse	CeLLM
<b>Model</b>	<b>8 models:</b> Mixtral(8×7B, 8×22B); Deepseek(67B); Qwen2(72B); Llama3(70B); GPT-4o, 4o-mini; C2S.	<b>12 models:</b> Qwen2.5(7B, 32B, 72B); Deepseek-V3,R1; GPT-4,4o,4o-mini,4.1,4.1-mini; C2S.	<b>20 models:</b> Qwen3(4B, 7B, 14B, 32B, 30B-A3B with thinking mode); Deepseek-V3.1; Llama(8B); GPT-oss(20B, 120B); GPT5; Gemini 2.5(flash,pro); C2S; Cell-o1; rBio.
<b>Metric</b>	<b>1 type:</b> classification (BLEU, F1).	<b>1 type:</b> classification (Precision, Recall, F1, Accuracy).	<b>3 type:</b> rank (Spearman, NDCG@k, Rank), FM-as-judge (CosSim, BertScore), classification (Precision, Recall, F1, Accuracy, Top-k overlap, ROUGE-1).

## R.2 OPTIONAL WINDOW AND K-SHOT STUDY

### R.2.1 OPTIONAL WINDOW

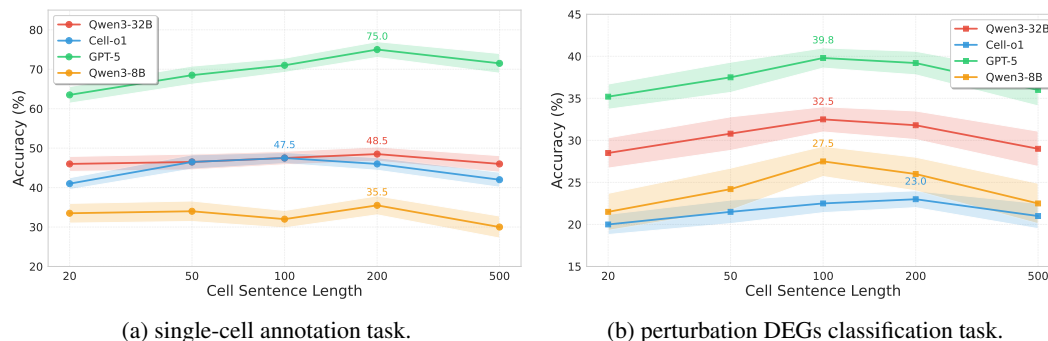


Figure 2: Effect of varying cell sentence lengths on model performance

### R.2.2 K-SHOT WITH MASK AND LEAK STUDY

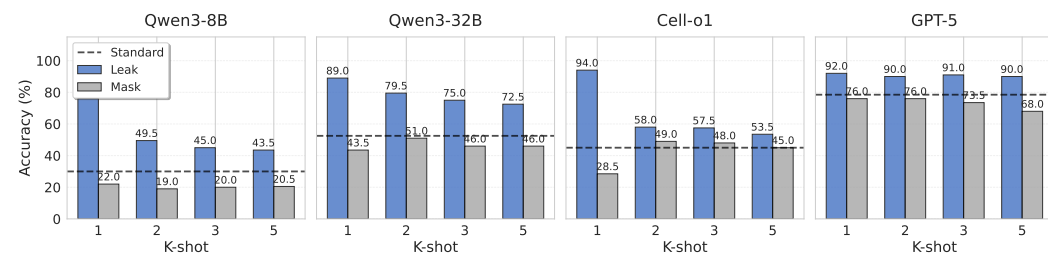


Figure 3: Performance of Mask and Leak few-shot strategies as a function of K.

## R.3 INSIGHTS AND ACTIONABLE TAKEAWAYS

Our benchmark yields the following observations and actionable recommendations:

- **Model choice should be task-dependent rather than uniformly “bigger is better”.**
  - For knowledge recall tasks such as gene annotation, all models perform similarly, and frontier closed-source models provide little additional gain. Lightweight open-source models are sufficient.
  - For cell annotation tasks across single-cell, batch, and spatial settings, strong open-source models (e.g., gpt-oss-120B, Qwen3-14B) and domain-specific models (cell-o1) can match or surpass closed-source models, indicating that practitioners can rely on open models where privacy or cost is critical.
  - For generative tasks requiring full expression modeling (cell perturbation generation and conditional cell generation), model scale matters significantly: small models hallucinate heavily, whereas GPT-5 and Gemini-2.5 achieve consistently better. Frontier models are preferable when reliable generative biology is needed.
- **Reasoning (“think”) modes should be used selectively.**
  - On sequential, topology-like tasks such as cell annotation with ordered gene lists, explicit reasoning modes (e.g., Qwen3 thinking models) generally improve performance.
  - On perturbation and differential expression tasks, think-mode often fails to help and can even degrade performance, likely because these tasks require integrating noisy signals rather than symbolic step-by-step reasoning.
  - **Takeaway:** Chain-of-thought benefits annotation-style reasoning over gene rankings, but should not be assumed to help all tasks; validation is needed per task family.

- 810
- 811 • **Input construction for “cell sentences” should use medium gene lengths rather than**
- 812 **naive long contexts.**
- 813 – Our gene list length ablation shows that 100–200 genes form an empirical “sweet
- 814 spot”: shorter lists lose marker information, while very long lists dilute salient signals.
- 815 – **Takeaway:** For cell-to-text pipelines, we recommend using 100–200 genes.
- 816 • **Multi-omics should not be concatenated naïvely; larger models exploit integration**
- 817 **more effectively.**
- 818 – In multi-omics integration experiments, combining CITE-seq and RNA-seq into a
- 819 longer prompt does not consistently improve performance, and may harm some mod-
- 820 els. Only the largest models (GPT-5, Gemini-2.5-pro) show clear benefits.
- 821 – **Takeaway:** Multi-omics integration for LLMs is non-trivial. Structured representa-
- 822 tions or architectures are needed rather than simple concatenation.
- 823
- 824 • **Prompt framing and few-shot design critically shape model behavior.**
- 825 – For perturbation and drug-response prediction, binary QA prompts induce a conser-
- 826 vative bias toward “No change” or “Sensitive”, resulting in near-random accuracy
- 827 despite high negative recall. Framing the task as choosing among candidate genes
- 828 yields much higher accuracy and clearer scale effects.
- 829 – Few-shot prompting on brain cell annotation shows that “leaky” demonstrations (con-
- 830 taining the target cell type) reliably improve accuracy, whereas “masked” demonstra-
- 831 tions hurt performance. Models tend to perform gene-set pattern matching rather than
- 832 pathway-level reasoning.
- 833 – **Takeaway:** (i) Use ranking/selection-style prompts instead of yes/no QA for
- 834 perturbation-style tasks; (ii) Carefully design few-shot examples to avoid label leak-
- 835 age and shallow gene-set matching.
- 836 • **Spatial information must be encoded carefully.**
- 837 – In spatial cell annotation, appending nearest-neighbor gene lists is often interpreted as
- 838 noise; only a few models show small gains.
- 839 – **Takeaway:** Naive concatenation of neighbor cells is insufficient. Future work should
- 840 explore structured (e.g., graph-based) or aggregated spatial encodings.
- 841
- 842 • **Evaluation design: combine classic metrics with FM-as-judge for biological realism.**
- 843 – Traditional classification and ranking metrics capture general correctness but miss bi-
- 844 ologically meaningful differences. Embedding-based distances using single-cell foun-
- 845 dation models (e.g., Geneformer) or biomedical text encoders (e.g., BioBERT) better
- 846 capture biological and semantic fidelity.
- 847 – **Takeaway:** Future virtual-cell models should be evaluated using both classic metrics
- 848 and biologically informed embedding spaces; CeLLM provides such a protocol.

849

850 The above summarized takeaways enable both model developers and end users to directly translate

851 CeLLM’s experimental findings into concrete decisions regarding model selection, input construc-

852 tion, prompt design, and evaluation strategies.

853 **Limitations and future work.** Although CeLLM provides a systematic evaluation of foundation

854 models across diverse single-cell and spatial transcriptomics tasks, our benchmark primarily fo-

855 cuses on single-turn, static model behavior. Future work will build upon CeLLM to explore cell-

856 aware agents whose capabilities extend beyond one-shot predictions—enabling multi-step decision-

857 making, iterative hypothesis generation and refinement, dynamic incorporation of new evidence,

858 and autonomous biological analysis.

#### 859 R.4 DISCUSSION OF CELL REPRESENTATION IN LLMs

860

861 Our choice to use the “cell sentence” format in the benchmark is motivated by three key reasons:

862 **Alignment with the prevailing paradigm and benchmark focus.** The cell sentence representation

863 is the most widely adopted format in SOTA Cell LLMs. For example, both the Cell2Sentence and

Cell-o1 models use this approach for downstream tasks. Since CeLLM aims to evaluate progress within this emerging paradigm, adopting this mainstream representation is essential.

**Information retention and interchangeability.** Although the representation is textual, it preserves critical biological information. The original C2S paper demonstrated a strong linear relationship between gene order in the cell sentence and gene expression levels in log space ( $R^2=0.85$ ). This indicates that little information is lost during the conversion. Such interchangeability allows leveraging the strengths of LLMs in natural language processing while retaining the ability to convert the representation back into gene-expression vectors for use in traditional single-cell analysis methods.

**Robustness.** Furthermore, work (Dhodapkar, 2022) related to cell text representations highlights that the ranking mechanism serves as a form of rank normalization. This property enhances the model’s robustness to noise, enabling more stable performance when processing single-cell data.

We observed that when given the input cell sentence, LLMs can automatically infer pathway-related information from the gene list in downstream task.

**LLM Output:** </analyze> The top 100 highly expressed genes in this cell include several long non-coding RNAs (lncRNAs) such as MALAT1, MIR99AHG, ENSG00000231557, ENSG00000225421, LINC00609, PCAT1, LINC00882, FTX, CASC15, LINC02899, and CASC2. These lncRNAs are often associated with regulatory functions in various cell types, including glial cells and neurons. Additionally, there are several genes related to ciliary structure and function, such as DNAH9, CFAP299, DNAH6, DNAH7, DNAH11, DNAH12, CFAP54, CFAP61, CFAP44, CFAP52, CFAP418-AS1, DNAI4, and HYDIN. These genes are commonly expressed in cells with motile cilia, such as ependymal cells and some types of epithelial cells. The presence of RORA and SOX5 suggests involvement in developmental processes, which is consistent with certain glial or neuronal cell types. NRXN1 and NCAM2 are involved in cell-cell adhesion and are commonly expressed in neurons and glia. Among the cell types listed, the ependymal cell is a strong candidate because of the presence of ciliary-related genes and the expression of RORA, SOX5, and NRXN1, all of which are relevant to ependymal cells, which are ciliated cells lining the ventricles of the brain and the central canal of the spinal cord. Furthermore, ependymal cells are known to express a variety of lncRNAs and have complex regulatory networks involving developmental genes. The expression profile does not strongly align with other listed cell types such as oligodendrocytes, astrocytes, fibroblasts, or macrophages, which have distinct marker genes not prominently represented here. </analyze>  
<answer>ependymal cell</answer>

In addition, to examine other types of cell representations, we test four representative models using different variants of the cell sentence: the original version, the version augmented with expression information, and the version incorporating gene co-expression information. The results on the cell-type annotation task are shown in Table 5.

Table 5: Performance comparison(%) of different cell representations on downstream task.

Model	cell sentence	with expression	with gene co-expression
Qwen3-8B	<b>72.14</b>	72.04	66.17
Qwen3-32B	<b>84.08</b>	83.08	77.61
Cell-o1	<b>38.81</b>	25.37	34.83
GPT-5	<b>89.55</b>	89.55	89.55

As the table indicates, providing additional information did not yield further performance gains; on the contrary, it even negatively affected the performance of smaller models. Regarding spatial information, in the spatial cell-annotation task in our paper, we conveyed spatial context to the LLM by constructing structured multiple cell sentences based on the distances of neighboring cells to the center cell. We found that not all models were able to make effective use of this spatial information. Therefore, for spatial transcriptomics tasks, the current cell sentence format still has limitations. Designing more effective and generalizable spatial cell-encoding strategies remains an important direction for future exploration.

## R.5 DISCUSSION OF DATASET LEAKAGE

**Data provenance and novel textualization.** Except for the gene-annotation task, all CeLLM tasks are constructed directly from *primary* single-cell or spatial omics matrices (scRNA-seq, CITE-seq, spatial transcriptomics). These raw count matrices are not natural-language documents and, to the best of our knowledge, are not included in any LLM pretraining corpora. Moreover, the inputs used in CeLLM are produced through our newly designed *cell-sentence* and *multi-omics-sentence* textualization pipeline, which serializes high-dimensional omics vectors into descriptive sentences. These templates and serialization rules did not exist prior to CeLLM, making it impossible for any LLM to have encountered identical input–output pairs during pretraining.

Table 6: Summary of the CeLLM benchmark.

Task name	Dataset	Input	Output	Metric
gene annotation	Collected gene GO annotation dataset	gene name	gene description	BertScore Precision, Recall, F1; ROUGE-1 Recall.
single-cell type annotation	Brain & Myeloid & Organoid	single cell sentence	cell type label	Precision; Recall; F1; Accuracy.
batch-level cell annotation	Cell Puzzle	Multi-cell sentence	cell type label list	Accuracy; Batch-level Accuracy; Uniqueness; Format.
spatial cell annotation	Brain1	Multi-cell sentence	cell type label	Precision; Recall; F1; Accuracy.
multi-omics integration	CITE-BMMC	Multi-omics sentence	cell type label	Precision; Recall; F1; Accuracy.
drug response prediction	CellVerse	Drug & single cell sentence	sensitive / Resistant	Precision; Recall; F1; Accuracy.
perturbation prediction (DE prediction)	Virtual Cell Challenge (VCC)	single-cell sentence & perturbation gene & gene / gene candidate	Yes/No (QA); gene name (CLS)	Rank; Precision; Recall; F1; Accuracy.
perturbation prediction (Dir prediction)	Virtual Cell Challenge (VCC)	single-cell sentence & perturbation gene & DE gene	Yes/No	Precision; Recall; F1; Accuracy.
perturbation prediction with HVGs	Virtual Cell Challenge (VCC)	control group & perturbed HVG cell sentence & perturbation gene	Yes/No (QA); gene name (CLS)	Precision; Recall; F1; Accuracy.
perturbed cell generation	Virtual Cell Challenge (VCC)	single cell sentence(HVG) & perturbation gene	single cell sentence(HVG)	Generformer cosine similarity; Top-k overlap; NDCG@k; Spearman@k.
condition cell generation	Lung & Liver & Blood	condition (organ, cell type)	single cell sentence	Generformer cosine similarity; Top-k overlap; NDCG@k; Spearman@k.

**Gene-annotation task and overlap with biological text.** For the gene-annotation task, we intentionally use Gene-Ontology-derived descriptions. These may partially overlap with public biological text, but this is by design: the goal is to evaluate whether an LLM can function as a gene-centric knowledge base. Exposure to biological corpora during pretraining is therefore part of the evaluated capability rather than an unfair advantage.

**Domain-specific models.** For domain-adapted models such as rBio, Cell-o1, and C2S, which are trained on textualized omics data, we ensure that our CeLLM test datasets do *not* overlap with the datasets used for training in their respective papers, thereby eliminating dataset-level leakage.

**Empirical evidence against contamination advantages.** Across tasks such as spatial cell annotation and differential-gene prediction, larger models do *not* consistently outperform smaller models. This pattern indicates that CeLLM tasks require genuine biological reasoning rather than memorization of any potential pretraining data, providing additional evidence that contamination is unlikely to influence our conclusions.

## R.6 SUMMARY OF CELLM BENCHMARK

Table 6 presents a summary of the CeLLM benchmark tasks, including the datasets used, the input and output formats, and the evaluation metrics.

**Datasets, Metrics, and Prompts.** Descriptions of the datasets, evaluation metrics, and prompt templates used for the tasks in the table are provided in Appendix A.3.2, Appendix A.2, and Appendix A.5.

**Input and Output Format.** The input consists of two components: a system prompt and a user prompt. The system prompt provides task instructions, guidance, and output rules, while the user prompt contains the specific query information. In the system prompt, we require the model to place its final prediction inside answer tags so that answers can be consistently extracted and evaluated.

**Model Sampling Parameters.** LLM responses are generated using temperature = 0.7 and top\_p = 0.95, following common settings that balance diversity with coherence.

**Data Processing.** CellLM evaluates LLMs in a zero-shot setting to assess their ability to handle cell-related tasks without task-specific fine-tuning; therefore, most tasks do not contain a training set. For few-shot experiments, we split the Brain dataset into 80% training and 20% testing using a fixed random seed (42), and construct few-shot prompts using examples from the training portion. We process cellular data into core text data for various tasks according to the following steps, with a random seed of 42 applied for data sampling parts.

- **Gene Annotation:** *Structure GO terms* (Molecular Function, Biological Process, Cellular Component) and their official definitions into reference answer text.
- **Multi-omics Integration:** *Gene Normalization* is performed on scRNA-seq data; *Transcriptomics* calculates the top-200 expressed genes for each cell to generate a cell sentence; *Proteomics* extracts all surface proteins with non-zero ADT expression (max 134), ranked descendingly.
- **Single-Cell Annotation:** *Gene Name Standardization* standardizes gene names; *CellSentence Encoding* converts each sampled cell’s expression profile into a top-k gene cell sentence.
- **Batch Cell Annotation:** *Batch Aggregation* calculates batch-level mean expression profiles; *Top-Genes* extracts top-k genes based on mean profiles; *Context Construction* packages batch information, top-genes, and candidate cell types into a text prompt.
- **Spatial Cell Annotation:** *Encoding* converts central and neighbor cells into top-k gene CellSentences; *Prompt Construction* concatenates these lists by distance based on task settings.
- **Drug Response Prediction:** *Gene Name Standardization* standardizes gene names; *CellSentence Encoding* converts profiles to top-k gene lists; *Prompt Construction* packages the cell sentence and drug name into a QA-style prompt.
- **Cell Perturbation:** *DE Data Sampling* uses Wilcoxon rank-sum test and Benjamini-Hochberg correction to identify significant DEGs, sampling Positive, Moderate, and Negative samples based on FDR thresholds (0.05, 0.1, 0.5); *CLS Task Construction* creates candidate sets with 1 strong positive, 2 moderate, and 2 negative samples; *Directional Sampling* categorizes significant DEGs (FDR < 0.05) as Increase or Decrease based on log2 fold change; *HVG Ranking* constructs HVG lists sorted by expression values before and after perturbation.
- **Condition Cell Generation:** *Mean Profiling* calculates mean expression profiles for specific cell types in organs after log1p transformation; *Label Generation* converts these profiles into top-200 gene lists as labels.

## R.7 LLM OUTPUT ON CELL GENERATION TASK.

```
<analyze>
We need to simulate the top 200 expressed genes in **plasmacytoid dendritic cells (pDCs)** from human blood.
These cells are specialized dendritic cells notable for strong expression of interferon-regulated genes
(IRGs), endosomal TLRs (especially TLR7, TLR9), antigen presentation machinery, and certain surface
receptors.
**Key features of blood pDCs:**
- **Lineage and identity markers:** IL3RA (CD123), CLEC4C (BDCA2), LILRA4 (ILT7), TCF4, SPIB.
- **Interferon response / antiviral genes:** ISG15, MX1, IFIT1, IFIT3, OAS1, OAS2, IFI6, IFI27, RSAD2 (viperin
).
- **Antigen presentation machinery:** HLA class II genes (HLA-DRA, HLA-DRB1, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA
-DQB1), plus HLA class I genes (HLA-A, HLA-B, HLA-C, B2M).
- **Endosomal TLRs:** TLR7, TLR9.
- **Co-stimulatory molecules:** CD74 (invariant chain), IRF7 (master regulator of type I IFN), IRF8, etc.
- **Surface receptors involved in migration and immune sensing:** CXCR3, CXCR4, CCR7, S1PR1, ITM2C, GZMB
sometimes reported in activated states.
- **Housekeeping / structural genes:** ACTB, GAPDH, ribosomal protein genes (RPLs, RPSs), mitochondrial genes
(MT-CO1, MT-ND1, etc.). These tend to dominate the absolute highest expression in scRNA-seq.
Thus, the simulated list should begin with ribosomal and mitochondrial genes, B2M, actin, GAPDH, before
transitioning into pDC-specific genes, HLA genes, interferon-stimulated genes (ISGs), and transcription
factors. To reach 200 genes, we include a mix of: ribosomal proteins, pDC markers, HLA genes, ISGs, and
essential metabolic genes.
</analyze>
<answer> Top 200 gene list.</answer>
```

## R.8 CASE STUDIES: ROBUSTNESS OF CELL2SENTENCE TO PROMPTS

```
Basic Question
Question:
Who are you ?
Cell2Sentence Answer:
```

1026

) are expressed in the following order:

1027

Cell 1:  
MALAT1 KCNIP

1028

1029

**Drug Response Prediction Task**

1030

**Question:**

Below is a list of {gene\_names} ordered by descending expression level in a cell. Your task is to predict its response to the drug Erlotinib: [resistant, sensitive]. Make your choice in format 'Final Answer: resistant' or 'Final Answer: sensitive'

1031

{gene\_names}: MALSU1 MT-TD PABPC1L FTLPL1 TADA1 S100A9 FTH1P1 TXN2 RPS12P15 MT-CYB RPS16P5 RPS27AP1 RPS20 RPS14P3 RPL35AP RPL38P3 RPS15AP1 TMSB10P1 RPS5P2 RPL13AP17 RPSAPI ATP5E RPL1P13 NCLN EEF1A1P1 ANXA2 RPL8P2 RPL14P1 GAPDHP1 KTN1-AS1 SERPINB5 COX6CP1 HSPPE1 DSTNP4 RPL37A CAPN7 CSTA SERBP1P1 RPL2P1 RPL36AL UBBP1 PFN1P1 TMEM259 SAP30 RPS27A RPS3A RAD51 RPL13A COX5BP6 GNB4 HNRNPA3 TMSB4XP1 CFL1P1 NDUFS7 RPS23P1 RPL36AP10 NOP14 MT-ND2 SRIP1 HSP90AB2P HIST1H4E MRP514 VANGL2 EFS8L2 PPIH HIST1H2BE SPTBN2 RPL5P1 RECK RTN4P1 ARHGAP21 RPL31P1 HNRNPKP1 CCDC82 RPS20P1 GCDH RPS8P3 CIR1P2 CSDE1 TALDO1 PPP1CB EIF4A3 RPL30 EIF3I RPL11P5 SLC26A4-AS1 MT-RNR2 CCNJ PTMAP2 TAX1BP3 LGALS7B RANBP1 RPS21P4 PPIC RHOC XRRAL1 SRA1 RPL34P1 HINT1P1 NEPH.

1032

1033

1034

**Cell2Sentence Answer:**.  
659\_SAPA1 SPRNA9 PSMA

1035

1036

1037

1038

**Batch Cell Annotation Task**

1039

**Question:**

Context: The cell is from a female at the 73-year-old stage, originating from the lung. The patient has been diagnosed with chronic obstructive pulmonary disease. The patient is a smoker. There is no cancer present.

1040

1041

Cell 1: MALAT1, MTCO2P22, MTATP6P29, MTND4LP30, MTND3P6, MTND2P13, MTCYBP19, MTND5P17, MTND1P22, RPL28, MIDEAS, PDE4B, SAT1, RPS2, RPS19, S100A9, RPS27, SMAP2, JUNB, CROCC, ASTN1, RPL10, S100A8, RPL37A, CD74, RPS12, RPS15, RPL18A, ELF1, RPLP2, ZNF331, RPL15, RPS3A, GLIPR2, RPL13, HOOK2, SH3BGR13, CYBA, TXNRD1, RPS28, UPPI, ATP5MG, FTL, LPP, RPL35A, RPL13A, RPL14, TNFAIP3, SERF2, HSPBAP1

1042

1043

1044

Cell 2: CCL21, MALAT1, MT2A, TMSB4X, RPL10, STEAP1B, VIM, RPL21, ASAP1, EEF1A1, RPS27, MTCO2P22, RPL32, TPT1, MTND4LP30, FTL, RHOJ, TM4SF1, CCSER1, RPLP1, RPL1, RPL34, RPL13A, RPL13A, IL6ST, MARCF3, MTND2P13, RPL37A, MTATP6P29, RPL34, DLCL1, RPS12, RPS27A, MTCYBP19, RBFOX2, TMSB10, MTND1P22, B2M, FOS, TGF2, RPS3A, RPS28, RPS6, ZFP36, RPL41, RPS15A, RPL39, MIR4435-2HG, RPL18A, PTMA, RPS18, RPS8

1045

1046

1047

Cell 3: FTL, MTATP6P29, MTND2P13, TMSB4X, MTCO2P22, MTND4LP30, CCSER1, APOC1, MTCYBP19, MTND3P6, MT2A, B2M, RPS19, BCAS4, CCDC30, CCL18, RPS9, ADGR13, S100A6, MTND1P22, ACTB, RPS18, MTND5P17, LGALS1, APOE, CD52, COLEC11, RPLP2, RPL34, RPL39, MT1X, RPL3, RPL27A, HSPA4, TBCA, RBMS1, IFI27, PRDX1, JPT1, R3HDM2, COX5B, NSA2, TMSB10, RPS25, S100A10, SELENOW, HLA-DRA, RPS6, S100A9, ELOB

1048

1049

1050

Cell 4: MT2A, MT1E, TMSB4X, MT1X, CCSER1, MALAT1, S100A6, AGER, MTND4LP30, MT1G, CAV1, ACTB, S100A10, B2M, DSTN, RPS12, MTCO2P22, RBFOX2, RPL21, EMP2, IL6ST, RPL41, GSTP1, MTCYBP19, ATF7IP2, H3-3B, CD55, BCAS4, RPL10, RPS27, KRT7, MTATP6P29, RPL34, STEAP1B, TPT1, RPLP1, EEF1A1, RPL37, RPS18, RPS28, RPL27A, RPS16, RPLP2, RPL37A, POLR2L, KRT18, NEDD4L, FTL, MTND2P13, RPL19

1051

1052

1053

Cell 5: MALAT1, MTATP6P29, RPLP1, MTCO2P22, MTND4LP30, RPL10, AREG, RPS6, B2M, RPL13, RPS24, TMSB10, RPS12, RPS27, SRGN, RPL37A, IL6ST, RPL34, MTND2P13, RPL41, RBFOX2, RPL21, RPL18A, TPT1, CD74, RPL28, RPL14, RPS8, RPL39, RPS4X, RPS15, HLA-DRA, TXNRD1, MTCYBP19, SIK3, SPALL1, PABPC1, RPLP2, ACTB, RPS2, RPL13A, RPS16, HLA-DRB1, RPS3A, FTL, RPL3, RPL9, RPL37, RPL19, EREG

1054

1055

1056

Cell 6: MALAT1, MT2A, MTND3P6, MTCO2P22, B2M, MTND4LP30, TMSB10, MTATP6P29, MTND2P13, MTCYBP19, INSR, ADAMTS9, ACKR1, MT1E, IFITM3, IFITM2, ACTB, MT1X, LDB2, CCL14, S100A6, IGFBBP7, DHFR, IL6ST, APP, RPS18, RPS17, RBFOX2, PITPNCL1, ITM2B, RPL10, CLDN5, MTND1P22, HNRNPH1, FOXO1, CTNNB1, SASH1, RPL41, RPL12, RPS16, VWF, MTHFD1L, FLNB, RPS3A, RPL34, PTMA, SPARCL1, RPS19, MTND5P17, RPS15A

1057

1058

1059

Cell 7: MALAT1, MTATP6P29, NEAT1, MTND3P6, MTCO2P22, MTND1P22, MTND2P13, MTND4LP30, MTCYBP19, SFTPB, MACROD2, SFTPC, PDE4B, ANK3, TANC2, MBNL1, ZNF385B, PACSIN2, DLG2, SHROOM3, MYO1D, BTBD9, MACF1, DPYD, ANXA1, USP54, LINC-PINT, CXCL2, NEDD4L, LMO7, PTPRG, DOCK5, HSPB1, TRIO, SGFP2, SYTL3, PLPP3, RAB7A, MAGI3, SNX25, FOSB, QKI, WSB1, DIAPH2, ZBTB20, ANKRD1, ACACA, LDLR, RASEF, ZSWIM6

1060

1061

1062

Cell 8: MALAT1, MTATP6P29, ZSWIM6, MTCO2P22, MTCYBP19, MT2A, MTND1P22, MACF1, MACROD2, SFTPB, FHIP1A-DT, MTND2P13, VMP1, MTND3P6, IL1R1, RPLP1, RPL13, ALCAM, PALLD, DST, MTND4LP30, HNRNPH1, IL6ST, NEDD4L, RPL10, SGPP2, RBFOX2, SMYD3, APP, RPL41, RPS18, SIK3, ITGA2, EFNA5, NXN, F3\_ENSG00000117525, RPS27, WWTR1, STAT3, ZFAND3, ACBD3, RPS12, TXNRD1, DENND5A, MTUS1, AOEPEP, CXCL2, GCLC, SDK1, RPL35A

1063

1064

1065

Cell 9: MT2A, TMSB10, ACTB, MT1M, EMP1, IGFBBP7, MTCO2P22, B2M, DEPP1, MTND4LP30, HLA-B, TMSB4X, IFITM2, TM4SF1, RPL10, EPAS1, PTPRG, RPLP2, MTND2P13, ACTG1, RPL37A, STEAP1B, RPS6, TPT1, RPS8, PECAM1, IFITM3, MT1E, CCSER1, H3-3B, FTL, CD93, RPL13A, MTATP6P29, S1PR1, CAV1, MTND3P6, IL6ST, TIMP3, NES, NAMPT, RPS12, PDZD2, MTND5P17, SYN3, RPL8, RPL18, RPS27, RPL28, RPL41

1066

1067

1068

Match the cells above to one of the following cell types:

1069

1070

1071

CD1c-positive myeloid dendritic cell  
classical monocyte  
endothelial cell of lymphatic vessel  
epithelial cell of lung  
macrophage  
malignant cell  
pulmonary alveolar type 1 cell  
pulmonary artery endothelial cell  
vein endothelial cell

1072

1073

**Cell2Sentence Answer:**of lung endothelial cell.  
mucosal brick ciliated tissue group.  
bas

1074

1075

1076

1077

1078

1079

**R.9 CASE STUDIES: WHEN LLM REASONING ALIGNS WITH KNOWN BIOLOGY?****R.9.1 EXAMPLES OF INCORRECT ANSWERS****True Label:** central nervous system macrophage

1073

**LLM Output:** <analyze>The top 100 highly expressed genes in this single-cell RNA sequencing profile include a mix of long non-coding RNAs (lncRNAs) such as \*\*MALAT1\*\*, \*\*NEAT1\*\*, and \*\*LINC00278\*\*, which are often associated with various cellular processes, including regulation of gene expression, splicing, and chromatin remodeling. Additionally, the profile contains genes involved in a variety of cellular functions, including signaling, metabolism, and structural components. However, a closer look at the expression profile reveals some key markers that can be used to infer the cell type: 1. \*\*QKI\*\* (Quaking) is a well-known marker for \*\*oligodendrocyte precursor cells (OPCs)\*\* and \*\*oligodendrocytes\*\*. It is involved in RNA splicing and is highly expressed in oligodendrocyte lineage cells. 2. \*\*MBNL1\*\* (Muscleblind-like 1) is another marker that has been associated with oligodendrocyte precursor cells and is involved in RNA splicing and regulation of alternative splicing events. 3. \*\*PLXND1\*\* (Plexin D1) has been implicated in oligodendrocyte development and myelination. 4. \*\*RUNX1\*\* is a transcription factor involved in the regulation of oligodendrocyte differentiation and myelination. 5. \*\*PICALM\*\* (Phosphatidylinositol binding clathrin assembly protein) is involved in endocytosis and has been associated with oligodendrocyte lineage cells. 6. \*\*COMMD10\*\* is a gene that has been linked to oligodendrocyte development and myelination. These markers, particularly \*\*QKI\*\* and \*\*MBNL1\*\*, are

1074

1075

1076

1077

1078

1079

1080 highly specific to oligodendrocyte lineage cells, and the presence of **RUNX1** and **PICALM** further  
 1081 supports the possibility of an oligodendrocyte precursor cell or oligodendrocyte identity. While **oligodendrocyte**  
 1082 and **oligodendrocyte precursor cell** are both plausible, the expression of **RUNX1**  
 1083 and **MBNL1** is more commonly associated with **oligodendrocyte precursor cells**, which are in the  
 1084 process of differentiating into mature oligodendrocytes. Therefore, based on the expression profile and  
 the presence of these specific markers, the most probable cell type is **oligodendrocyte precursor cell**.  
 <analyze>  
 <answer>**oligodendrocyte precursor cell**</answer>

1085 **True Label:** choroid plexus epithelial cell  
 1086 **LLM Output:** <analyze>The given list of top 100 genes includes several markers and pathways that can help  
 1087 infer the cell type. Let's examine some of the key genes and their known associations: 1. **TRPM3**:  
 1088 This gene encodes a transient receptor potential cation channel, which is expressed in various cell  
 1089 types, including neurons and glial cells, but it is not specific to any one cell type. 2. **HTR2C**:  
 1090 This gene codes for the serotonin receptor 2C, which is expressed in neurons and some glial cells, but  
 1091 again not specific to one cell type. 3. **MALAT1**: This is a long non-coding RNA (lncRNA) known to be  
 1092 highly expressed in many cell types, including neurons and glial cells, and is involved in splicing  
 1093 regulation. 4. **RORA**: This gene is a transcription factor involved in multiple biological processes  
 1094 and is expressed in various cell types, including neurons, astrocytes, and macrophages. 5. **MSI2**:  
 1095 This gene is a marker for oligodendrocyte precursor cells (OPCs) and is highly expressed in these cells.  
 1096 6. **CFAP54**: This gene is associated with ciliary function and is expressed in various cell types,  
 1097 including some glial and epithelial cells. 7. **SLC13A4**: This gene encodes a sodium-dependent  
 1098 dicarboxylate transporter and is expressed in various tissues, including the brain. 8. **SYNE1**: This  
 1099 gene is involved in the formation of the nuclear envelope and is expressed in various cell types,  
 1100 including neurons and glial cells. 9. **ZBTB20**: This gene is a transcription factor that plays a role  
 1101 in development and is expressed in various cell types. 10. **NEAT1**: This is another lncRNA that is  
 1102 highly expressed in many cell types and is involved in nuclear organization and RNA processing. 11. **GPM6A**:  
 1103 This gene is involved in cell adhesion and is expressed in various cell types, including  
 1104 neurons and glial cells. 12. **SLC4A10**: This gene encodes a sodium bicarbonate cotransporter and is  
 1105 expressed in various tissues. 13. **DMD**: This gene is associated with Duchenne muscular dystrophy and  
 1106 is primarily expressed in skeletal muscle cells, but it is also expressed in some glial cells. 14. **TENM2**:  
 1107 This gene is involved in axon guidance and is expressed in neurons and some glial cells. 15. **POU2AF1**:  
 1108 This gene is a transcription factor that is expressed in various cell types, including  
 1109 neurons and glial cells. 16. **SLC2A12**: This gene encodes a glucose transporter and is expressed in  
 1110 various cell types, including neurons and glial cells. 17. **PLCB4**: This gene encodes a phospholipase  
 1111 C and is expressed in various cell types, including neurons and glial cells. 18. **PROS1**: This gene is  
 1112 involved in cell signaling and is expressed in various cell types, including neurons and glial cells.  
 The presence of **MSI2** is a strong indicator of **oligodendrocyte precursor cells (OPCs)**, as it is a  
 1113 well-known marker for these cells. Other genes like **TRPM3**, **HTR2C**, **SYNE1**, **TENM2**, and **SLC2A12**  
 1114 are also expressed in OPCs or related cell types. The overall expression profile suggests a  
 cell with characteristics of glial cells, particularly those involved in myelination and development.  
 Considering all of this, the most probable cell type is **oligodendrocyte precursor cell**.</analyze>  
 <answer>**oligodendrocyte precursor cell**</answer>

## 1104 R.9.2 EXAMPLES OF CORRECT ANSWERS

1105 **True Label:** oligodendrocyte  
 1106 **LLM Output:** <analyze>The given list of top 100 highly expressed genes provides a strong indication of the cell  
 1107 type based on the presence of known marker genes and functional pathways. Several of the top genes in  
 1108 this list are well-established markers of oligodendrocytes, the myelinating cells of the central nervous  
 1109 system. For example: MBP (Myelin Basic Protein) is a hallmark gene of oligodendrocytes, essential for  
 1110 myelin sheath formation. PLP1 (Proteolipid protein 1) is another critical myelin component,  
 1111 predominantly expressed in oligodendrocytes. QKI (Quaking) is a RNA-binding protein involved in  
 1112 myelination and is specifically expressed in oligodendrocytes. PCDH9 (Protocadherin 9) is also  
 1113 associated with oligodendrocytes and myelination. CTNNA3 (Catenin alpha 3) is involved in cell adhesion  
 1114 and has been linked to oligodendrocyte development. SLC24A2 (Solute carrier family 24 member 2) is  
 1115 implicated in oligodendrocyte function and myelin maintenance. Other genes like DST (Dystonin), SLC44A1,  
 1116 CDH11, CNTNAP2, and RTN4 (Nogo-A) are also known to be expressed in oligodendrocytes or are involved in  
 1117 processes related to myelination or axon-glial interactions. These genes are not typically expressed at  
 1118 high levels in other cell types such as astrocytes, macrophages, fibroblasts, or endothelial cells. The  
 1119 expression profile is highly specific to oligodendrocytes, making this the most likely cell type.</  
 1120 analyze>  
 <answer>**oligodendrocyte**</answer>

1121 **True Label:** ependymal cell  
 1122 **LLM Output:** <analyze> The top 100 highly expressed genes in this cell include several long non-coding RNAs (lncRNAs) such as MALAT1, MIR99AHG, ENSG00000231557, ENSG00000225421, LINC00609, PCAT1, LINC00882, PTX, CASC15, LINC02899, and CASC2. These lncRNAs are often associated with regulatory functions in various cell types, including glial cells and neurons. Additionally, there are several genes related to ciliary structure and function, such as DNAH9, CFAP299, DNAH6, DNAH7, DNAH11, DNAH12, CFAP54, CFAP61, CFAP44, CFAP52, CFAP418-AS1, DNAI4, and HYDIN. These genes are commonly expressed in cells with motile cilia, such as ependymal cells and some types of epithelial cells. The presence of RORA and SOX5 suggests involvement in developmental processes, which is consistent with certain glial or neuronal cell types. NRXN1 and NCAM2 are involved in cell-cell adhesion and are commonly expressed in neurons and glia. Among the cell types listed, ependymal cell is a strong candidate because of the presence of ciliary-related genes and the expression of RORA, SOX5, and NRXN1, all of which are relevant to ependymal cells, which are ciliated cells lining the ventricles of the brain and the central canal of the spinal cord. Furthermore, ependymal cells are known to express a variety of lncRNAs and have complex regulatory networks involving developmental genes. The expression profile does not strongly align with other listed cell types such as oligodendrocytes, astrocytes, fibroblasts, or macrophages, which have distinct marker genes not prominently represented here. </analyze>  
 <answer>**ependymal cell**</answer>

## A.1 MODELS

Table 7: The release dates and sources of the evaluated models.

Model	Release Date	Source
<i>Closed-source LLMs</i>		
GPT-5	2025-08	<a href="https://platform.openai.com/docs/models/gpt-5">https://platform.openai.com/docs/models/gpt-5</a>
Gemini-2.5-pro	2025-06	<a href="https://ai.google.dev/gemini-api/docs/models#gemini-2.5-pro">https://ai.google.dev/gemini-api/docs/models#gemini-2.5-pro</a>
Gemini-2.5-flash	2025-06	<a href="https://ai.google.dev/gemini-api/docs/models#gemini-2.5-flash">https://ai.google.dev/gemini-api/docs/models#gemini-2.5-flash</a>
<i>Domain-specific LLMs</i>		
cell-o1	2025-06	<a href="https://huggingface.co/ncki/Cell-o1">https://huggingface.co/ncki/Cell-o1</a>
rBio1	2025-08	<a href="https://virtualcellmodels.cziscience.com/model/rbio">https://virtualcellmodels.cziscience.com/model/rbio</a>
C2S	2025-04	<a href="https://huggingface.co/vandijklab/C2S-Scale-Pythia-1b-pt">https://huggingface.co/vandijklab/C2S-Scale-Pythia-1b-pt</a>
<i>Open-source LLMs</i>		
Deepseek-V3.1	2025-09	<a href="https://huggingface.co/deepseek-ai/DeepSeek-V3.1">https://huggingface.co/deepseek-ai/DeepSeek-V3.1</a>
gpt-oss-120B	2025-08	<a href="https://huggingface.co/openai/gpt-oss-120b">https://huggingface.co/openai/gpt-oss-120b</a>
gpt-oss-20B	2025-08	<a href="https://huggingface.co/openai/gpt-oss-20b">https://huggingface.co/openai/gpt-oss-20b</a>
Llama-3.1-8B	2024-12	<a href="https://huggingface.co/meta-llama/Llama-3.1-8B">https://huggingface.co/meta-llama/Llama-3.1-8B</a>
Qwen3-4B	2025-08	<a href="https://huggingface.co/Qwen/Qwen3-4B">https://huggingface.co/Qwen/Qwen3-4B</a>
Qwen3-8B	2025-08	<a href="https://huggingface.co/Qwen/Qwen3-8B">https://huggingface.co/Qwen/Qwen3-8B</a>
Qwen3-14B	2025-08	<a href="https://huggingface.co/Qwen/Qwen3-14B">https://huggingface.co/Qwen/Qwen3-14B</a>
Qwen3-30B-A3B	2025-08	<a href="https://huggingface.co/Qwen/Qwen3-30B-A3B">https://huggingface.co/Qwen/Qwen3-30B-A3B</a>
Qwen3-32B	2025-08	<a href="https://huggingface.co/Qwen/Qwen3-32B">https://huggingface.co/Qwen/Qwen3-32B</a>

## A.2 METRICS

In this appendix, we provide the formal definitions of all evaluation metrics used for CeLLM evaluation, corresponding to the three categories in the main text: classification, ranking, and foundation model-based evaluation.

### A.2.1 CLASSIFICATION METRICS

For classification-style tasks (e.g., cell type annotation and drug response prediction), we adopt standard metrics. Let TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (2)$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (3)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \quad (4)$$

**Top- $k$  overlap.** For multi-label or candidate set prediction tasks, we compute the overlap between predicted and reference sets:

$$\text{Top-}k = \frac{|\text{Top}_k(\text{pred}) \cap \text{Top}_k(\text{ref})|}{k}. \quad (5)$$

**ROUGE-1 recall.** Given a reference set  $R$  and a predicted set  $P$ , ROUGE-1 recall is defined as:

$$\text{ROUGE-1 Recall} = \frac{|\text{Unigrams}(R) \cap \text{Unigrams}(P)|}{|\text{Unigrams}(R)|}. \quad (6)$$

**Batch-level accuracy.** The proportion of batches where all predicted labels exactly match the ground truth.

1188 A.2.2 RANKING METRICS  
1189

1190 For tasks requiring ranked outputs, such as perturbation prediction or condition generation, we adopt  
1191 the following.

1192 **Spearman’s rank correlation** ( $\rho$ ). For  $n$  items, let  $r_i$  and  $s_i$  denote the ranks of item  $i$  in the  
1193 predicted and reference lists:  
1194

$$1195 \rho = 1 - \frac{6 \sum_{i=1}^n (r_i - s_i)^2}{n(n^2 - 1)}. \quad (7)$$

1196  
1197 **Normalized Discounted Cumulative Gain (NDCG@k)**. Given graded relevance scores  $rel_i$  at  
1198 rank position  $i$ , the DCG@k and NDCG@k are:  
1199

$$1200 \text{DCG@}k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i + 1)}, \quad (8)$$

$$1201 \text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}, \quad (9)$$

1202 where  $\text{IDCG@}k$  denotes the maximum achievable DCG@k (ideal ranking).  
1203

1204 In our experiments, we instantiate relevance using logarithmic gain:  $rel_i = 1/\log_2(1 + r_i)$ , where  
1205  $r_i$  is the ground-truth rank of the item at position  $i$  (or 0 if unranked).  
1206

1207 **Rank**. The Rank metric measures the average ranking of the model’s predicted answer within the  
1208 candidate list for differential expression gene classification. A lower value is better, indicating that  
1209 the model tends to select ground truth answers with higher rankings (i.e., genes more likely to be  
1210 differentially expressed).  
1211

1212 A.2.3 FOUNDATION MODEL-BASED EVALUATION  
1213

1214 Beyond statistical consistency, we also employ domain-specific foundation models (FMs) as evalu-  
1215 ators.  
1216

1217 **Geneformer-based cosine similarity**. Given a cell-sentence gene list as input (with genes not  
1218 in the Geneformer vocabulary filtered out), we extract the CLS token embedding from the final  
1219 Transformer layer of the Geneformer model and use it as the cell representation  $\mathbf{z}$ . Let  $\mathbf{z}_{\text{pred}} \in \mathbb{R}^d$   
1220 and  $\mathbf{z}_{\text{ref}} \in \mathbb{R}^d$  denote the embeddings of the predicted and reference cells obtained from a pretrained  
1221 geneformer. We compute the cosine similarity:  
1222

$$1223 \text{CosSim}(\mathbf{z}_{\text{pred}}, \mathbf{z}_{\text{ref}}) = \frac{\mathbf{z}_{\text{pred}}^\top \mathbf{z}_{\text{ref}}}{\|\mathbf{z}_{\text{pred}}\| \|\mathbf{z}_{\text{ref}}\|}. \quad (10)$$

1224 This metric quantifies the biological alignment of predictions in the latent space.  
1225

1226 **BERTScore with BioBERT**. For textual outputs (e.g., gene function annotations), we use  
1227 BioBERT embeddings (Lee, 2020). Let  $\mathbf{x} = (x_1, \dots, x_m)$  be the reference token embeddings  
1228 and  $\mathbf{y} = (y_1, \dots, y_n)$  the prediction embeddings. The pairwise cosine similarity is  
1229

$$1230 S_{ij} = \frac{x_i^\top y_j}{\|x_i\| \|y_j\|}. \quad (11)$$

1231 The BERTScore calculation adopted in this paper is as follows:  
1232

$$1233 \text{BERTScore-Recall} = \frac{1}{m} \sum_{i=1}^m \max_j S_{ij}, \quad \text{BERTScore - Precision} = \frac{1}{n} \sum_{j=1}^n \max_i S_{ij}, \quad (12)$$

$$1234 \text{BERTScore - F1} = 2 \cdot \frac{\text{BERTScore - Precision} \cdot \text{BERTScore - Recall}}{\text{BERTScore - Precision} + \text{BERTScore - Recall}}. \quad (13)$$

1235

#### 1242 A.2.4 OTHER METRICS

1243

1244 **Format Validity.** The proportion of outputs that follow the required response format.

1245

1246 **Answer Uniqueness.** The average proportion of unique cell type predictions per batch.

1247

1248 This family of metrics enables a holistic evaluation by combining correctness, ranking consistency,  
1249 and biological or semantic plausibility.

1250

#### 1251 A.3 DATASETS

1252

1253 This section introduces the datasets used in CELLM and their processing workflow. We first de-  
1254 scribe the global gene symbol standardization, then outline the reference datasets, and finally present  
1255 the construction of task-specific text datasets.

1256

##### 1257 A.3.1 GENE SYMBOL STANDARDIZATION

1258

1259 To ensure cross-dataset consistency, we established a unified gene symbol mapping. Official map-  
1260 pings from **GENCODE** and **HGNC** (Ensembl Gene ID  $\rightarrow$  HGNC Symbol) were merged, followed  
1261 by deduplication, case normalization, alias resolution, and one-to-many conflict handling, where  
1262 only the primary symbol was retained. Importantly, many raw datasets contain *previous symbols*  
1263 and *alias symbols*, which can lead to ambiguities or even gene duplication across studies. Our  
1264 standardization pipeline resolves these issues by mapping all variants to the authoritative **HGNC**  
1265 **primary symbol**, thereby producing a clean and human-readable representation of gene names. All  
1266 datasets were standardized prior to downstream processing; unmapped genes were preserved but not  
1267 discarded. Compared to earlier approaches (which did not strictly standardize gene symbols), this  
1268 procedure substantially reduces noise and ensures the construction of text samples that are maxi-  
1269 mally suitable for LLMs (Frankish et al., 2025; Tweedie et al., 2021).

1269

##### 1270 A.3.2 REFERENCE DATASETS

1271

1272 We employed a diverse set of authoritative resources covering single-cell, spatial, multi-omics, and  
1273 perturbation contexts:

1273

- 1274 • **Virtual Cell Challenge (VCC) (cell perturbation):** The dataset (Roohani et al., 2025) was  
1275 generated using single-cell functional genomics, producing approximately 300,000 single-  
1276 cell RNA-seq profiles by silencing 300 carefully selected genes using CRISPR interference  
1277 (CRISPRi). To obtain the single-cell gene expression profiles, 10x Genomics GEM-X Flex  
and Illumina sequencing were utilized.

1278

- 1279 • **Gene annotation dataset (gene Annotation):** The benchmark dataset comprises 50 genes  
1280 annotations. These genes annotations were built based on a search of the GO database to  
1281 test the model’s fundamental biological knowledge.

1281

- 1282 • **Brain (single-cell annotation):** BICCN/Allen whole-brain atlas (MERFISH & scRNA-  
1283 seq) (Yao, 2023; Zhang, 2023).

1283

- 1284 • **Myeloid (single-cell annotation):** 10x PBMC 68k (Donor A) (10x Genomics, 2017;  
1285 Zheng, 2017).

1285

- 1286 • **Organoid (single-cell annotation):** Human brain organoid scRNA-seq datasets (Velasco,  
1287 2019; He, 2024).

1288

- 1289 • **CellPuzzles (batch-level annotation):** A multi-batch benchmark built on CELLXGENE  
1290 Discover/Census resources (Yang & Gomez, 2024; cel, 2025).

1290

- 1291 • **Brain1 (spatial cell annotation):** Brain slices from the STOFM benchmark (Zhao et al.,  
1292 2025).

1292

- 1293 • **CellVerse (drug response prediction):** A benchmark suite of cell/drug tasks; we use the  
1294 Drug Response Prediction subset (Fang, 2025b).

1294

- 1295 • **Lung / Liver / Blood ( condition cell generation):** HLCA (lung), Aizarani liver atlas  
(liver), and PBMC 68k (blood) (Sikkema, 2023; Aizarani, 2019; 10x Genomics, 2017).

- **CITE-BMMC (multi-omics integration):** OpenProblems multimodal benchmark (Luecken, 2021; 2025).

### A.3.3 TASK-SPECIFIC TEXT DATA GENERATION

Based on the standardized symbols and reference datasets, we unify diverse tasks into textual input–output interfaces, with specific evaluation objectives and metrics.

**Cell Annotation.** *Goal:* Assign biological cell types to single-cell transcriptomes. *Single-cell annotation:* Brain, Myeloid, and Organoid datasets were transformed into gene-ranked lists (“gene sentences”), with hierarchical cell-type labels as supervision (Yao, 2023; Zhang, 2023; 10x Genomics, 2017; Zheng, 2017; Velasco, 2019; He, 2024). Metrics include **accuracy**, **recall**, **batch-level accuracy**, **uniqueness**, and **response length**, with optional text metrics (ROUGE, BERTScore). *Batch-level annotation:* Using CellPuzzles, we aggregate top-genes and metadata within each batch, requiring globally consistent batch-level predictions; report both cell- and batch-level metrics (Yang & Gomez, 2024; cel, 2025). *Spatial annotation:* Using Brain1 (SToFM), each cell/spot is represented by a gene sentence augmented with absolute coordinates and neighborhood graphs; evaluate accuracy/recall and spatial consistency.

**Drug Response.** *Goal:* Predict cellular response to drug perturbations or answer drug-related queries. Based on CellVerse DRP, inputs combine drug/perturbation context with cell samples, rewritten as QA-style or classification prompts; outputs are categorical responses or explanatory text (Fang, 2025b). Metrics: **accuracy**, **recall** (classification); **ROUGE**, **BERTScore** (text).

**Gene Annotation.** *Goal:* Explain gene function using authoritative knowledge bases. We employ Gene Ontology (GO) and NCBI resources (AmiGO portal; NCBI Gene; GEO). Inputs are gene lists or queries; outputs are functional categories, term definitions, or enrichment results. Metrics: text similarity (**ROUGE**, **BERTScore**) and structured matching (strict/soft) (The Gene Ontology Consortium, 2023; Carbon et al., 2009; Sayers et al., 2024; Clough et al., 2024).

**Cell Generation.** *Goal:* Learn conditional generation of cell expression profiles. Using HLCA (lung), Aizarani liver atlas, and PBMC 68k, inputs are (organ, cell type, condition) and outputs are bulk-like mean profiles rendered as ranked gene lists. Metrics: **top- $k$  overlap**, **Spearman correlation**, **nDCG**, **cosine similarity**, plus optional distributional metrics (KS, EMD) (Sikkema, 2023; Aizarani, 2019; 10x Genomics, 2017).

**Perturbation Prediction.** *Goal:* Model perturbation outcomes with curated perturbation resources. We consider four sub-tasks: (i) *DEG prediction*, (ii) *direction prediction* (up/down), (iii) *condition prediction*, and (iv) *perturbed cell generation*. Inputs are (cell type + perturbation); outputs are predictions or generated profiles. We reference SCPERTURB and the NeurIPS OP3 perturbation challenge as representative sources for data/task design (Peidli, 2024; Szałata, 2024). Metrics include ranking (**Spearman@ $k$** , **nDCG**, **top- $k$  overlap**, **cosine similarity**) and classification (**accuracy**, **recall**).

**Multi-omics Integration.** *Goal:* Perform cross-modality prediction, matching, and joint embedding while correcting batch effects. Using OpenProblems, tasks include (i) *modality prediction* (RNA  $\leftrightarrow$  ADT/ATAC), (ii) *modality matching*, and (iii) *joint embedding*. Evaluation covers prediction accuracy/correlation, matching precision/recall, and embedding quality (structure preservation and batch removal) (Luecken, 2021; 2025).

## A.4 EXPERIMENTAL RESULTS

Table 8: Results of the single cell annotation task.

Setting	Brain				Myeloid				Organoid			
Metric	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
<i>Closed-source LLMs</i>												
GPT-5	81.06	65.17	67.10	67.50	12.37	18.43	12.01	20.00	65.26	<b>65.20</b>	<b>63.58</b>	<b>71.50</b>
Gemini-2.5-pro	80.98	65.57	66.61	67.50	17.12	20.73	13.67	20.00	<b>65.94</b>	63.87	62.57	66.00
Gemini-2.5-flash	76.16	45.57	50.13	49.00	13.97	16.28	11.36	19.00	61.92	44.85	46.10	48.00
<i>Domain-specific LLMs</i>												
cell-o1	54.31	47.90	44.00	47.00	13.83	<b>22.20</b>	14.67	21.00	22.31	20.40	17.29	23.50
rBio1	18.31	16.72	14.25	16.00	1.45	4.80	1.97	5.50	11.15	13.81	11.93	15.00
C2S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>												
Deepseek-V3.1	64.51	19.90	25.77	20.50	<b>22.77</b>	8.52	10.28	10.00	52.88	22.58	30.97	24.50
gpt-oss-120B	81.27	<b>75.31</b>	<b>76.08</b>	<b>75.50</b>	19.91	22.25	<b>18.09</b>	<b>26.00</b>	62.14	60.53	58.02	65.00
gpt-oss-20B	<b>86.16</b>	39.33	50.51	41.00	22.47	18.05	16.49	20.50	60.57	38.21	44.61	41.00
Llama-3.1-8B	40.73	20.55	19.38	20.00	2.73	6.42	3.37	7.00	24.62	21.16	18.04	25.00
Qwen3-4B	37.10	21.73	16.69	22.50	10.34	4.44	3.62	5.00	15.85	14.71	12.89	17.00
Qwen3-4B-think	43.12	32.60	26.88	34.00	4.60	8.60	3.01	9.00	21.66	19.52	18.27	22.50
Qwen3-8B	37.11	28.15	23.45	29.50	6.05	8.69	5.12	9.50	21.74	22.57	19.74	27.00
Qwen3-8B-think	52.65	34.75	30.45	37.00	10.39	13.80	8.44	12.00	37.21	30.61	24.24	33.00
Qwen3-14B	68.61	48.08	46.88	49.00	11.37	16.46	9.68	15.50	28.76	28.85	23.34	35.00
Qwen3-14B-think	70.06	56.40	55.34	57.50	13.46	20.09	11.96	19.00	43.34	38.77	35.67	44.50
Qwen3-30B-A3B	49.90	35.50	33.70	37.50	12.30	12.28	9.37	13.50	26.64	29.09	22.75	35.50
Qwen3-30B-A3B-think	60.33	49.45	48.15	52.00	19.12	18.06	12.95	20.00	34.13	31.75	25.55	38.50
Qwen3-32B	51.70	43.96	42.45	45.00	8.84	13.37	9.30	15.00	31.04	31.77	28.43	37.50
Qwen3-32B-think	62.72	56.27	53.70	58.00	11.12	14.84	9.69	16.50	31.99	35.98	30.91	43.00

Table 9: Results of the cell puzzles task.

Model	Cell-level Acc	Batch-level Acc	Uniqueness	Format
<i>Closed-source LLMs</i>				
GPT-5	63.04	14.16	<b>100.00</b>	<b>100.00</b>
Gemini-2.5-pro	66.35	20.64	97.17	97.17
Gemini-2.5-flash	34.20	7.12	81.05	81.10
<i>Domain-specific LLMs</i>				
cell-o1	<b>71.84</b>	<b>31.14</b>	99.88	<b>100.00</b>
rBio-1-3B	6.28	0.00	52.24	82.28
C2S-1B	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>				
Deepseek-V3.1	58.85	15.16	92.24	92.33
gpt-oss-120B	60.66	18.08	94.77	94.79
gpt-oss-20B	2.13	0.27	4.46	4.47
Llama-3.1-8B	2.70	0.00	7.05	15.71
Qwen3-4B-think	23.82	0.09	71.15	97.81
Qwen3-8B-think	31.21	1.37	83.71	99.63
Qwen3-30B-A3B-think	37.07	1.74	78.65	96.71
Qwen3-32B-think	42.68	4.02	94.75	99.73

Table 10: Results of the spatial cell annotation task.

Setting Metric	non				top5				top10				sample5			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
<i>Closed-source LLMs</i>																
GPT-5	24.00	27.37	21.29	21.02	27.00	33.72	29.00	<b>25.99</b>	20.00	30.68	23.27	21.18	19.00	25.79	22.82	19.19
Gemini-2.5-pro	32.00	26.76	30.00	25.54	23.00	25.96	23.20	21.17	23.00	22.78	27.19	21.58	29.00	27.50	<b>28.75</b>	<b>25.47</b>
Gemini-2.5-flash	27.00	24.06	26.01	23.26	10.00	19.00	11.96	13.40	10.00	21.67	8.94	10.85	12.00	25.55	11.36	13.72
<i>Domain-specific LLMs</i>																
cell-ol	11.00	14.45	8.23	9.11	6.00	7.09	8.82	4.54	7.00	4.70	4.41	4.21	10.00	10.14	6.44	6.39
rBio1	16.00	8.21	7.19	5.47	19.00	7.21	6.98	4.31	15.00	1.55	4.93	2.35	18.00	3.91	6.63	4.05
C2S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>																
Deepseek-V3.1	12.00	18.41	10.90	12.85	11.00	21.77	10.76	12.84	7.00	15.92	8.98	7.44	12.00	20.49	7.87	10.34
gpt-oss-120B	32.00	<b>37.17</b>	24.76	<b>25.58</b>	30.00	<b>36.15</b>	19.48	21.84	<b>35.00</b>	<b>42.62</b>	<b>25.25</b>	<b>27.63</b>	<b>33.00</b>	<b>32.42</b>	22.71	22.74
gpt-oss-20B	21.00	28.95	14.38	18.21	12.00	11.97	7.26	8.01	22.00	30.85	14.48	18.54	18.00	23.52	12.78	16.04
Llama-3.1-8B	11.00	3.10	3.96	3.07	14.00	5.53	7.43	5.08	15.00	9.19	10.42	8.30	11.00	5.17	6.49	4.53
Qwen3-4B	22.00	18.03	12.69	10.25	18.00	4.58	7.85	4.44	13.00	11.50	6.77	6.63	16.00	6.77	6.49	4.12
Qwen3-4B-think	31.00	19.88	16.89	13.95	22.00	14.52	10.47	8.73	11.00	6.70	4.44	3.74	15.00	6.00	6.88	5.43
Qwen3-8B	23.00	16.27	13.79	12.10	16.00	6.94	8.99	6.46	19.00	11.74	9.53	7.98	18.00	12.66	10.21	8.56
Qwen3-8B-think	23.00	10.52	14.20	11.07	22.00	11.17	12.26	9.83	24.00	13.25	13.23	11.56	24.00	15.78	12.99	11.89
Qwen3-30B-A3B	25.00	15.48	17.62	13.46	20.00	8.95	11.18	9.22	20.00	16.77	14.88	12.69	25.00	22.14	16.65	16.25
Qwen3-30B-A3B-think	<b>33.00</b>	21.57	<b>27.42</b>	20.29	32.00	17.44	22.05	17.21	27.00	18.68	20.33	16.81	30.00	19.84	22.31	18.33
Qwen3-32B	29.00	23.28	23.34	21.64	23.00	19.88	16.29	15.67	25.00	18.98	18.91	16.79	20.00	19.69	22.78	17.21
Qwen3-32B-think	32.00	27.56	24.66	20.76	<b>33.00</b>	21.78	<b>24.45</b>	20.00	31.00	21.67	21.08	18.49	31.00	19.65	22.05	18.22

Table 11: Results of the multi-omics cell annotation task.

Setting Metric	scCite				scRNA				Integrated			
	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
<i>Closed-source LLMs</i>												
GPT-5	16.90	16.90	11.58	18.50	19.37	23.72	18.95	48.00	29.88	31.96	26.71	49.00
Gemini-2.5-pro	30.39	<b>30.68</b>	<b>26.35</b>	<b>33.00</b>	<b>31.27</b>	24.60	24.84	45.50	<b>39.05</b>	<b>34.68</b>	<b>34.05</b>	<b>53.50</b>
Gemini-2.5-flash	<b>32.50</b>	15.21	18.86	12.50	24.92	12.47	15.22	24.00	29.49	17.79	19.46	23.50
<i>Domain-specific LLMs</i>												
cell-ol	1.81	1.01	0.70	4.00	14.27	18.50	14.81	42.50	7.97	7.95	7.32	24.00
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
C2S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>												
Deepseek-V3.1	17.78	19.01	15.20	29.00	17.08	6.81	8.77	17.00	18.90	12.64	13.05	21.50
gpt-oss-120B	3.83	4.71	3.08	4.00	22.18	<b>24.68</b>	<b>22.83</b>	<b>52.50</b>	32.93	30.81	27.63	44.00
gpt-oss-20B	10.49	2.68	3.46	3.00	21.86	15.74	16.37	42.00	20.73	12.82	12.98	30.50
Llama-3.1-8B	5.60	9.27	4.58	10.00	7.86	10.66	7.83	25.50	13.39	12.85	9.88	19.00
Qwen3-4B	3.44	2.69	2.24	7.00	4.50	9.81	5.15	24.00	2.19	1.58	1.41	3.50
Qwen3-4B-think	6.82	1.61	1.94	3.00	10.35	7.86	6.89	22.00	3.77	1.61	1.54	3.00
Qwen3-8B	2.24	2.46	1.56	2.50	9.94	7.72	7.31	25.50	9.04	3.54	3.90	9.00
Qwen3-8B-think	9.22	7.29	4.22	7.50	13.30	14.55	12.08	31.50	9.66	12.86	8.98	20.00
Qwen3-14B	9.51	9.51	6.50	12.00	10.83	14.06	8.92	30.50	15.39	7.25	7.82	14.50
Qwen3-14B-think	10.33	6.25	6.15	7.50	17.97	13.18	12.23	37.00	19.42	12.43	10.60	25.00
Qwen3-30B-A3B	5.72	7.70	4.73	10.00	9.53	10.89	9.05	37.50	9.77	7.94	7.22	26.00
Qwen3-30B-A3B-think	1.44	1.69	1.17	3.50	11.48	14.86	10.70	38.00	14.21	17.88	12.77	33.50
Qwen3-32B	9.26	10.88	5.91	9.50	12.84	12.79	10.80	36.50	12.52	12.60	8.56	23.50
Qwen3-32B-think	8.54	11.04	6.39	10.00	10.21	15.94	11.04	40.50	10.77	14.74	9.80	32.00

Table 12: Results of the differential expression task (CLS &amp; QA).

Setting	CLS				QA									
Model	CLS		Overall				Pos				Neg			
	Rank	Acc	Recall	Prec	F1	Acc	Recall	Prec	F1	Acc	Recall	Prec	F1	Acc
<i>Closed-source LLMs</i>														
GPT-5	1.98	<b>49.20</b>	<b>65.30</b>	<b>65.30</b>	72.82	<b>62.19</b>	36.60	85.92	85.92	51.33	94.00	59.72	59.72	73.04
Gemini-2.5-pro	2.07	44.20	58.20	58.20	67.81	51.89	<b>94.60</b>	54.87	54.87	<b>69.46</b>	21.80	<b>80.74</b>	<b>80.74</b>	34.33
Gemini-2.5-flash	2.16	34.60	55.40	55.40	63.21	51.72	86.20	54.49	54.49	66.77	24.60	71.93	71.93	36.66
<i>Domain-specific LLMs</i>														
cell-o1	2.71	21.00	49.30	49.30	46.85	37.06	93.40	49.63	49.63	64.82	5.20	44.07	44.07	9.30
rBiol	2.52	28.00	51.60	51.60	52.45	47.02	22.20	53.88	53.88	31.44	81.00	51.01	51.01	62.60
C2S	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>														
Deepseek-V3.1	2.14	11.40	20.50	20.50	73.29	29.25	9.20	85.19	85.19	16.61	31.80	61.39	61.39	41.90
gpt-oss-120B	<b>1.95</b>	47.20	58.70	58.70	<b>75.47</b>	51.13	19.00	<b>95.96</b>	<b>95.96</b>	31.72	<b>98.40</b>	54.97	54.97	70.54
gpt-oss-20B	1.99	26.60	51.30	51.30	70.50	50.00	21.20	85.48	85.48	33.97	81.40	55.53	55.53	66.02
Llama-3.1-8B	2.39	28.20	51.20	51.20	52.23	46.40	20.80	53.33	53.33	29.93	81.60	51.13	51.13	62.87
Qwen3-4B	2.34	33.40	56.50	56.50	58.08	54.33	34.60	61.57	61.57	44.30	78.40	54.60	54.60	64.37
Qwen3-4B-think	2.41	30.40	56.80	56.80	59.03	54.89	35.00	62.95	62.95	44.99	78.60	55.12	55.12	64.80
Qwen3-8B	2.31	35.40	58.60	58.60	58.64	58.56	55.40	59.19	59.19	57.23	61.80	58.08	58.08	59.88
Qwen3-8B-think	2.43	31.60	59.40	59.40	65.39	55.02	28.20	75.00	75.00	40.99	90.60	55.79	55.79	69.05
Qwen3-30B-A3B	2.30	36.60	55.40	55.40	60.83	49.01	20.00	68.49	68.49	30.96	90.80	53.16	53.16	67.06
Qwen3-30B-A3B-think	2.32	34.80	56.00	56.00	70.38	46.58	14.00	87.50	87.50	24.14	98.00	53.26	53.26	<b>98.00</b>
Qwen3-32B	2.33	34.40	57.30	57.30	66.92	50.23	19.60	79.67	79.67	31.46	95.00	54.16	54.16	68.99
Qwen3-32B-think	2.33	34.80	55.90	55.90	65.96	47.65	16.20	78.64	78.64	26.87	95.60	53.29	53.29	68.43

Table 13: Results of the direction of change task.

Model	Overall				Increase				Decrease			
	Acc	Rec	Prec	F1	Acc	Rec	Prec	F1	Acc	Rec	Prec	F1
<i>Closed-source LLMs</i>												
GPT-5	55.20	<b>55.80</b>	55.84	55.19	62.33	49.27	50.54	61.14	49.27	61.14	<b>61.14</b>	54.56
Gemini-2.5-pro	52.60	53.57	54.12	52.66	64.10	43.04	48.74	59.49	43.04	59.49	59.49	49.95
Gemini-2.5-flash	49.00	49.49	56.60	52.24	54.85	44.14	50.92	<b>62.27</b>	44.14	62.27	62.27	51.66
<i>Domain-specific LLMs</i>												
cell-o1	46.10	46.86	47.93	46.55	55.07	38.64	43.63	52.23	38.64	52.23	52.23	44.42
rBiol	53.20	52.41	52.80	52.50	43.83	60.99	48.77	56.83	60.99	56.83	56.83	58.83
C2S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>												
Deepseek-V3.1	14.20	14.86	54.17	21.71	22.03	7.69	45.66	62.69	7.69	62.69	62.69	13.70
gpt-oss-120B	55.40	55.00	55.57	<b>55.28</b>	50.66	59.34	51.69	59.45	59.34	59.45	59.45	59.40
gpt-oss-20B	49.00	48.10	55.82	51.45	38.33	57.88	51.79	59.85	57.88	59.85	59.85	58.85
Llama-3.1-8B	39.70	37.67	47.81	39.59	15.64	59.71	42.26	53.36	59.71	53.36	53.36	56.35
Qwen3-4B	52.70	50.75	52.45	49.87	29.52	71.98	49.08	55.82	71.98	55.82	55.82	62.88
Qwen3-4B-think	53.10	52.75	55.06	53.88	48.90	56.59	50.92	59.20	56.59	59.20	59.20	57.87
Qwen3-8B	53.30	51.93	53.02	51.78	37.00	66.85	49.70	56.33	66.85	56.33	56.33	61.14
Qwen3-8B-think	55.20	54.89	54.88	54.88	51.54	58.24	50.65	59.11	58.24	59.11	59.11	58.67
Qwen3-30B-A3B	<b>56.40</b>	53.80	<b>57.58</b>	51.57	25.55	<b>82.05</b>	<b>57.43</b>	57.73	82.05	57.73	57.73	<b>67.78</b>
Qwen3-30B-A3B-think	53.40	52.04	52.22	51.54	37.22	66.85	48.29	56.15	66.85	56.15	56.15	61.04
Qwen3-32B	51.80	51.72	51.96	51.78	50.88	52.56	47.53	56.39	52.56	56.39	56.39	54.41
Qwen3-32B-think	53.30	<b>54.27</b>	54.41	53.16	<b>64.76</b>	43.77	48.92	59.90	<b>64.76</b>	59.90	59.90	50.58

Table 14: Results of the perturbation prediction with HVGs task.

CLS		QA									
Model	Acc	Acc	Recall	Prec	F1	No Rec	Yes Rec	No Prec	Yes Prec	No F1	Yes F1
<i>Closed-source LLMs</i>											
GPT-5	47.33	<b>53.67</b>	<b>53.67</b>	54.80	<b>50.75</b>	78.00	29.33	52.47	57.14	62.73	38.77
Gemini-2.5-pro	23.33	39.33	39.33	<b>60.62</b>	45.17	24.00	54.67	<b>65.45</b>	55.78	35.12	55.22
Gemini-2.5-flash	7.33	11.67	11.67	55.74	19.02	8.67	14.67	59.09	52.38	15.12	22.92
<i>Domain-specific LLMs</i>											
cell-o1	47.33	51.33	51.33	52.41	47.56	24.00	<b>78.67</b>	53.73	51.08	33.18	<b>61.94</b>
rBio1	52.67	46.33	46.33	45.66	44.17	66.00	26.67	47.37	43.96	55.15	33.20
C2S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>											
Deepseek-V3.1	18.0	17.67	17.67	35.71	22.21	34.67	0.67	57.14	14.29	43.15	1.27
gpt-oss-120B	47.33	49.0	49.0	59.75	39.16	90.67	7.33	50.75	<b>68.75</b>	65.07	13.25
gpt-oss-20B	8.67	23.33	23.33	57.02	26.39	44.67	2.00	54.03	60.00	48.91	3.87
Llama-3.1-8B	50.67	48.00	48.00	42.12	35.54	92.00	4.00	48.94	35.29	63.89	7.19
Qwen3-4B	39.33	48.33	48.33	44.73	36.79	91.33	5.33	49.46	40.0	64.17	9.41
Qwen3-4B-think	7.33	29.00	29.00	39.74	28.64	54.67	3.33	48.24	31.25	51.25	6.02
Qwen3-8B	49.33	50.0	50.0	50.54	44.96	80.67	19.33	50.21	50.88	61.89	28.02
Qwen3-8B-think	48.00	51.33	51.33	57.49	38.74	96.67	6.00	50.70	64.29	<b>66.51</b>	10.98
Qwen3-30B-A3B	46.67	48.00	48.00	39.65	34.21	94.00	2.00	49.30	30.00	64.68	3.75
Qwen3-30B-A3B-think	48.67	49.67	49.67	24.92	33.18	<b>99.33</b>	0.00	49.83	0.00	66.37	0.00
Qwen3-32B	50.67	49.67	49.67	48.91	38.75	92.00	7.33	50.00	47.83	64.79	12.72
Qwen3-32B-think	<b>54.67</b>	50.0	50.0	50.0	38.87	92.67	7.33	50.0	50.0	64.95	12.79

Table 15: Results of the perturbed cell generation task.

Model	Cosine Sim.	Top@50	Top@100	nDCG@50	nDCG@100	Spearman@50	Spearman@100
<i>Closed-source LLMs</i>							
GPT-5	99.00	77.57	83.02	96.55	96.90	72.62	76.47
Gemini-2.5-pro	95.85	42.23	51.08	53.72	57.62	63.18	70.95
Gemini-2.5-flash	81.03	6.44	19.69	12.69	24.24	11.35	13.23
<i>Domain-specific LLMs</i>							
cell-o1	74.31	23.89	18.55	54.12	37.35	62.48	62.22
rBio1	77.66	39.25	35.17	53.58	46.86	66.47	65.60
C2S	-	0.00	0.00	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>							
Deepseek-V3.1	<b>99.23</b>	85.88	<b>89.44</b>	98.41	<b>98.75</b>	86.66	89.56
gpt-oss-120B	94.45	6.39	20.29	9.77	18.64	24.68	27.89
gpt-oss-20B	78.87	5.48	10.91	6.54	10.15	47.11	34.78
Llama-3.1-8B	82.76	26.20	33.68	45.45	46.15	43.02	52.17
Qwen3-4B	90.66	70.43	71.55	80.59	78.94	83.36	83.84
Qwen3-4B-think	68.65	22.25	20.57	38.55	29.91	50.49	50.46
Qwen3-8B	89.04	34.20	37.75	58.55	51.80	50.80	41.84
Qwen3-8B-think	76.36	18.52	17.97	47.00	34.90	33.60	37.40
Qwen3-30B-A3B	91.09	53.44	53.02	59.25	58.70	73.01	72.06
Qwen3-30B-A3B-think	74.84	27.35	25.59	47.52	37.76	35.05	43.03
Qwen3-32B	97.45	72.21	72.61	90.39	85.71	81.08	79.18
Qwen3-32B-think	97.73	<b>86.07</b>	87.27	<b>99.02</b>	96.94	<b>88.85</b>	<b>89.86</b>

Table 16: Results of the drug response task.

Setting	Overall				Sensitive			Resistant		
Metric	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Precision	Recall	F1
<i>Closed-source LLMs</i>										
GPT-5	24.68	48.75	32.77	48.75	0.00	<b>97.50</b>	0.00	49.37	0.00	65.55
Gemini-2.5-pro	41.51	49.38	34.12	49.38	1.25	<b>97.50</b>	33.33	49.68	2.41	<b>65.82</b>
Gemini-2.5-flash	40.30	16.88	21.82	16.88	5.00	28.75	36.36	44.23	8.79	34.85
<i>Domain-specific LLMs</i>										
cell-o1	47.58	30.63	36.49	30.63	18.75	42.50	42.86	52.31	26.09	46.90
rBio1	48.48	46.88	44.17	46.88	<b>71.25</b>	22.50	48.31	48.65	<b>57.58</b>	30.77
C2S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>										
Deepseek-V3.1	50.00	<b>50.00</b>	39.75	50.00	8.75	91.25	50.00	50.00	14.89	64.60
gpt-oss-120B	<b>55.16</b>	50.62	36.72	<b>50.62</b>	3.75	<b>97.50</b>	<b>60.00</b>	50.32	7.06	66.38
gpt-oss-20B	52.01	<b>50.00</b>	41.70	50.00	11.25	88.75	52.94	51.08	18.56	64.84
Llama-3.1-8B	53.57	26.88	34.96	26.88	20.00	33.75	57.14	50.00	29.63	40.30
Qwen3-4B	47.26	41.88	37.66	41.88	11.25	72.50	47.37	47.15	18.18	57.14
Qwen3-4B-think	41.01	43.75	36.44	43.75	8.75	78.75	35.00	47.01	14.00	58.88
Qwen3-8B-think	44.67	48.75	36.62	48.75	5.00	92.50	40.00	49.33	8.89	64.35
Qwen3-14B	42.06	43.75	39.83	43.75	17.50	70.00	37.84	46.28	23.93	55.72
Qwen3-14B-think	43.46	46.25	38.83	46.25	11.25	81.25	39.13	47.79	17.48	60.19
Qwen3-30B-A3B	43.63	16.88	24.32	16.88	16.25	17.50	44.83	42.42	23.85	24.78
Qwen3-30B-A3B-think	38.26	46.25	35.23	46.25	5.00	87.50	28.57	47.95	8.51	61.95
Qwen3-32B	50.82	45.00	<b>46.24</b>	45.00	31.25	58.75	53.19	48.45	<b>39.37</b>	53.11
Qwen3-32B-think	35.27	46.25	33.69	46.25	2.50	90.00	22.22	48.32	4.49	62.88

Table 17: Results of the conditional cell generation task.

Model	Cosine Sim.	Top@50	Top@100	nDCG@50	nDCG@100	Spearman@50	Spearman@100
<i>Closed-source LLMs</i>							
GPT-5	93.56	24.21	33.82	<b>97.74</b>	<b>94.70</b>	25.93	13.18
Gemini-2.5-pro	<b>93.75</b>	<b>36.47</b>	<b>40.32</b>	86.30	89.01	41.61	34.72
Gemini-2.5-flash	74.70	7.37	9.50	32.23	31.72	11.33	12.70
<i>Domain-specific LLMs</i>							
cell-o1	60.99	4.37	3.47	47.83	32.10	-5.67	26.30
rBio1	63.47	0.53	0.66	52.81	35.38	50.00	23.33
C2S	-	0.00	0.00	0.00	0.00	0.00	0.00
<i>Open-source LLMs</i>							
Deepseek-V3.1	89.80	11.63	17.00	62.25	68.36	6.56	18.75
gpt-oss-120B	91.18	14.58	22.45	65.60	66.85	8.03	18.77
gpt-oss-20B	84.03	5.58	10.71	37.47	36.70	17.62	15.65
Llama-3.1-8B	60.86	2.11	1.39	55.08	34.81	17.43	31.19
Qwen3-4B	68.13	0.95	0.68	44.01	28.95	15.00	<b>52.50</b>
Qwen3-4B-think	65.75	1.32	1.32	41.03	29.15	<b>55.43</b>	32.08
Qwen3-8B	80.50	2.89	2.82	70.94	61.18	6.73	30.85
Qwen3-8B-think	71.68	3.37	2.84	70.68	49.63	-17.55	-2.06
Qwen3-30B-A3B	75.62	4.11	4.68	75.96	57.56	34.70	29.42
Qwen3-30B-A3B-think	77.67	2.95	2.71	75.18	54.93	-4.71	6.62
Qwen3-32B	89.78	6.37	8.24	84.37	79.37	3.64	19.58
Qwen3-32B-think	80.91	7.00	8.34	84.17	63.65	-5.21	6.11

Table 18: Results of the gene annotation task.

Model	BERTScore P	BERTScore R	BERTScore F1	ROUGE-1 R
<i>Closed-source LLMs</i>				
GPT-5	67.19	68.98	68.00	43.48
Gemini-2.5-pro	67.61	68.87	68.15	58.70
Gemini-2.5-flash	68.10	69.18	68.56	41.30
<i>Domain-specific LLMs</i>				
cell-o1	64.47	68.00	65.95	63.04
rBio1	<b>68.91</b>	69.62	<b>69.17</b>	30.43
C2S	68.86	59.47	63.74	19.57
<i>Open-source LLMs</i>				
Deepseek-V3.1	68.00	<b>69.86</b>	68.83	41.30
gpt-oss-120B	65.34	68.41	66.73	71.74
gpt-oss-20B	61.74	66.28	63.82	<b>95.65</b>
Llama-3.1-8B	65.94	69.03	67.37	63.04
Qwen3-4B	58.97	67.01	62.66	41.30
Qwen3-4B-think	66.80	68.80	67.70	52.17
Qwen3-8B	68.67	69.84	69.18	52.17
Qwen3-8B-think	67.75	68.90	68.25	43.48
Qwen3-30B-A3B	67.97	68.92	68.35	43.48
Qwen3-30B-A3B-think	67.82	69.17	68.40	34.78
Qwen3-32B	67.78	68.90	68.25	56.52
Qwen3-32B-think	67.35	69.71	68.43	39.13

Table 19: Results of the few shot ablation task.

Model	leak				mask			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
gpt-5-chat-latest	<b>72.00</b>	<b>69.26</b>	<b>78.74</b>	<b>71.65</b>	63.00	60.51	<b>72.58</b>	<b>61.36</b>
cell_o1	51.00	52.18	54.44	48.97	<b>54.50</b>	<b>55.33</b>	61.09	<b>53.23</b>
qwen3-32b	48.50	46.59	56.09	46.61	37.00	35.10	42.89	33.02
qwen3-8b	36.50	35.15	47.99	32.23	31.00	29.62	40.94	26.93

Table 20: Results of the robustness ablation task.

Model	0%				20%				40%			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
gpt-5-chat-latest	<b>67.50</b>	<b>65.17</b>	<b>81.06</b>	<b>67.10</b>	<b>68.00</b>	<b>66.39</b>	<b>81.12</b>	<b>67.96</b>	<b>68.50</b>	<b>66.59</b>	<b>81.43</b>	<b>68.45</b>
cell_o1	47.00	47.90	54.31	44.00	45.00	46.04	56.56	43.03	43.00	43.76	53.63	40.89
qwen3-32b	45.00	43.96	51.70	42.45	44.00	43.47	50.77	41.44	48.00	47.37	58.63	46.23
qwen3-8b	29.50	28.15	37.11	23.45	35.00	34.17	41.90	29.97	34.50	33.20	40.56	29.91
Model	60%				80%				100%			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
gpt-5-chat-latest	<b>67.50</b>	<b>65.77</b>	79.62	<b>68.53</b>	<b>68.50</b>	<b>66.53</b>	78.02	<b>68.85</b>	<b>70.00</b>	<b>68.15</b>	<b>80.36</b>	<b>70.52</b>
cell_o1	47.00	48.24	53.14	45.38	44.50	44.61	<b>55.94</b>	42.40	46.50	47.35	47.83	42.75
qwen3-32b	44.00	42.78	<b>56.99</b>	43.00	41.00	39.99	50.25	37.84	43.50	42.94	51.91	41.92
qwen3-8b	33.50	32.48	38.05	27.50	30.50	29.28	32.75	24.54	31.00	29.37	35.07	25.99

Table 21: Results of the gene list length ablation task.

Model	20				50				100				200				500			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
gpt-5-chat-latest	<b>63.50</b>	<b>63.10</b>	<b>75.36</b>	<b>63.14</b>	<b>68.50</b>	<b>66.77</b>	<b>80.64</b>	<b>67.79</b>	<b>71.00</b>	<b>68.19</b>	<b>80.61</b>	<b>69.04</b>	<b>75.00</b>	<b>71.42</b>	<b>84.75</b>	<b>72.32</b>	<b>71.50</b>	<b>67.81</b>	<b>80.76</b>	<b>68.89</b>
cell_o1	41.00	43.20	52.29	39.19	46.50	47.48	57.17	45.05	47.50	47.97	47.83	42.87	46.00	46.75	52.81	42.42	42.00	42.25	47.26	38.66
qwen3-32b	48.00	47.41	57.79	46.05	46.50	45.65	52.63	43.85	46.50	45.92	54.91	44.62	47.00	46.47	60.51	45.86	46.00	44.73	55.51	43.40
qwen3-8b	33.50	33.47	37.34	29.47	34.00	33.38	40.54	28.06	30.00	28.85	38.16	25.20	34.50	33.46	43.74	31.13	30.00	28.37	37.19	24.87

1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727

## A.5 PROMPT TEMPLATES

### Cell Type Annotation

#### System Prompt:

You are a helpful biology expert. You now need to determine the most probable cell type of a single cell based on its expression profile, which is represented by the top {n\_genes} highly expressed genes sorted in descending order.

#### Task:

- **Goal:** Given the expression profile of a single cell, identify its most probable cell type.
- **Context:**
  1. A list of the top {n\_genes} genes with the highest expression in one single cell. These highly expressed genes serve as a proxy for cell identity and can be used to infer the most likely cell type.
  2. The gene list is obtained from **single-cell RNA sequencing**, with genes sorted in descending order of expression.

**Answer Choices:** {cell\_type\_list}

#### Guidance:

- Use biological knowledge to reason about what cell type the given expression profile most likely represents.
- The provided gene list reflects the single-cell expression profile, which serves as a reference for inferring biological pathways and regulatory interactions.
- Decide independently which perspective (e.g., marker genes, pathways, expression specificity) to use for reasoning.

#### Output Rules:

- Put your reasoning process inside `<analyze></analyze>` tags.
- Put your final decision inside `<answer></answer>` tags.
- The `<answer>` tag must contain exactly one option from the provided list.

#### User Prompt:

From single-cell RNA sequencing, the expression profile of this cell was ranked in descending order, and the top {n\_genes} highly expressed genes were selected:  
{cell\_sentence}

Please tell me what the cell type is.

1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781

### Batch Cell Type Annotation Task

#### System Prompt:

You are an expert assistant specialized in cell type annotation. You will be given a batch of N cells from the same donor, where each cell represents a unique cell type. For each cell, the top expressed genes are provided in descending order of expression. Using both the gene expression data and donor information, determine the correct cell type for each cell. You will also receive a list of N candidate cell types, and each candidate must be assigned to exactly one cell. Ensure that you consider all cells and candidate types together, rather than annotating each cell individually.

Include your detailed reasoning within `<think></think>` tags, and provide your final answer within `<answer></answer>` tags. The final answer should be a single string listing the assigned cell types in order, separated by “—”.

#### User Context:

The cell is from a female at the 73-year-old stage, originating from the lung. The patient has been diagnosed with chronic obstructive pulmonary disease. The patient is a smoker. There is no cancer present.

#### Cells:

- Cell 1: ...
- Cell 2: ...
- Cell 3: ...
- Cell 4: ...
- Cell 5: ...
- Cell 6: ...
- Cell 7: ...
- Cell 8: ...
- Cell 9: ...

#### Candidate Cell Types:

- classical monocyte
- endothelial cell of lymphatic vessel
- epithelial cell of lung
- fibroblast of lung
- macrophage
- malignant cell
- non-classical monocyte
- pulmonary artery endothelial cell
- smooth muscle cell

1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835

## Multi-Omics Cell Type Annotation

### System Prompt:

You are a helpful biology expert.

### Task

- **Goal:** Predict the most probable cell type for a single cell using integrated transcriptomic and proteomic profiles from multi-omics sequencing.
- **Context:** You will be provided with the following information:
  1. A transcriptomic profile containing the top most highly expressed protein-coding genes (non-zero expression), sorted in descending order by expression level.
  2. A proteomic profile listing all non-zero surface proteins detected via CITE-seq (134 possible targets), sorted in descending order by protein expression level.
  3. The complete list of candidate cell types.

### Guidance

Use your biological knowledge to reason about the relationships between transcriptomic and proteomic signals, including marker genes, protein markers, and cell-type-specific pathways.

### Output Rules

- Put your reasoning process inside `<analyze></analyze>` tags.
- Put your final answer inside `<answer></answer>` tags.
- The `<answer>` tag must contain exactly one valid cell type selected from the provided candidate list.

### User Prompt:

Task: Single-Cell Type Annotation Using Multi-Omics Data.

You are given paired transcriptomic and proteomic profiles from a single cell, generated using multi-omics single-cell sequencing (including CITE-seq for surface protein detection).

- The transcriptomic data includes the top 200 most highly expressed protein-coding genes (non-zero expression), sorted in descending order by expression level.
- The proteomic data includes all 134 measurable surface proteins via CITE-seq, with only non-zero detections provided, sorted in descending order by protein expression level.

Possible cell types: {cell\_type\_list}

Transcriptomic profile (top 200 genes): {gene\_top\_names}

Proteomic profile (non-zero surface proteins, ranked by expression): {protein\_top\_names}

Based on this combined multi-omics signature, please predict the most likely cell type.

1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889

## Perturbation QA

### System Prompt:

You are a helpful biology expert. You are given a task of assessing whether CRISPRi knock-down of a perturbation gene leads to a statistically significant change in the expression of another gene.

### Task

- **Goal:** Determine if the queried target gene shows a significant expression change after perturbing the specified gene using CRISPRi.
- **Context:** You will be provided with the following information:
  1. The **background cell line** where the perturbation is applied.
  2. A list of **highly expressed genes** in this cell line, obtained by averaging single-cell expression profiles across all cells (a bulk-like profile representing the baseline transcriptional state). The list is sorted in descending order of mean expression.
  3. The **perturbation condition** (which gene is knocked down by CRISPRi).
  4. The **target gene** whose potential expression change should be assessed.

### Guidance

- Use your biological knowledge to reason about the potential relationship between the perturbation gene and the target gene.
- The provided gene list reflects the average expression profile of the {cell\_line}, serving as a reference for inferring biological pathways and regulatory interactions.
- Knocking down a gene may disrupt its associated pathways, which can in turn affect the expression of the target gene.
- While most genes typically remain unchanged under a given perturbation, in this testing setup the probability has been artificially adjusted—around **50%** of the queries involve target genes that do show significant changes.
- Based on your biological knowledge, independently determine from which perspective to conduct the analysis.

### Output Rules

- Put your reasoning process inside `<analyze></analyze>` tags.
- Put your final decision inside `<answer></answer>` tags.
- The `<answer>` tag must contain one of the following:
  - `<answer>Yes</answer>`: if the target gene shows a statistically significant change.
  - `<answer>No</answer>`: if the target gene does not show a statistically significant change.

### User Prompt:

In {cell\_line} cells, we apply CRISPRi knockdown of {perturbation}.  
Here is the list of highly expressed genes in this background, sorted by mean expression (high → low):  
{cell\_sentence}

Does {gene} show a significant change in expression under this perturbation?

1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943

## Perturbation CLS

### System Prompt:

You are a helpful biology expert. You are given a task of assessing whether CRISPRi knock-down of a perturbation gene leads to a statistically significant change in the expression of another gene.

### Task

- **Goal:** From the provided candidate genes, select the **single most likely gene** that shows a statistically significant expression change after CRISPRi perturbation of the specified gene.
- **Context:** You will be provided with:
  1. The **background cell line** where the perturbation is applied.
  2. A list of **highly expressed genes** in this cell line, obtained by averaging single-cell expression profiles across all cells (bulk-like baseline). The list is sorted in descending order of mean expression.
  3. The **perturbation condition** (which gene is knocked down by CRISPRi).
  4. A **candidate set of gene names**: {candidates}.

### Guidance

- Use biological knowledge to reason about mechanistic links (e.g., pathways, transcription factor–target relationships, co-expression, compensatory mechanisms, or cell-type specificity).
- The background gene list reflects the average expression profile of {cell\_line}, which can guide inference of pathway or regulatory interactions.
- In this testing setup, the candidate set follows a fixed rule:
  - **One gene** is strongly differentially expressed ( $FDR < 0.05$ , Wilcoxon test with BH correction).
  - **Two genes** are weakly or borderline differentially expressed ( $0.05 \leq FDR < 0.1$ ).
  - **Two genes** are not differentially expressed ( $FDR > 0.5$ ).
- Your task is to identify the **single gene** that is most likely the truly significant DEG.
- You must choose **exactly one** gene from {candidates}.

### Output Rules

- Put your reasoning inside `<analyze></analyze>`. Keep it concise but biologically grounded.
- Put your final decision inside `<answer></answer>`.
- The `<answer>` tag must contain **only one exact gene symbol** selected from {candidates}.

### User Prompt:

In {cell\_line} cells, we apply CRISPRi knockdown of {perturbation}.  
Here is the background high-expression gene list (high → low):  
{cell\_sentence}

From the candidates {candidates}, which gene most likely shows a significant expression change?

1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997

## Perturbation Direction Prediction

### System Prompt:

You are a helpful biology expert. You are given a task of determining the direction of expression change caused by a CRISPRi perturbation.

### Task

- **Goal:** Decide whether a queried gene shows an **increase (Increase)** or **decrease (Decrease)** in expression after knockdown of a target gene by CRISPRi.
- **Context:** You will be provided with the following information:
  1. The **background cell line** where the perturbation is applied.
  2. A list of **highly expressed genes** in this cell line, obtained by averaging single-cell expression profiles across all cells (a bulk-like profile representing the base-line transcriptional state). The list is sorted in descending order of mean expression (genes appearing later have lower expression).
  3. The **perturbation condition** (which gene is knocked down by CRISPRi).
  4. The **target gene** (already determined to be differentially expressed) whose direction of change should be assessed.

### Guidance

- Use your biological knowledge to reason about the potential relationship between the perturbation gene and the target gene.
- The provided gene list reflects the average expression profile of the {cell\_line}, serving as a reference for inferring biological pathways and regulatory interactions.
- Knocking down a gene may disrupt its associated pathways, which can in turn affect the expression of the target gene.
- The target gene has already been identified as significantly differentially expressed based on the log2 fold change from the Wilcoxon BH test. Your task is to further determine whether it is significantly increased or decreased.
- Based on your biological knowledge, independently determine from which perspective to conduct the analysis.

### Output Rules

- Put your reasoning process inside `<analyze></analyze>` tags.
- Put your final decision inside `<answer></answer>` tags.
- The `<answer>` tag must contain one of the following:
  - `<answer>Increase</answer>`: if the target gene will be significantly increased after CRISPRi perturbation.
  - `<answer>Decrease</answer>`: if the target gene will be significantly decreased after CRISPRi perturbation.

### User Prompt:

In {cell\_line} cells, we apply CRISPRi knockdown of {perturbation}.

Here is the list of highly expressed genes in this background, sorted by mean expression (high → low):

{cell\_sentence}

Does {gene} go Increase or Decrease in expression under this perturbation?

1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051

### Perturbation CLS with HVGs

#### System Prompt:

You are a helpful biology expert. You are given a task of identifying which of two candidate genes is the true perturbation that caused the observed expression change.

#### Task

- **Goal:** Choose the correct perturbation gene between the two given candidates.
- **Context:** You will be provided with:
  1. The background cell line.
  2. A **control gene list** (200 genes ranked by mean expression).
  3. A **perturbed gene list** (200 genes ranked by mean expression after perturbation).
  4. Two candidate perturbation genes: {choice\_1} and {choice\_2}.

#### Guidance

- Each gene list is constructed by HVG analysis followed by selection of the **top 200 bulk-expressed genes**.
- These lists capture the most informative signatures of control and perturbation states.
- Compare the control and perturbed lists carefully.
- Use biological reasoning: pathway involvement, regulatory modules, upstream regulators, compensatory shifts.
- Only one candidate is correct.

#### Output Rules

- Put your reasoning process inside `<analyze></analyze>` tags.
- Put your final decision inside `<answer></answer>` tags.
- The `<answer>` must be exactly one of the two candidates: `<answer>{choice_1}</answer>` or `<answer>{choice_2}</answer>`.

#### User Prompt:

Background: {cell\_line}.

Control gene list:

{ctrl\_gene\_list}

Perturbed gene list:

{pert\_gene\_list}

Which perturbation gene is correct, {choice\_1} or {choice\_2}?

2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105

## Perturbation QA with HVGs

### System Prompt:

You are a helpful biology expert. You are given a task of determining whether the observed difference between the control and perturbed gene expression profiles is due to CRISPRi knockdown of a specific gene.

### Task

- **Goal:** Judge whether the queried perturbation gene is the true cause of the expression profile change.
- **Context:** You will be provided with:
  1. The background cell line.
  2. A **control gene list** (200 genes ranked by mean expression).
  3. A **perturbed gene list** (200 genes ranked by mean expression after perturbation).
  4. A queried **perturbation gene**.

### Guidance

- Each gene list is constructed by first performing highly variable gene (HVG) analysis, then selecting the **top 200 genes by bulk mean expression**.
- These gene lists represent the most biologically informative subsets of the expression state.
- Use your biological knowledge: consider pathways, regulatory networks, transcription factors, co-expression, compensatory responses, etc.
- Decide whether the observed differences are consistent with knockdown of the queried gene.
- About half of the queries are true perturbations (Yes), half are false (No).

### Output Rules

- Put your reasoning process inside `<analyze></analyze>` tags.
- Put your final decision inside `<answer></answer>` tags.
- The `<answer>` must be either `<answer>Yes</answer>` or `<answer>No</answer>`.

### User Prompt:

Background: {cell\_line}.

Control gene list:

{ctrl\_gene\_list}

Perturbed gene list:

{pert\_gene\_list}

Was this caused by knockdown of {query}?

2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159

## Perturbed Cell Generation

### System Prompt:

You are a helpful biology expert. You are given a task of predicting how the gene expression profile will change under a perturbation.

### Task

- **Goal:** Given the background cell line, the control gene list, and a perturbation gene, predict the perturbed gene list.
- **Context:** You will be provided with:
  1. The background cell line.
  2. A **control gene list** (200 genes ranked by mean expression).
  3. A **perturbation gene** to be knocked down by CRISPRi.

### Guidance

- Each gene list is obtained by HVG analysis, then the **top 200 genes by bulk mean expression** are selected and ranked.
- These gene lists are informative signatures of cell states.
- Use your biological knowledge: pathways, transcriptional regulation, compensatory activation, stress responses, metabolic rewiring, etc.
- Always reason explicitly before generating the perturbed gene list.
- The perturbed gene list should still contain 200 genes.

### Output Rules

- Put your reasoning process inside `<analyze></analyze>` tags.
- Put your final output inside `<answer></answer>` tags.
- The `<answer>` must be a numbered list of exactly 200 entries, one gene per line, like:

```
<answer>
1. GENE_A
2. GENE_B
...
200. GENE_Z
</answer>
```

### User Prompt:

Background: {cell\_line}.

Control gene list:

{ctrl\_gene\_list}

Perturbation: knockdown of {query}.

Predict the perturbed gene list.

2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209  
2210  
2211  
2212  
2213

### Drug Response Prediction Task

#### System Prompt:

You are a helpful biology expert. You are an expert who knows a lot about single cell biology and genomics and will help me solve a series of tasks related to single cell data analysis.

#### Guidance

- If there is provided gene lists, they represent the top  $\{n\_genes\}$  highly expressed genes of each cell, ranked in descending order. These serve as representative expression profiles for the cells and can be used to infer their likely identities.
- Based on your expertise, decide independently which perspectives (e.g., marker genes, tissue architecture, cell-cell interactions) are most informative for this analysis.

#### Output Rules

- Put your reasoning process inside `<analyze></analyze>` tags.
- Put your final decision inside `<answer></answer>` tags.

#### User Prompt:

Below is a list of  $\{gene\_names\}$  ordered by descending expression level in a cell. Your task is to predict its response to the drug Erlotinib: [resistant, sensitive]. Make your choice in format `Final Answer: resistant` or `Final Answer: sensitive`.

$\{gene\_names\}$ : MALSU1 MT-TD MT-CYB MT-CO2 RBM4 MT-ND4L ...

2214  
2215  
2216  
2217  
2218  
2219  
2220  
2221  
2222  
2223  
2224  
2225  
2226  
2227  
2228  
2229  
2230  
2231  
2232  
2233  
2234  
2235  
2236  
2237  
2238  
2239  
2240  
2241  
2242  
2243  
2244  
2245  
2246  
2247  
2248  
2249  
2250  
2251  
2252  
2253  
2254  
2255  
2256  
2257  
2258  
2259  
2260  
2261  
2262  
2263  
2264  
2265  
2266  
2267

## Condition Cell Generation

### System Prompt:

#### Task

- **Goal:** Generate a pseudo cell by predicting the top-expressed genes in a specific cell type and organ context.
- **Context:** You will be provided with the following information:
  1. The organ from which the cell originates: {organ}.
  2. The cell type of interest: {cell\_type}.
  3. The number of top highly expressed genes to generate: 200.
  4. The ground truth expression profile was derived from real single-cell RNA sequencing (scRNA-seq) data, where gene expression levels were averaged across cells of the same type and transformed using log1p. Your task is to simulate this profile by listing genes in descending order of expected expression.

#### Guidance

Use your biological knowledge to reason about the typical transcriptional signature of the specified cell type within the given organ.

#### Output Rules

- You should include your analysis in `<analyze></analyze>`.
- Your final answer should be enclosed in XML tags:

```
<answer>
1. gene_1
2. gene_2
3. gene_3
...
i. gene_i
...
200. gene_200
</answer>
```

- `gene_i` denotes the gene with the *i*-th highest expression in that cell; you should report this gene using the standard HGNC symbol.

#### User Prompt:

Your task is to generate a pseudo cell by outputting the gene expression rank in descending order.

- The cell is from organ: {organ}.
- The cell type is: {cell\_type}.
- Generate the top 200 genes with the highest expression levels, sorted in descending order.
- Output using standard gene symbols.
- Ground truth was extracted from real scRNA-seq data, averaged across cells after log1p transformation.

2268  
2269  
2270  
2271  
2272  
2273  
2274  
2275  
2276  
2277  
2278  
2279  
2280  
2281  
2282  
2283  
2284  
2285  
2286  
2287  
2288  
2289  
2290  
2291  
2292  
2293  
2294  
2295  
2296  
2297  
2298  
2299  
2300  
2301  
2302  
2303  
2304  
2305  
2306  
2307  
2308  
2309  
2310  
2311  
2312  
2313  
2314  
2315  
2316  
2317  
2318  
2319  
2320  
2321

## Gene Annotation

### System Prompt:

You are a specialized expert in Gene Ontology (GO) annotation and human molecular biology, with proficiency in retrieving and structuring accurate biological data.

### Task

- **Goal:** Perform a comprehensive Gene Ontology (GO) annotation for a given human gene.
- **Context:** You will be provided with a single human gene Symbol (e.g., TP53).

### Guidance

- Your annotation must be biologically accurate and consistent with up-to-date GO databases.
- The response must include the following information for the target gene:
  1. **Official Gene Name:** The full, official name.
  2. **Gene Type:** The biotype of the gene (e.g., `protein_coding`, `pseudogene`).
  3. **Summary:** A concise description of the gene's primary function and biological role.
  4. **GO (Gene Ontology):** A detailed listing of the three subcategories:
    - **Molecular Function**
    - **Biological Process**
    - **Cellular Component**
- For each GO term listed, you must provide both its name and its official definition.
- If no reliable annotations are available for a specific GO category, explicitly state: `no reliable annotation is available`.

### Output Rules

Place your annotation information inside `<answer></answer>` tags. No markdown formatting.

### User Prompt:

Provide a full GO annotation for the human gene `{gene}`.

2322  
2323  
2324  
2325  
2326  
2327  
2328  
2329  
2330  
2331  
2332  
2333  
2334  
2335  
2336  
2337  
2338  
2339  
2340  
2341  
2342  
2343  
2344  
2345  
2346  
2347  
2348  
2349  
2350  
2351  
2352  
2353  
2354  
2355  
2356  
2357  
2358  
2359  
2360  
2361  
2362  
2363  
2364  
2365  
2366  
2367  
2368  
2369  
2370  
2371  
2372  
2373  
2374  
2375

### Spatial Cell Type Annotation

#### System Prompt:

You are a helpful biology expert. You are tasked with performing a **spatially-informed cell annotation task**. Based on the expression profiles (represented by lists of highly expressed genes) of a **center cell** and its **{N} neighboring cells**, you must infer the most probable type of the center cell.

#### Task:

- **Goal:** Given the expression profile of a center cell and its **{N}** spatial neighbors, identify the most probable cell type of the center.
- **Context:** You will be provided with the following information:
  1. One **center cell** with its top **{n\_genes}** expressed genes.
  2. **{N} neighboring cells**, each represented by their own top expressed genes.
  3. All gene lists are obtained from **single-cell RNA sequencing**.
  4. The expression profiles will always follow the order: center cell followed by nearest neighbors.
  5. The neighboring cells are the closest cells to the center cell in spatial proximity.

You must select the answer from the following list:  
{cell\_type\_list}

#### Guidance:

- Use your biological knowledge to reason about the potential relationships between the **center cell** and its **spatial neighbors**.
- Since neighboring cells are the closest in spatial distance, they may share microenvironmental or tissue-level context that should inform your prediction.
- The provided gene lists represent the top **{n\_genes}** highly expressed genes of each cell, ranked in descending order. These serve as representative expression profiles for the cells and can be used to infer their likely identities.
- You should make full use of the spatial information while reasoning.
- Based on your expertise, decide independently which perspectives (e.g., marker genes, tissue architecture, cell-cell interactions) are most informative for this analysis.

#### Output Rules:

- Put your reasoning process inside `<analyze></analyze>` tags.
- Put your final decision inside `<answer></answer>` tags.
- The `<answer>` tag must contain one of the following formats: `<answer>{{cell_type}}</answer>`, where `{{cell_type}}` must exactly match one of the options in the provided list.

#### User Prompt:

Center cell (genes high  $\rightarrow$  low):  
{gene\_0}

Neighbors:

- First neighbor: {gene\_1}
- Second neighbor: {gene\_2}
- ...
- **{N}**-th neighbor: {gene\_{N}}

What is the type of the center cell?