scACORN: Context-engineered agent orchestration of specialized small language models for single-cell transcriptomic interpretation

Arash Rasti-Meymandi[1,5], Sepideh Nahali[2], Eustache Paramithiotis[4], Angela M. Cheung[1], Elham Dolatabadi[3,5*]

[1*]Department of Medicine, University of Toronto, Toronto, Ontario, Canada.
[2]School of Health Policy and Management, Faculty of Health, York University, Toronto, Ontario, Canada.
[3]York University, Toronto, Ontario, Canada.
[4]CellCarta, Montreal, Quebec, Canada.
[5]Vector Institute, Toronto, Ontario, Canada.

*Corresponding author(s). E-mail(s): edolatab@yorku.ca;

Abstract

Single-cell atlases now exceed 66 million cells, but turning a ranked expression profile and a free-form biological question into a reliable, evidence-grounded answer remains unsolved. Scaling a single model does not resolve this, because single-cell interpretation is a heterogeneous family of tasks whose correct answer depends on tissue, cohort, perturbation and annotation resolution. Here we present scACORN, an agentic alternative to monolithic single-cell language models that combines specialized small language models with context-engineered agent orchestration for their selection and composition at inference time. Each expert is built in two stages: domain-aligned contrastive adaptation fits a pretrained cell-to-text backbone to the transcriptomic geometry of a target dataset, and geometry-preserving specialization learns question-conditioned biological completions without eroding that geometry. A fixed orchestrating language model agent then selects and combines experts under a natural-language playbook that is itself optimized from textual feedback, with no gradient updates to the orchestrator. Across 10 Tabula Sapiens tissues, domain alignment raised transfer macro-F1 from 0.36 to 0.64 and Recall@5 from 0.87 to 0.97; specialized experts reached 0.89 mean exact-match annotation accuracy; and playbook optimization reduced

unsupported gene citations from 14.5% to 3.5%. Our findings support specialization and orchestration as complementary responses to the heterogeneity and evidentiary demands of single-cell analysis.

Keywords: single-cell transcriptomics, language models, agentic systems, small language models, contrastive learning

1 Introduction

Single-cell atlases have made cellular heterogeneity measurable at scale (Regev et al. 2017; Rood et al. 2025; Tabula Sapiens Consortium 2022). Between 2020 and 2025 the Human Cell Atlas grew more than elevenfold, from 5.8 million to over 66 million cells and beyond 300 TB (Human Cell Atlas Data Portal 2020; Human Cell Atlas 2025). These resources have enabled detailed characterization of the cellular identities, states and responses that shape tumor heterogeneity, immunotherapy response and human development (Tirosh et al. 2016; Sade-Feldman et al. 2018; Cao et al. 2020). Yet atlas-scale analysis requires more than assigning a label to an individual expression profile. It requires resolving closely related lineages, interpreting transcriptional programs in the context of tissue, disease and perturbation, comparing cells across experimental conditions, and linking biological conclusions to the genes that support them. As single-cell datasets increase in both scale and biological diversity (Regev et al. 2017; Tabula Sapiens Consortium 2022), the principal computational challenge is shifting from data generation and representation toward reproducible, context-sensitive and evidence-grounded interpretation (Svensson et al. 2018; Lähnemann et al. 2020).

Foundation models offer a potential route to scale this interpretation. Geneformer (Theodoris et al. 2023), scGPT (Cui et al. 2024) and scFoundation (Hao et al. 2024) learn general-purpose transcriptomic representations, whereas Cell2Sentence and Cell-o1 connect single-cell profiles to natural-language generation and reasoning (Levine et al. 2024; Fang et al. 2025). However, increasing model scale does not, by itself, resolve the interpretation problem. Zero-shot transfer remains weak, perturbation prediction does not beat linear baselines, and task-specific models frequently match or exceed far larger ones (Kedzierska et al. 2025; Ahlmann-Eltze et al. 2025). Single-cell interpretation comprises a heterogeneous set of tasks, and a single biological question may require several forms of expertise. The significance of an expression program may vary with tissue, cohort, modality, perturbation and annotation resolution (Luecken et al. 2022; Domínguez Conde et al. 2022; Abdelaal et al. 2019; Hrovatin et al. 2025). Monolithic adaptation therefore conflates two requirements: acquiring precise local expertise and determining which expertise applies to each profile and question.

Here we formulate single-cell interpretation as the *agentic composition of specialized* biological expertise. Given one or more transcriptomic profiles represented by ranked gene lists and a free-form biological question, the system must identify the relevant biological subproblems, assign each profile to the appropriate specialists, and synthesize a coherent answer in which molecular claims remain traceable to the corresponding measurements. We introduce scACORN (**s**ingle-**c**ell **A**gentic **C**ontext **OR**chestration

Network), an agentic specialist–orchestrator designed to meet these requirements (Fig. 1). scACORN constructs reusable cell-to-text experts from target-domain data and selects and composes them at the level of individual profiles. In contrast to conventional tool-using agents, which operate over a predefined collection of fixed tools (Yao et al. 2023; Schick et al. 2023), scACORN builds a repertoire of specialized biological experts and dynamically orchestrates them according to the biological context and question at hand.

scACORN separates the acquisition of local biological expertise from the policy used to select and integrate that expertise. Experts are built from a pretrained single-cell small language model (scSLM) backbone in two stages. Domain-aligned contrastive adaptation fits a pretrained cell-to-text backbone to the transcriptomic geometry of a target dataset, establishing competence within a defined domain. Geometry-preserving specialization then converts that competence into question-conditioned predictions and gene-level rationales while explicitly penalizing drift from the aligned representation. The resulting tissue-, dataset-, modality- or task-specific models form the set of experts available to the orchestrator. Composition is performed by one of two fixed language-model orchestrators, GPT-5.4 mini or Claude Sonnet 4.5, guided by an explicit natural-language playbook that is optimized from structured textual feedback rather than gradients (Zhang et al. 2025; OpenAI 2026; Anthropic 2025). Because experts are constructed independently and share a textual interface, the framework provides a modular basis for incorporating additional specialists.

In total we trained 20 adapters spanning 10 Tabula Sapiens tissues (Tabula Sapiens Consortium 2022), generated and scored approximately 170,000 held-out cell annotations against 12 baseline systems, and evaluated composition on 80 held-out free-form questions under two orchestrating language models. Domain alignment increased mean cell-type classification macro-F1 from 0.356 to 0.643 and Recall@5 from 0.868 to 0.968, bringing a language-model representation into the range of purpose-built ranked-gene encodings. Geometry-preserving specialization produced tissue experts with 0.90 pooled exact cell-type accuracy and an evidence-gene F1 of 0.82, while retaining hidden-space Recall@10 at 0.969 at the reported precision. On held-out free-form questions, including 42 that required two or three experts, playbook optimization reduced unsupported support-gene citations from 14.5% to 3.5% and increased ontology-aware agreement from 0.754 to 0.843. scACORN achieved the highest ontology agreement, evidence alignment and nuanced-answer quality among the evaluated general-purpose and transcriptomic models.

2 Results

We evaluated scACORN (Fig. 1) across the three linked stages of its workflow; *First*, we asked whether contrastive domain alignment could adapt a pretrained cell-to-text backbone to the transcriptomic structure of a target tissue. *Second*, we tested whether the aligned backbone could be converted into a cell-level reasoning expert that produces accurate labels and grounded marker rationales. *Third*, we evaluated whether textual orchestration could answer free-form questions by selecting and combining the appropriate experts from a library of specialized models.

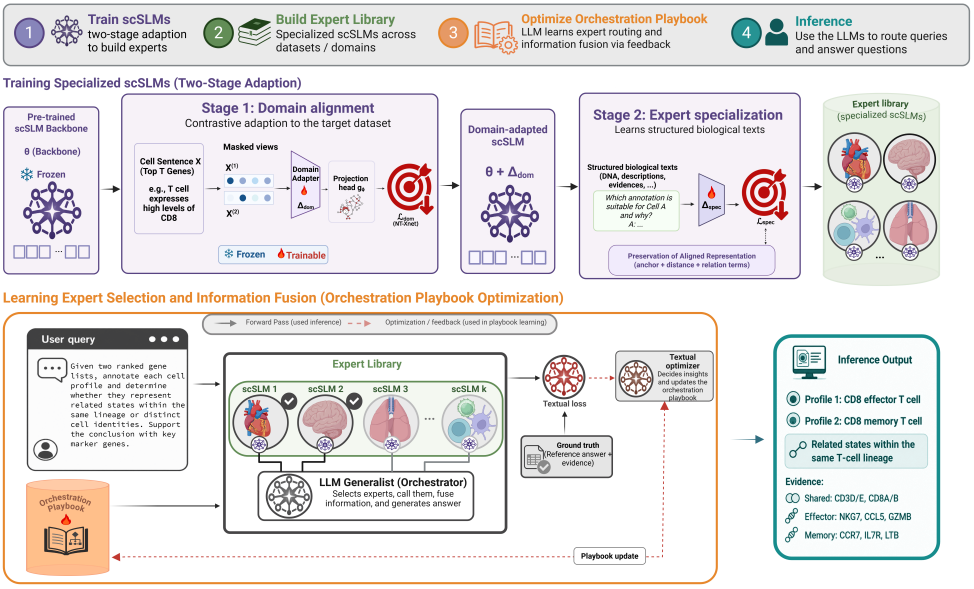

Fig. 1. Overview of scACORN (single-cell Agentic Context ORchestration Network). 1, Domain-aligned contrastive adaptation fits a pretrained cell-to-text backbone to a target transcriptomic domain, and geometry-preserving specialization converts it into a question-conditioned expert. **2**, The trained experts form the collection available to the orchestrator. **3**, The orchestration playbook is optimized from textual feedback while the expert and orchestrator parameters remain fixed. **4**, At inference, the orchestrator selects one or more experts and synthesizes their outputs into a free-form answer with profile-linked molecular evidence. Solid arrows indicate inference flow, and dashed arrows indicate feedback used for playbook optimization.

2.1 Experimental Design

All experiments used Tabula Sapiens as a common biological substrate (Tabula Sapiens Consortium 2022). The domain-alignment benchmark included cells from 10 tissues. To reduce leakage from experimentally linked observations, cells belonging to the same acquisition group, including cells from the same 10x run, were assigned to the same training, validation or test partition. The expert-specialization benchmark evaluated tissue-specific biological completions on held-out cells from the same 10-tissue collection. External model comparisons used matched held-out profiles within each tissue so that all methods received the same ranked-gene inputs.

The orchestration benchmark comprised 80 fixed held-out questions: 38 single-expert questions and 42 compositional questions. Single-expert questions required one tissue-specialized expert, whereas compositional questions required the selection and combination of two or three experts. The reference expert set associated with each question was used only by the evaluator for routing diagnostics. All compared systems received the same profiles, questions and evaluation rubric.

To test whether orchestration performance depended on a particular model scale or provider, we evaluated `GPT-5.4 mini` as a compact OpenAI orchestrator and `Claude Sonnet 4.5` as a larger-capacity Anthropic orchestrator (OpenAI 2026; Anthropic

2025). The set of experts, held-out questions and evaluation procedure were fixed across the two configurations.

For domain alignment, we compared Rank PCA, TF-IDF followed by truncated SVD, the unadapted pretrained cell-to-text backbone, the hidden representations produced after domain alignment, and the corresponding learned projection space. For expert specialization, we compared scACORN experts with open-source instruction-tuned models from the Mistral, Qwen, Gemma and Llama families (Jiang et al. 2023; Qwen Team 2025; Yang et al. 2025; Gemma Team 2024; Dubey et al. 2024), closed-source models (OpenAI 2024, 2025), and specialized transcriptomic language models (Fang et al. 2025; Rizvi et al. 2025). Primary metrics were aligned with the claim tested by each component: transfer macro-F1 and Recall@5 for domain alignment; exact label agreement and evidence grounding for expert specialization; and ontology-aware agreement, evidence alignment, marker coverage and nuanced answer quality for orchestration. Paired model comparisons used sample-level randomization tests against the strongest non-scACORN baseline. Formal metric definitions are provided in Appendix C.

2.2 Domain Alignment Fits A Language-Model Backbone to A Target Single-Cell Domain

Because expert specialization operates on the adapted LLM state rather than on TF-IDF, PCA or another fixed feature space, we evaluated domain alignment as an upstream component of expert construction. Specifically, we asked whether contrastive adaptation (Fig. 2) improved the local structure of the unfitted cell-to-text representation and brought it into the range of conventional ranked-gene representations before question-conditioned specialization.

Across the 10 tissues, the unfitted scSLM backbone obtained a mean classification macro-F1 of 0.356 ± 0.088. Domain alignment increased macro-F1 to 0.643 ± 0.125 for the adapted hidden features and 0.640 ± 0.171 for the learned projection space. TF-IDF + SVD reached 0.716 ± 0.085, and Rank PCA reached 0.651 ± 0.137 (Fig. 2(a1,a3)). Thus, contrastive adaptation substantially improved the pretrained language-model representation and produced classification structure comparable to the conventional representations, while retaining a state that could be reused by the downstream SLM expert.

The retrieval analysis showed a similar pattern (Fig. 2(a2)). Adapted hidden features achieved the highest mean Recall@5, 0.968 ± 0.022, compared with 0.868 ± 0.078 for the unfitted backbone, 0.963 ± 0.045 for TF-IDF + SVD, 0.962 ± 0.053 for Rank PCA and 0.951 ± 0.052 for the learned projection space. These results do not establish domain alignment as a new general-purpose embedding method; rather, they show that it fits the cell-to-text backbone to the target biological domain while retaining competitive local separation.

Representative prostate and stomach embeddings provided a qualitative view of the same effect (Fig. 2(b1–b8)). The unfitted scSLM representation showed weaker cell-type organization, whereas the adapted hidden features formed more coherent local neighborhoods.

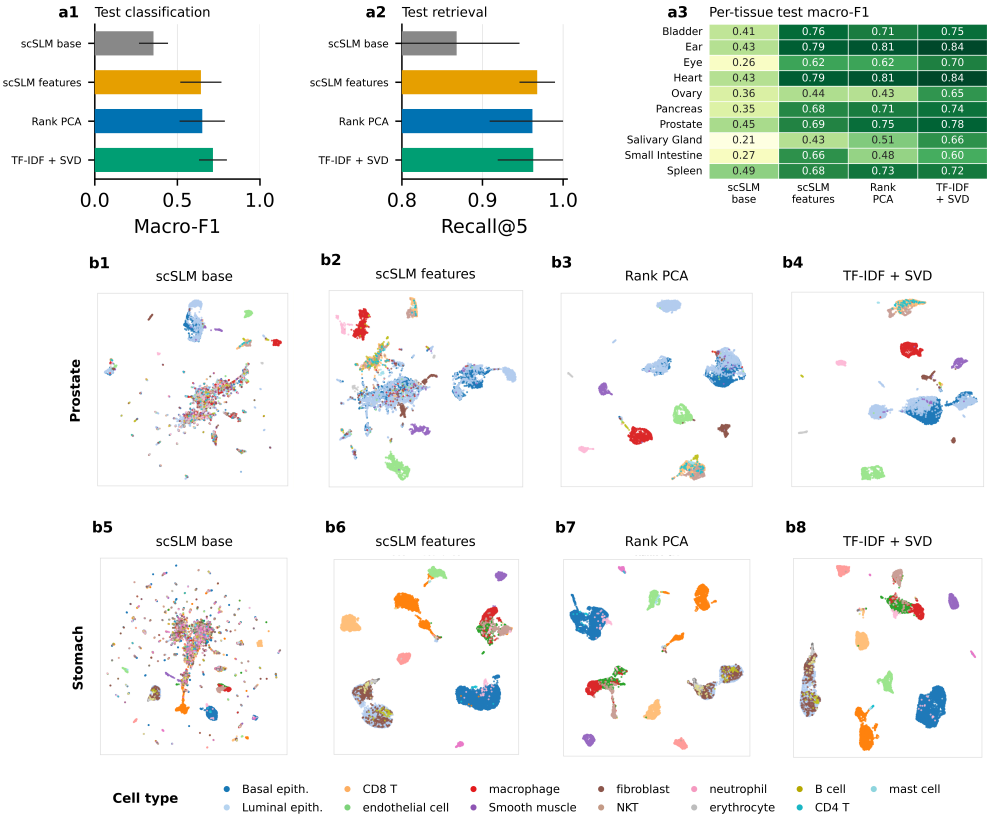

Fig. 2. Domain alignment improves cell-type structure in cell-to-text representations. **a1**, Test classification macro-F1 for the unadapted single-cell small language model (scSLM) backbone, domain-aligned scSLM hidden features, Rank PCA and TF-IDF + SVD representations. **a2**, Test retrieval Recall@5 for the same four representations. **a3**, Per-tissue test macro-F1 across the 10 evaluated tissues. Bars in **a1** and **a2** show the mean across tissues; error bars indicate the s.d. across tissues. **b1**–**b4**, UMAP projections (McInnes et al. 2018) of prostate cells represented by the unadapted scSLM backbone, domain-aligned hidden features, Rank PCA and TF-IDF + SVD, respectively. **b5**–**b8**, Corresponding projections for stomach cells. Each point represents one cell, and colors denote reference cell-type annotations.

2.3 Expert Specialization Improves cell-type Reasoning and Molecular Grounding

We next asked whether the domain-aligned backbone could be converted into a question-conditioned biological expert. Across the 10-tissue expert-specialization benchmark, scACORN experts achieved a mean exact-match rate of 0.893 ± 0.096. Performance varied across tissues, ranging from 0.990 in bladder to 0.719 in small intestine and 0.760 in salivary gland. Detailed tissue-level results are reported in Appendix A.

scACORN also outperformed substantially larger general-purpose and transcriptomics-oriented models on label quality. Across the full evaluation, scACORN reached 0.95 canonical accuracy, 0.90 exact accuracy, 0.95 ontology credit,

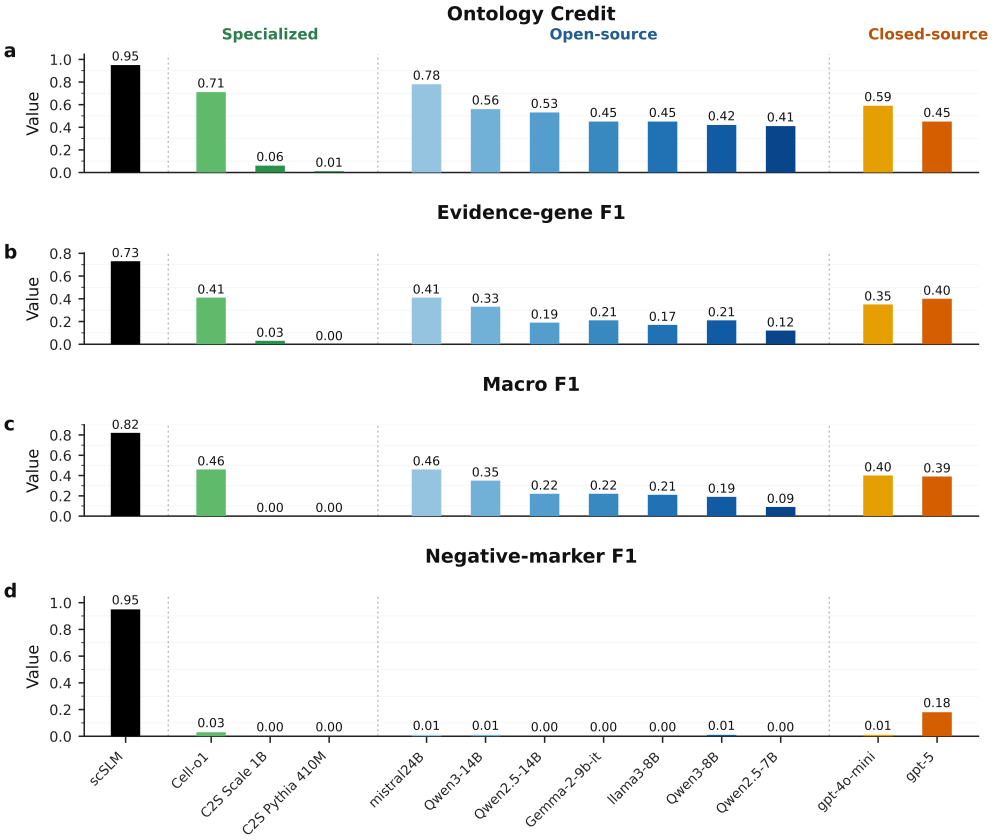

Fig. 3. Expert specialization improves cell-type reasoning and molecular grounding. a–d, Performance of the scACORN's experts (scSLMs) compared with specialized transcriptomic models, open-source instruction-tuned language models and closed-source language models on matched held-out cell profiles. **a**, Ontology credit, which awards partial credit to predictions that are biologically related to the reference cell type. **b**, Evidence-gene F1, measuring agreement between genes cited in the generated rationale and the reference supporting-gene set. **c**, Macro-F1 across cell-type classes. **d**, Negative-marker F1, measuring agreement between predicted and reference genes used to exclude competing cell identities. Higher values are better; values are shown above the bars.

0.73 macro-F1 and 0.96 weighted F1. The ontology-credit and macro-F1 comparisons are summarized in Fig. 3(a,b). The strongest open-source baseline, `Mistral-Small-3.2-24B-Instruct-2506`, reached 0.76 canonical accuracy and 0.41 macro-F1. The strongest specialized baseline, Cell-o1, reached 0.70 canonical accuracy and 0.41 macro-F1, whereas the best closed-source result reached 0.54 canonical accuracy and 0.59 ontology credit. The larger difference in macro-F1 is consistent with improved performance beyond the dominant cell classes.

The advantage of expert specialization extended to the molecular evidence in the generated answers (Fig. 3 (c,d)). scACORN obtained evidence-gene precision, recall and F1 of 0.82, 0.82 and 0.82, respectively. The strongest open-source baseline reached an evidence-gene F1 of 0.46, the strongest closed-source baseline reached 0.40, and

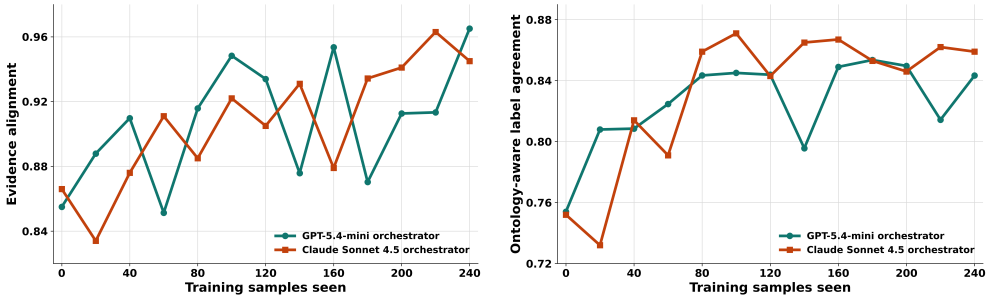

Fig. 4. Textual playbook optimization improves evidence grounding and ontology-aware agreement under two orchestrators. Left, input-grounded support-gene score at successive playbook-optimization checkpoints. The unsupported support-gene citation rate is one minus this score. Right, ontology-aware label agreement at the same checkpoints. Curves show results for the GPT-5.4-mini and Claude Sonnet 4.5 orchestrators. Dashed horizontal lines indicate performance with the corresponding initial playbook. Each checkpoint was evaluated on the same fixed held-out set; the final endpoint was specified in advance, and intermediate checkpoints were not used to select the reported playbook.

Cell-o1 reached 0.46. scACORN also obtained negative-marker precision, recall and F1 of 0.96, 0.94 and 0.95, respectively.

Qualitative examples further illustrated the distinction between label prediction and grounded reasoning. scACORN identified bladder urothelial cells using epithelial programs including *KRT19*, *CLDN4* and *TACSTD2*; fibroblasts using *CFD*, *DCN* and *MGP*; and heart endothelial cells using *PECAM1*, *VWF* and *ENG*. In a multi-profile question, the model separated an ileal enterocyte profile supported by *APOA1*, *APOA4*, *SLC5A1* and *FABP2* from a pericyte profile supported by *RGS5*, *NOTCH3* and *CALD1*. Full qualitative interactions are provided in Appendix D.

2.4 Textual Orchestration Improves Free-Form Biological Question Answering

Finally, we evaluated whether a language-model orchestrator guided by an optimized textual playbook could select and combine specialized experts to answer free-form biological questions. The held-out benchmark included both questions requiring one expert and compositional questions requiring two or three expert calls. The latter setting requires the system to determine which domain-specific knowledge is relevant, associate each profile with the appropriate expert and synthesize evidence across multiple expert outputs.

As indicated in Fig. 4, playbook optimization improved the orchestrator's use of expert evidence and the compatibility of its predictions with the reference labels (Fig. 4). For GPT-5.4 mini, the input-grounded support-gene score increased from 0.855 with the initial playbook to 0.965 at the prespecified final endpoint, reducing the unsupported support-gene citation rate from 0.145 to 0.035. Ontology-aware agreement increased from 0.754 to 0.843. Claude Sonnet 4.5 showed a similar overall improvement. Although performance varied across intermediate versions of the playbook, the final version outperformed the initial version on both measures.

Table 1. Textual orchestration performance across orchestrators, question types and playbook conditions. a, Single-expert questions ($n = 38$), each requiring one tissue-specialized expert. **b**, Compositional questions ($n = 42$), each requiring two or three experts. For scACORN, the model in parentheses denotes the orchestrating language model. Values are mean scores across held-out questions; higher is better. The best result for each metric within each panel is shown in bold. Empty indicates that no playbook instructions were supplied; optimized denotes the final optimized playbook. Evidence align., weighted set-F1 agreement between predicted and reference supporting and negative-marker sets; Marker cov., weighted recall of reference supporting and negative markers mentioned in the generated answer.

Category	Model	Ontology ↑	Evidence align. ↑	Marker cov. ↑	Nuanced ↑
a. Single-expert questions					
Open-source	Mistral-Small-3.2-24B-Instruct	0.593	0.311	0.506	0.435
	Qwen3-14B	0.460	0.200	0.342	0.371
	Llama-3.1-8B-Instruct	0.216	0.135	0.296	0.273
Closed-source	GPT-4o-mini	0.480	0.308	0.388	0.400
	GPT-5	0.633	0.290	0.865	0.486
Specialized	Cell-o1	0.411	0.259	0.556	0.476
	scACORN (GPT-5.4-mini; empty)	0.780	0.537	0.902	0.699
	scACORN (GPT-5.4-mini; optimized)	0.832	0.632	0.904	0.747
	scACORN (Claude Sonnet 4.5; empty)	0.751	0.543	0.914	0.709
	scACORN (Claude Sonnet 4.5; optimized)	**0.841**	**0.678**	**0.922**	**0.753**
b. Compositional questions					
Open-source	Mistral-Small-3.2-24B-Instruct	0.491	0.301	0.461	0.377
	Qwen3-14B	0.569	0.193	0.375	0.381
	Llama-3.1-8B-Instruct	0.338	0.113	0.209	0.280
Closed-source	GPT-4o-mini	0.543	0.257	0.334	0.384
	GPT-5	0.773	0.364	0.807	0.520
Specialized	Cell-o1	0.499	0.305	0.544	0.482
	scACORN (GPT-5.4-mini; empty)	0.614	0.019	0.793	0.513
	scACORN (GPT-5.4-mini; optimized)	0.853	0.508	0.702	0.741
	scACORN (Claude Sonnet 4.5; empty)	0.622	0.139	**0.808**	0.553
	scACORN (Claude Sonnet 4.5; optimized)	**0.858**	**0.535**	0.716	**0.744**

Across all 80 held-out questions, Claude-orchestrated scACORN achieved an ontology score of 0.850, an evidence-alignment score of 0.603, a marker-coverage score of 0.814 and a nuanced-answer score of 0.748. The corresponding GPT-5.4-mini results were 0.843, 0.567, 0.798 and 0.744. Claude Sonnet 4.5 therefore produced the strongest ontology, evidence-alignment and nuanced-answer results among the evaluated systems. GPT-5, the strongest standalone closed-source baseline, obtained 0.707, 0.328, 0.834 and 0.504 and retained the highest aggregate marker coverage. Cell-o1 obtained 0.457, 0.283, 0.550 and 0.480.

Even with an empty playbook, the initial prompt sometimes led the orchestrator to select an appropriate expert. Playbook optimization made this selection more reliable, which helps explain the large improvement in evidence alignment for compositional questions: once the appropriate experts were selected, their outputs were already strongly aligned with the reference evidence. The other metrics changed more moderately, and marker coverage was less consistent.

For single-expert questions ($n = 38$), the optimized Claude-orchestrated scACORN achieved the highest score on all four metrics. For compositional questions ($n = 42$), it achieved the highest ontology, evidence-alignment and nuanced-answer scores, whereas the empty-playbook Claude condition had the highest marker coverage (Table 1).

With optimized playbooks, Claude Sonnet 4.5 outperformed GPT-5.4 mini on all four metrics in both subsets.

3 Discussion

We introduce scACORN, which constructs specialized scSLMs and uses a context-engineered agent to orchestrate their selection and composition for single-cell transcriptomic interpretation. Each expert is trained within a defined biological domain, while the orchestrator selects and combines those relevant to the profiles and question at hand. The shared textual interface preserves the provenance of molecular evidence by linking each contribution to its source profile and expert. The resulting design supports broader biological questions through the composition of focused expertise. This modularity also allows the system to accommodate new experts without altering its overall structure. For example, a new scSLM can be constructed by applying domain alignment and expert specialization to a new tissue, dataset or task, then added to the available expert set together with a textual description of its scope and input requirements. Because the playbook specifies how and when experts should be invoked, the orchestrator can consider the new specialist at inference, although routing reliability as the expert set expands remains to be evaluated.

Single-cell interpretation requires local adaptation without erasing the transcriptomic structure that distinguishes related cell populations (Hrovatin et al. 2025; Li and Hoiem 2016). Our findings support this rationale. Domain alignment organized the language-model backbone around tissue-specific transcriptomic structure, providing a biologically appropriate starting point for expert specialization. The preservation objective then allowed experts to learn question-conditioned outputs without measurable loss of this structure under the present evaluation. This was particularly important for distinguishing closely related populations, whose identities depend on both expressed markers and the absence of competing lineage programs (Abdelaal et al. 2019; Domínguez Conde et al. 2022). Accordingly, the compact experts improved not only label quality but also the recovery of supporting and exclusionary evidence relative to substantially larger models. At the system level, playbook optimization primarily improved how this evidence was used, substantially reducing unsupported gene citations across both single- and multi-expert questions.

This study provides an initial validation of scACORN in a controlled, multi-tissue setting and identifies several priorities for further evaluation. Deriving all experts from Tabula Sapiens enabled matched comparisons across tissues and system components; validation in independent atlases, disease and perturbation cohorts, and data generated using different protocols will now be important for establishing robustness to dataset shift. Similarly, our benchmarks isolate cell identification and evidence synthesis, providing a foundation for prospective studies of whether the framework can support biological discovery and generate hypotheses that withstand expert review and experimental validation. The current set of experts also offers a tractable testbed for orchestration, whereas larger collections will introduce challenges related to overlapping expertise, conflicting outputs and inference cost. Our trace-based routing controls indicate non-random expert selection, but complete generative reruns

under oracle, random and disabled routing will be needed to determine its causal contribution to answer quality. Finally, gene-level grounding provides a transparent measure of whether cited evidence is present in the input or reference set, but should not be interpreted as evidence of biological specificity, mechanism or causality.

Future work should evaluate broader sets of experts across independent datasets, disease and perturbation cohorts, rare-cell populations and multimodal measurements such as CITE-seq, spatial transcriptomics and chromatin accessibility. As the number of experts grows, orchestration will need to account explicitly for expert overlap, disagreement, uncertainty and computational cost. This reframes biological model scaling as a problem of organizing and coordinating expertise rather than continually enlarging a single model.

4 Methods

4.1 Problem Formulation and Framework Overview

We consider grounded cell-level biological question answering. The index i denotes an example or cell. Each example contains a ranked single-cell transcriptomic profile x_i, a natural-language question q_i, optional task context c_i, and, during supervised training, a target textual answer y_i. Given one or more profiles and a question, the objective is to generate a biologically compatible answer whose supporting evidence is traceable to the corresponding observed profile. The system must also identify the relevant domain-specific expert or combination of experts and avoid citing genes that are unsupported by the associated ranked input. The target answer may contain a cell-type label, supporting genes, negative markers, tissue context, alternative interpretations, or a concise biological rationale.

scACORN separates representation adaptation, expert construction, and runtime reasoning into three sequential components. Domain alignment first adapts a pretrained cell-to-text backbone to the transcriptomic structure of a target dataset or tissue. Geometry-preserving expert specialization then learns question-conditioned biological completions without erasing the aligned representation. Finally, textual expert orchestration registers the resulting specialists as tools and selects one or more of them through an explicitly optimized playbook.

Unless stated otherwise, r indexes gene rank, m indexes target tokens, v indexes augmented views, k indexes experts, p indexes input profiles, j indexes tool calls, and t indexes textual-optimization iterations. The same expert-construction procedure is applied independently to each domain-specific expert; we omit the expert identifier during domain alignment and specialization for clarity.

4.2 Cell-to-Text Representation

Each cell is represented by its top T expressed genes,

$$x_i = (g_{i,1}, g_{i,2}, \ldots, g_{i,T}),$$

where $g_{i,r}$ is the gene at rank r for cell i, and genes are ordered by decreasing expression. We use $T = 200$ in all experiments. A deterministic prompt template $\mathcal{T}$ converts the ranked profile, question, and optional context into a textual prompt:

$$s_i = \mathcal{T}(x_i, q_i, c_i).$$

During domain alignment, no question or task context is supplied, so the corresponding prompt is $\mathcal{T}(x_i, \varnothing, \varnothing)$.

Let θ denote the frozen parameters of the pretrained causal language-model backbone, and let Δ denote an active adapter or ordered collection of adapters. For a prompt s, we write $\mathbf{h}_{\theta,\Delta}(s) \in \mathbb{R}^d$ for the hidden state of its final non-padding token, where d is the backbone hidden dimension. The full sequence-level definition is provided in Appendix B.1. This formulation follows the cell-to-text paradigm and supports parameter-efficient domain alignment and expert specialization (Levine et al. 2024; Rizvi et al. 2025).

4.3 Domain-Aligned Contrastive Adaptation

Domain alignment adjusts the pretrained backbone to the local transcriptomic structure of a target domain before supervised reasoning is introduced. Let $\mathcal{A}$ denote the stochastic augmentation operator applied to ranked-gene profiles. For each cell x_i, we draw two independent views,

$$\widetilde{x}_i^{(1)}, \widetilde{x}_i^{(2)} \sim \mathcal{A}(x_i),$$

where $\mathcal{A}$ applies biologically conservative perturbations, including gene dropout, local rank swaps, and random truncation or subsampling.

The augmented profiles are converted into cell-to-text prompts and encoded using a trainable domain-alignment LoRA adapter Δ_{dom} (Hu et al. 2022). A projection head $g_\phi : \mathbb{R}^d \to \mathbb{R}^{d_z}$, parameterized by ϕ, maps each prompt representation to a d_z-dimensional contrastive space. For any nonzero vector $\mathbf{u}$, define $\text{norm}(\mathbf{u}) = \mathbf{u}/\|\mathbf{u}\|_2$, where $\|\cdot\|_2$ is the Euclidean norm. The normalized embedding for view $v \in \{1, 2\}$ is

$$\mathbf{z}_i^{(v)} = \text{norm}\left(g_\phi\left(\mathbf{h}_{\theta,\Delta_{\text{dom}}}\left(\mathcal{T}(\widetilde{x}_i^{(v)}, \varnothing, \varnothing)\right)\right)\right).$$

For a mini-batch of N cells, we optimize the NT-Xent objective (Chen et al. 2020):

$$\mathcal{L}_{\text{dom}} = \text{NTXent}\left(\left\{\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}\right\}_{i=1}^{N}; \tau\right),$$

where N is the number of cells in the contrastive mini-batch and $\tau > 0$ is the temperature. The two views of the same cell form a positive pair, whereas all other views in the batch act as negatives. The expanded objective is given in Appendix B.2.

Only Δ_{dom} and ϕ are optimized. In our implementation, LoRA modules are applied to the attention projections in the upper half of the transformer layers. The projection head is used only during domain-alignment training and representation evaluation.

Checkpoints are selected using validation Recall@10 in the projection space under cosine similarity. This criterion reflects the purpose of domain alignment: adapting the language-model representation to the target domain rather than directly optimizing cell-type prediction.

4.4 Geometry-Preserving Expert Specialization

Expert specialization converts the domain-aligned backbone into a question-conditioned biological reasoning expert. The domain-alignment adapter Δ_{dom} remains active and frozen, while a second adapter Δ_{spec} is trained for the target completion tasks.

For training example i, the target answer is tokenized as

$$y_i = (y_{i,1}, \ldots, y_{i,M_i}),$$

where M_i is the number of target tokens and $y_{i,<m}$ denotes the prefix preceding token m. For a specialization mini-batch $\mathcal{B}_{\text{spec}}$, we optimize the completion-only causal language-modeling loss

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{\sum_{i \in \mathcal{B}_{\text{spec}}} M_i} \sum_{i \in \mathcal{B}_{\text{spec}}} \sum_{m=1}^{M_i} \log P_{\theta,\Delta_{\text{dom}},\Delta_{\text{spec}}}\left(y_{i,m} \mid s_i, y_{i,<m}\right).$$

Here, $P_{\theta,\Delta_{\text{dom}},\Delta_{\text{spec}}}$ denotes the model next-token distribution. Prompt tokens are excluded from the loss. Training targets contain the final cell label and, when available, supporting genes, negative markers, tissue information, confidence, and a short rationale.

4.5 Representation-Preservation Objective

Expert specialization can improve completion quality while distorting the domain geometry learned through domain alignment, reflecting the broader problem of representation drift during sequential adaptation (Li and Hoiem 2016). We therefore compare the normalized prompt representation of the current specialized expert with that of the frozen domain-aligned reference model. The three preservation terms draw on pointwise feature alignment, moment matching, and relational knowledge preservation (Sun and Saenko 2016; Park et al. 2019).

We preserve the domain-aligned representation at three complementary levels:

1. **pointwise alignment**, which preserves each prompt representation;
2. **distribution alignment**, which preserves batch means and componentwise variances; and
3. **relational alignment**, which preserves pairwise similarities among prompts in the same mini-batch.

The complete specialization objective is

$$\mathcal{L}_{\text{spec}} = \mathcal{L}_{\text{SFT}} + \lambda_{\text{anchor}}\mathcal{L}_{\text{anchor}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}},$$

where $\lambda_{\text{anchor}}, \lambda_{\text{dist}}, \lambda_{\text{rel}} \geq 0$ weight the pointwise, distributional, and relational preservation losses, respectively. Their exact definitions are provided in Appendix B.3.

Only Δ_{spec} is updated during expert specialization. Checkpoints are selected using validation completion loss, while hidden-space retrieval on a held-out domain split is monitored as an auxiliary measure of representation retention.

4.6 Textual Expert Orchestration

After expert specialization, each specialized model is made available to the orchestrator as a textual tool, following the broader tool-augmented agent paradigm in which an LLM selects and invokes external capabilities during reasoning (Yao et al. 2023; Schick et al. 2023).

Let

$$\mathcal{R} = \{\rho_1, \ldots, \rho_K\}$$

denote the set of K experts, where descriptor ρ_k specifies expert k's dataset or tissue, supported tasks, adapter location, and routing description. A query may contain P ranked-gene profiles,

$$\mathcal{X} = \{x^{(1)}, \ldots, x^{(P)}\}.$$

Given a free-form question q, profile set $\mathcal{X}$, optional context c, expert set $\mathcal{R}$, and textual playbook $\mathcal{P}$, the selected fixed orchestrating LLM defines a playbook-conditioned textual policy Π:

$$(\mathcal{C}, a) = \Pi\left(q, \mathcal{X}, c, \mathcal{R}; \mathcal{P}\right).$$

Here,

$$\mathcal{C} = [(k_1, p_1), \ldots, (k_J, p_J)]$$

is the ordered tool trace containing J expert calls, $k_j \in \{1, \ldots, K\}$ is the expert selected at call j, $p_j \in \{1, \ldots, P\}$ is the profile supplied to that expert, and a is the synthesized final answer.

We instantiated this policy separately with GPT-5.4 mini and Claude Sonnet 4.5. In both configurations, the orchestrator parameters remained fixed and only the textual playbook was updated. Both configurations used the same experts and tool interface.

The policy receives descriptions of all available experts but does not receive the identity of the correct expert. It must therefore determine which expert or combination of experts is appropriate for the query. This component is the mechanism that enables accurate responses to free-form questions: instead of relying on a single general model for all biological contexts, the orchestrator maps each question and ranked-gene profile to the most suitable specialized expert or expert combination in the available set.

Expert execution.

The shared backbone is loaded once, and expert-specific domain-alignment and specialization adapters are activated on demand. Expert outputs are converted to a common textual schema containing the predicted answer, supporting genes, negative markers, tissue context, confidence, and execution metadata. The common schema enables heterogeneous experts to be composed without aligning their hidden representations.

4.7 Textual Optimization of the Playbook

The playbook $\mathcal{P}$ contains explicit natural-language rules for expert routing, tool construction, evidence synthesis, ambiguity handling, and abstention. We optimize this textual context directly while keeping the parameters of the orchestrating LLM fixed. This design follows recent work that treats prompts or textual context as optimizable objects using natural-language feedback rather than parameter gradients (Pryzant et al. 2023; Yang et al. 2024; Yuksekgonul et al. 2025; Zhang et al. 2025).

At optimization iteration t, example i is processed with the current playbook $\mathcal{P}_t$:

$$(\mathcal{C}_{i,t}, a_{i,t}) = \Pi\left(q_i, \mathcal{X}_i, c_i, \mathcal{R}; \mathcal{P}_t\right).$$

A textual-loss module then compares the output with the available training supervision:

$$\left(\ell_{i,t}^{\text{text}}, \delta_{i,t}\right) = \mathcal{L}_{\text{text}}\left(a_{i,t}, \mathcal{C}_{i,t}; y_i, \mathcal{E}_i^{\star}, \mathcal{X}_i\right).$$

Here, $\ell_{i,t}^{\text{text}} \in \mathbb{R}_{\geq 0}$ is a scalar error summary, $\delta_{i,t}$ is a structured natural-language diagnosis, and $\mathcal{E}_i^{\star} \subseteq \{1, \ldots, K\}$ is the set of oracle experts used only by the training-time evaluator.

For an optimization mini-batch $\mathcal{B}_t^{\text{ACE}}$, the diagnoses are distilled into recurrent, actionable insights:

$$\mathbf{g}_t^{\text{text}} = \text{Distill}\left(\left\{\delta_{i,t} : i \in \mathcal{B}_t^{\text{ACE}}\right\}\right).$$

The operator Distill summarizes recurring diagnoses, and $\mathbf{g}_t^{\text{text}}$ is the resulting textual gradient. It is not a numerical derivative; it is a natural-language surrogate that identifies how the current playbook should change to reduce routing, grounding, synthesis, or abstention errors.

A textual optimizer updates the playbook:

$$\mathcal{P}_{t+1} = \text{Opt}_{\text{text}}\left(\mathcal{P}_t, \mathbf{g}_t^{\text{text}}\right),$$

where Opt_{text} is the rule-editing operator. It can add, rewrite, merge, remove, or reorder rules while eliminating redundant and contradictory instructions. Detailed formalization of the textual loss, edit space, and discrete optimization interpretation is given in Appendix B.4.

Oracle experts, reference answers, textual losses, and optimizer feedback are unavailable at inference time. The final policy operates only on the user request, input profiles, expert descriptions, expert outputs, and the optimized playbook.

4.8 Extensibility

The framework is designed for expert-level extensibility: separately constructed specialists communicate through a shared textual interface. This modularity follows the principle of reusable and composable adapters (Houlsby et al. 2019; Pfeiffer et al. 2020, 2021). The present study evaluated a fixed set of experts; expansion to previously unseen experts remains future work.

Table 2. Stage 2 cell-type annotation performance across ten Tabula Sapiens tissues. Exact match is the primary task metric. Evidence-grounding reports the fraction of ground-truth evidence genes recovered by the generated rationale.

Tissue	N_{test}	Exact match	Evidence-grounding	Mean latency (s)
Bladder	1532	0.990	0.910	11.61
Ear	1693	0.960	0.910	10.67
Eye	894	0.802	0.693	11.17
Heart	656	0.956	0.899	10.67
Ovary	1768	0.950	0.895	11.49
Pancreas	1449	0.937	0.757	11.50
Prostate	916	0.912	0.759	11.47
Salivary Gland	657	0.760	0.483	12.05
Small Intestine	637	0.719	0.920	11.47
Spleen	2859	0.942	0.923	11.56

Table 3. Representation-preservation ablation during expert specialization. Validation loss measures task fitting (lower is better), whereas hidden-space Recall@10 measures retention of the domain-aligned neighborhood structure (higher is better). The frozen domain-aligned representation had a pre-specialization Recall@10 of 0.969.

Expert-specialization variant	Validation loss	Hidden-space Recall@10
Without preservation regularization	1.185	0.915
With preservation regularization	**1.143**	**0.969**

A Additional Expert-Specialization Results

Table 2 reports the tissue-level exact-match and grounding results for the expert-specialization benchmark. The table is placed in the appendix to keep the main Results focused on aggregate findings and cross-model comparisons.

A.1 Representation-Preservation Ablation

We compared expert specialization with and without the representation-preservation objective. The frozen domain-aligned representation achieved a held-out hidden-space Recall@10 of 0.969 before specialization. Without preservation regularization, Recall@10 decreased to 0.915. The regularized expert retained Recall@10 at 0.969 and achieved a lower validation loss than the unregularized model (1.143 versus 1.185; Table 3).

B Technical Details of the Method

B.1 Backbone and Prompt Representation

Let s be a tokenized prompt, let $L(s)$ be its number of non-padding tokens, and let d be the hidden dimension of the causal language-model backbone. Given frozen backbone

parameters θ and active adapter parameters Δ, the backbone produces

$$\mathbf{H}_{\theta,\Delta}(s) \in \mathbb{R}^{L(s)\times d},$$

where row r contains the hidden state of prompt token r. We use the final non-padding hidden state as the prompt representation:

$$\mathbf{h}_{\theta,\Delta}(s) = \mathbf{H}_{\theta,\Delta}(s)_{L(s),:} \in \mathbb{R}^d.$$

The notation

$$\mathbf{h}_{\theta,\Delta_{\text{dom}},\Delta_{\text{spec}}}(s)$$

denotes the same backbone with the domain-alignment adapter Δ_{dom} followed by the specialization adapter Δ_{spec}. The backbone parameters θ remain frozen throughout both adaptation components.

B.2 Expanded Domain-Alignment Objective

For cell i and view $v \in \{1, 2\}$, define

$$s_i^{(v)} = \mathcal{T}\left(\widetilde{x}_i^{(v)}, \varnothing, \varnothing\right).$$

Let $g_\phi : \mathbb{R}^d \to \mathbb{R}^{d_z}$ be the projection head, where d_z is the contrastive embedding dimension. The normalized projected representation is

$$\mathbf{z}_i^{(v)} = \frac{g_\phi\left(\mathbf{h}_{\theta,\Delta_{\text{dom}}}(s_i^{(v)})\right)}{\left\| g_\phi\left(\mathbf{h}_{\theta,\Delta_{\text{dom}}}(s_i^{(v)})\right)\right\|_2}.$$

For a contrastive mini-batch of N cells, define the set of $2N$ augmented-view indices

$$\mathcal{I}_{\text{view}} = \{(i, v) : i \in \{1, \ldots, N\},\ v \in \{1, 2\}\}.$$

For view v, let $\bar{v} = 3 - v$ denote the other view of the same cell. Because all projected vectors are normalized, cosine similarity between vectors $\mathbf{u}$ and $\mathbf{v}$ is

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v}.$$

The domain-alignment loss is

$$\mathcal{L}_{\text{dom}} = -\frac{1}{2N} \sum_{(i,v)\in\mathcal{I}_{\text{view}}} \log \frac{\exp\left(\text{sim}\left(\mathbf{z}_i^{(v)}, \mathbf{z}_i^{(\bar{v})}\right)/\tau\right)}{\sum\limits_{\substack{(j,u)\in\mathcal{I}_{\text{view}} \\ (j,u)\neq(i,v)}} \exp\left(\text{sim}\left(\mathbf{z}_i^{(v)}, \mathbf{z}_j^{(u)}\right)/\tau\right)},$$

where $\tau > 0$ is the contrastive temperature. The other view of the same cell is the positive sample, and every remaining view in the mini-batch is a negative sample.

B.3 Expanded Expert-Specialization Objective

Let $\mathcal{B}_{\text{spec}}$ be a specialization mini-batch and let $B = |\mathcal{B}_{\text{spec}}|$ be its number of examples. For example i, the target completion is

$$y_i = (y_{i,1}, \ldots, y_{i,M_i}),$$

where M_i is its token length. The token-normalized completion loss is

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{\sum_{i \in \mathcal{B}_{\text{spec}}} M_i} \sum_{i \in \mathcal{B}_{\text{spec}}} \sum_{m=1}^{M_i} \log P_{\theta, \Delta_{\text{dom}}, \Delta_{\text{spec}}} \left(y_{i,m} \mid s_i, y_{i,<m} \right).$$

For each prompt s_i, define the normalized current representation

$$\mathbf{e}_i^{\text{cur}} = \frac{\mathbf{h}_{\theta, \Delta_{\text{dom}}, \Delta_{\text{spec}}}(s_i)}{\left\| \mathbf{h}_{\theta, \Delta_{\text{dom}}, \Delta_{\text{spec}}}(s_i) \right\|_2}$$

and the frozen domain-aligned reference

$$\mathbf{e}_i^{\text{ref}} = \text{sg} \left[\frac{\mathbf{h}_{\theta, \Delta_{\text{dom}}}(s_i)}{\| \mathbf{h}_{\theta, \Delta_{\text{dom}}}(s_i) \|_2} \right],$$

where $\text{sg}[\cdot]$ denotes stop-gradient.

B.3.1 Pointwise Alignment

The pointwise anchor loss is

$$\mathcal{L}_{\text{anchor}} = 1 - \frac{1}{B} \sum_{i \in \mathcal{B}_{\text{spec}}} \left(\mathbf{e}_i^{\text{cur}} \right)^\top \mathbf{e}_i^{\text{ref}}.$$

B.3.2 Distribution Alignment

For state $\sigma \in \{\text{cur}, \text{ref}\}$, define the mini-batch mean

$$\boldsymbol{\mu}^\sigma = \frac{1}{B} \sum_{i \in \mathcal{B}_{\text{spec}}} \mathbf{e}_i^\sigma$$

and componentwise variance

$$\mathbf{v}^\sigma = \frac{1}{B} \sum_{i \in \mathcal{B}_{\text{spec}}} \left(\mathbf{e}_i^\sigma - \boldsymbol{\mu}^\sigma \right)^{\odot 2},$$

where $(\cdot)^{\odot 2}$ denotes elementwise squaring. The distribution-alignment loss is

$$\mathcal{L}_{\text{dist}} = \left\| \boldsymbol{\mu}^{\text{cur}} - \boldsymbol{\mu}^{\text{ref}} \right\|_2^2 + \left\| \mathbf{v}^{\text{cur}} - \mathbf{v}^{\text{ref}} \right\|_2^2.$$

B.3.3 Relational Alignment

Choose any fixed ordering $i_1, \dots, i_B$ of the examples in $\mathcal{B}_{\text{spec}}$. For $\sigma \in \{\text{cur}, \text{ref}\}$, stack their normalized representations row-wise:

$$\mathbf{E}^{\sigma} = \begin{bmatrix} (\mathbf{e}^{\sigma}_{i_1})^{\top} \\ \vdots \\ (\mathbf{e}^{\sigma}_{i_B})^{\top} \end{bmatrix} \in \mathbb{R}^{B \times d}.$$

The pairwise cosine-similarity matrix is

$$\mathbf{S}^{\sigma} = \mathbf{E}^{\sigma}(\mathbf{E}^{\sigma})^{\top}.$$

Let $\mathbf{1}_B \in \mathbb{R}^B$ be the all-ones vector, $\mathbf{I}_B \in \mathbb{R}^{B \times B}$ the identity matrix, and

$$\mathbf{W} = \mathbf{1}_B \mathbf{1}_B^{\top} - \mathbf{I}_B$$

the off-diagonal mask. The relational loss is

$$\mathcal{L}_{\text{rel}} = \frac{1}{B(B-1)} \left\| \mathbf{W} \odot \left(\mathbf{S}^{\text{cur}} - \mathbf{S}^{\text{ref}} \right) \right\|_F^2,$$

where $\odot$ denotes elementwise multiplication and $\| \cdot \|_F$ is the Frobenius norm. The diagonal is excluded because both similarity matrices have unit self-similarity.

The full specialization objective is

$$\mathcal{L}_{\text{spec}} = \mathcal{L}_{\text{SFT}} + \lambda_{\text{anchor}} \mathcal{L}_{\text{anchor}} + \lambda_{\text{dist}} \mathcal{L}_{\text{dist}} + \lambda_{\text{rel}} \mathcal{L}_{\text{rel}}.$$

B.4 Detailed Textual Optimization Formulation

Let

$$\mathcal{D}_{\text{ACE}} = \{(q_i, \mathcal{X}_i, c_i, y_i, \mathcal{E}_i^{\star})\}_{i=1}^{n_{\text{ACE}}}$$

denote the orchestration optimization set, where n_{ACE} is the number of optimization examples and $\mathcal{E}_i^{\star} \subseteq \{1, \dots, K\}$ is the oracle expert set available only to the training-time evaluator.

At iteration $t \in \{0, \dots, T_{\text{opt}} - 1\}$, the textual policy produces

$$(\mathcal{C}_{i,t}, a_{i,t}) = \Pi\left(q_i, \mathcal{X}_i, c_i, \mathcal{R}; \mathcal{P}_t\right).$$

The textual-loss module returns

$$\left(\ell_{i,t}^{\text{text}}, \delta_{i,t}\right) = \mathcal{L}_{\text{text}}\left(a_{i,t}, \mathcal{C}_{i,t}; y_i, \mathcal{E}_i^{\star}, \mathcal{X}_i\right),$$

where $\ell_{i,t}^{\text{text}} \in \mathbb{R}_{\geq 0}$ is the scalar loss and $\delta_{i,t}$ is its structured textual diagnosis.

For optimization mini-batch $\mathcal{B}_t^{\text{ACE}} \subseteq \{1, \dots, n_{\text{ACE}}\}$, the diagnoses are distilled as

$$\mathbf{g}_t^{\text{text}} = \text{Distill}\left(\left\{\delta_{i,t} : i \in \mathcal{B}_t^{\text{ACE}}\right\}\right).$$

Although termed a textual gradient, $\mathbf{g}_t^{\text{text}}$ is not obtained by differentiation. It is a natural-language surrogate that summarizes a direction of improvement in the discrete playbook space.

The playbook is updated by

$$\mathcal{P}_{t+1} = \text{Opt}_{\text{text}}\left(\mathcal{P}_t, \mathbf{g}_t^{\text{text}}\right).$$

The allowed edit actions form the set

$$\mathcal{U}_{\text{edit}} = \{\textsc{Add}, \textsc{Rewrite}, \textsc{Merge}, \textsc{Delete}, \textsc{Reorder}\}.$$

These actions alter routing rules, profile-specific constraints, evidence-synthesis instructions, and abstention criteria.

B.4.1 Discrete Optimization Interpretation

Let

$$\mathcal{Q}_{\text{rule}} = \{u_1, \ldots, u_{M_{\text{rule}}}\}$$

be an abstract collection of M_{rule} candidate textual rules, and let

$$\mathbf{b} = (b_1, \ldots, b_{M_{\text{rule}}}) \in \{0,1\}^{M_{\text{rule}}}$$

indicate which candidate rules are active. The playbook induced by $\mathbf{b}$ is

$$\mathcal{P}(\mathbf{b}) = \{u_m : b_m = 1\}.$$

Under this abstraction, playbook selection can be written as

$$\mathbf{b}^\star = \arg\min_{\mathbf{b}\in\{0,1\}^{M_{\text{rule}}}} \mathcal{J}\left(\mathcal{P}(\mathbf{b})\right),$$

where the empirical evaluation objective is

$$\mathcal{J}(\mathcal{P}) = \frac{1}{n_{\text{ACE}}} \sum_{i=1}^{n_{\text{ACE}}} \ell_{\text{eval}}\left(\Pi\left(q_i, \mathcal{X}_i, c_i, \mathcal{R}; \mathcal{P}\right), y_i, \mathcal{E}_i^\star\right).$$

Here, ℓ_{eval} is the scalar evaluator used to compare a policy output with its reference answer and oracle expert set.

The implemented optimizer operates over a richer space than binary rule selection because rules can also be rewritten, merged, or specialized. Let $\mathcal{N}(\mathcal{P}_t)$ denote the textual neighborhood induced by one or more actions in $\mathcal{U}_{\text{edit}}$, and let $\widehat{\mathcal{J}}$ denote the textual optimizer's local surrogate of $\mathcal{J}$. Each update can be viewed as

$$\mathcal{P}_{t+1} \approx \arg\min_{\mathcal{P}\in\mathcal{N}(\mathcal{P}_t)} \widehat{\mathcal{J}}\left(\mathcal{P} \mid \mathbf{g}_t^{\text{text}}\right).$$

After T_{opt} optimization iterations, the final playbook is

$$\mathcal{P}^\star = \mathcal{P}_{T_{\text{opt}}}.$$

For an unseen query, inference is

$$(\mathcal{C}, a) = \Pi\left(q, \mathcal{X}, c, \mathcal{R}; \mathcal{P}^\star\right).$$

No reference answer, oracle expert identity, textual loss, or optimizer feedback is used during inference.

C Metric Definitions

C.1 Domain-Alignment Metrics

Let $\mathcal{I}_{\text{test}}^{\text{dom}}$ be the index set of held-out cells in the domain-alignment benchmark, and let $\mathcal{Y}_{\text{dom}}$ be its set of evaluated cell-type labels. Transfer predictions are obtained using a k_{clf}-nearest-neighbor classifier, where k_{clf} denotes the fixed number of neighbors used by the classifier.

For class $c \in \mathcal{Y}_{\text{dom}}$, let TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives, respectively. Precision, recall, and F1 are

$$\text{Prec}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \qquad \text{Rec}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c},$$

and

$$\text{F1}_c = \frac{2\,\text{Prec}_c\,\text{Rec}_c}{\text{Prec}_c + \text{Rec}_c}.$$

Macro-F1 is

$$\text{Macro-F1} = \frac{1}{|\mathcal{Y}_{\text{dom}}|} \sum_{c \in \mathcal{Y}_{\text{dom}}} \text{F1}_c.$$

For Recall@5, let $\mathcal{N}_5(i)$ be the set of five nearest training cells to held-out cell i, and let ℓ_r^{dom} denote the reference cell-type label of any training or test cell indexed by r. We define

$$\text{Recall@5} = \frac{1}{|\mathcal{I}_{\text{test}}^{\text{dom}}|} \sum_{i \in \mathcal{I}_{\text{test}}^{\text{dom}}} \mathbf{1}\left[\exists j \in \mathcal{N}_5(i) \text{ such that } \ell_j^{\text{dom}} = \ell_i^{\text{dom}}\right],$$

where $\mathbf{1}[\cdot]$ is the indicator function.

C.2 Expert-Specialization Metrics

Let $\mathcal{I}_{\text{test}}^{\text{spec}}$ be the index set of held-out specialization examples. Let $\hat{\ell}_i^{\text{spec}}$ and ℓ_i^{spec} denote the generated and reference final labels for example i. Exact-match accuracy is

$$\text{Exact-match} = \frac{1}{|\mathcal{I}_{\text{test}}^{\text{spec}}|} \sum_{i \in \mathcal{I}_{\text{test}}^{\text{spec}}} \mathbf{1}\left[\hat{\ell}_i^{\text{spec}} = \ell_i^{\text{spec}}\right].$$

Let $\hat{\mathcal{E}}_i$ be the normalized, deduplicated set of genes returned in the generated evidence field, $\mathcal{G}_i$ the set of genes in the ranked input profile, and $\mathcal{E}_i$ the reference evidence set. The per-example evidence-in-input score is

$$a_i^{\text{input}} = \begin{cases} \dfrac{|\hat{\mathcal{E}}_i \cap \mathcal{G}_i|}{|\hat{\mathcal{E}}_i|}, & |\hat{\mathcal{E}}_i| > 0, \\ 0, & |\hat{\mathcal{E}}_i| = 0. \end{cases}$$

The dataset-level evidence-in-input rate is

$$\text{Evidence-in-input} = \frac{1}{|\mathcal{I}_{\text{test}}^{\text{spec}}|} \sum_{i \in \mathcal{I}_{\text{test}}^{\text{spec}}} a_i^{\text{input}}.$$

The per-example evidence-grounding score is

$$a_i^{\text{reference}} = \begin{cases} \dfrac{|\hat{\mathcal{E}}_i \cap \mathcal{E}_i|}{|\mathcal{E}_i|}, & |\mathcal{E}_i| > 0, \\ 0, & |\mathcal{E}_i| = 0. \end{cases}$$

The dataset-level evidence-grounding rate is

$$\text{Evidence-grounding} = \frac{1}{|\mathcal{I}_{\text{test}}^{\text{spec}}|} \sum_{i \in \mathcal{I}_{\text{test}}^{\text{spec}}} a_i^{\text{reference}}.$$

C.3 Orchestration Metrics

Let $\mathcal{I}_{\text{test}}^{\text{orch}}$ be the index set of held-out orchestration examples. The primary orchestration metrics quantify label compatibility, evidence agreement, marker coverage and the combined quality of the generated answer. Routing diagnostics are reported separately to isolate expert selection from answer synthesis.

C.3.1 Ontology-Aware Label Agreement

Let $\hat{\ell}_i^{\text{orch}}$ and ℓ_i^{orch} denote the predicted and reference cell-type labels for orchestration example i. Let $\nu(\cdot)$ be the fixed label-normalization mapping used by the evaluator. Define

$$s_{\text{ont}}\left(\hat{\ell}_i^{\text{orch}}, \ell_i^{\text{orch}}\right) = \begin{cases} 1, & \nu(\hat{\ell}_i^{\text{orch}}) = \nu(\ell_i^{\text{orch}}), \\ \alpha_{\text{ont}}, & \left(\nu(\hat{\ell}_i^{\text{orch}}), \nu(\ell_i^{\text{orch}})\right) \in \mathcal{C}_{\text{ont}}, \\ 0, & \text{otherwise}, \end{cases}$$

where $\mathcal{C}_{\text{ont}}$ is the fixed set of ontology-compatible label pairs and $0 < \alpha_{\text{ont}} < 1$ is the fixed partial-credit value. Hierarchical cell-type ontologies provide standardized relations for such comparisons (Diehl et al. 2016).

The dataset-level metric is

$$\text{OntologyAgree} = \frac{1}{|\mathcal{I}_{\text{test}}^{\text{orch}}|} \sum_{i \in \mathcal{I}_{\text{test}}^{\text{orch}}} s_{\text{ont}} \left(\hat{\ell}_i^{\text{orch}}, \ell_i^{\text{orch}} \right).$$

C.3.2 Input-Grounded Support-Gene Score

Let $\hat{\mathcal{S}}_i$ be the normalized, deduplicated set of support genes returned for orchestration example i, and let $\mathcal{G}_i^{\text{orch}}$ be the set of genes in the corresponding ranked input profile. The per-example score is

$$a_i^{\text{ground}} = \begin{cases} \dfrac{|\hat{\mathcal{S}}_i \cap \mathcal{G}_i^{\text{orch}}|}{|\hat{\mathcal{S}}_i|}, & |\hat{\mathcal{S}}_i| > 0, \\ 0, & |\hat{\mathcal{S}}_i| = 0. \end{cases}$$

Assigning zero to an empty support set prevents a model from obtaining a perfect grounding score by omitting evidence entirely.

The dataset-level score is

$$\text{InputGroundedSupport} = \frac{1}{|\mathcal{I}_{\text{test}}^{\text{orch}}|} \sum_{i \in \mathcal{I}_{\text{test}}^{\text{orch}}} a_i^{\text{ground}}.$$

The complementary input-unsupported citation rate is

$$\text{UnsupportedCitationRate} = 1 - \text{InputGroundedSupport}.$$

This metric evaluates whether cited genes are licensed by the observed input profile. It does not measure whether an input-present gene is biologically specific for the predicted label; biological label compatibility is evaluated separately through OntologyAgree.

C.3.3 Evidence Alignment

Evidence alignment measures whether the structured evidence returned by the model matches the reference support and exclusionary-marker fields. Let $\hat{\mathcal{S}}_i$ and $\mathcal{S}_i$ denote the predicted and reference support-gene sets for orchestration example i, and let $\hat{\mathcal{N}}_i$ and $\mathcal{N}_i$ denote the corresponding predicted and reference negative-marker sets. For any two sets A and B, define

$$\text{F1}_{\text{set}}(A, B) = \begin{cases} 1, & |A| = 0 \text{ and } |B| = 0, \\ 0, & |A| = 0 \text{ xor } |B| = 0, \\ \dfrac{2\,|A \cap B|}{|A| + |B|}, & \text{otherwise.} \end{cases}$$

The support-gene and negative-marker alignment scores are

$$a_i^{\text{sup}} = \text{F1}_{\text{set}}\left(\hat{\mathcal{S}}_i, \mathcal{S}_i\right), \qquad a_i^{\text{neg}} = \text{F1}_{\text{set}}\left(\hat{\mathcal{N}}_i, \mathcal{N}_i\right).$$

The per-example evidence-alignment score is

$$a_i^{\text{evid}} = 0.9\, a_i^{\text{sup}} + 0.1\, a_i^{\text{neg}},$$

The 0.9/0.1 weighting prioritizes supporting genes because they provide direct positive evidence for the assigned identity, while retaining negative markers as secondary exclusionary evidence. The dataset-level metric is

$$\text{EvidenceAlign} = \frac{1}{|\mathcal{I}_{\text{test}}^{\text{orch}}|} \sum_{i \in \mathcal{I}_{\text{test}}^{\text{orch}}} a_i^{\text{evid}}.$$

C.3.4 Marker Mention Coverage

Marker mention coverage measures whether the free-form rationale mentions the reference support and negative markers. Let r_i be the generated free-form answer. For a set of markers M, define

$$\text{MentionRecall}(r_i, M) = \begin{cases} 1, & |M| = 0, \\ \dfrac{1}{|M|} \sum_{m \in M} \mathbf{1}\left[m \text{ appears in } r_i\right], & |M| > 0. \end{cases}$$

The per-example marker-coverage score is

$$a_i^{\text{marker}} = 0.9\, \text{MentionRecall}(r_i, \mathcal{S}_i) + 0.1\, \text{MentionRecall}(r_i, \mathcal{N}_i),$$

and the dataset-level score is

$$\text{MarkerCoverage} = \frac{1}{|\mathcal{I}_{\text{test}}^{\text{orch}}|} \sum_{i \in \mathcal{I}_{\text{test}}^{\text{orch}}} a_i^{\text{marker}}.$$

C.3.5 Expert Routing Quality

Let $\mathcal{C}_i$ be the set of experts actually called for example i, and let $\mathcal{E}_i^\star$ be the oracle expert set stored in the request metadata. If no oracle set is available, the evaluator uses the candidate expert set $\mathcal{E}_i^{\text{cand}}$. Define the target expert set

$$\mathcal{T}_i = \begin{cases} \mathcal{E}_i^\star, & |\mathcal{E}_i^\star| > 0, \\ \mathcal{E}_i^{\text{cand}}, & \text{otherwise.} \end{cases}$$

The routing recall and precision are

$$\text{Recall}_i^{\text{expert}} = \frac{|\mathcal{C}_i \cap \mathcal{T}_i|}{|\mathcal{T}_i|}, \qquad \text{Precision}_i^{\text{expert}} = \frac{|\mathcal{C}_i \cap \mathcal{T}_i|}{|\mathcal{C}_i|},$$

with the empty-set cases handled by assigning perfect routing only when no expert is required and no expert is called. Let $h_i^{\text{expert}} = 1$ if at least one target expert is called and zero otherwise. The routing-quality score is

$$a_i^{\text{route}} = 0.5\, h_i^{\text{expert}} + 0.3\, \text{Recall}_i^{\text{expert}} + 0.2\, \text{Precision}_i^{\text{expert}}.$$

The reported routing quality is the mean over the held-out orchestration set,

$$\text{RoutingQuality} = \frac{1}{|\mathcal{I}_{\text{test}}^{\text{orch}}|} \sum_{i \in \mathcal{I}_{\text{test}}^{\text{orch}}} a_i^{\text{route}}.$$

For routing-only controls, the same score is recomputed after replacing the observed expert-call trace with oracle, random or empty expert-call traces. These controls isolate expert selection and do not regenerate the final answer.

C.3.6 Nuanced Overall Score

The nuanced overall score is the weighted evaluator composite used to summarize the quality of each orchestration answer. Let a_i^{abstain} be the abstention-calibration score, a_i^{explain} be the token-F1 similarity between the generated rationale and reference rationale, and a_i^{ground} be the input-grounded support-gene score defined above. The per-example nuanced score is

$$\begin{aligned} a_i^{\text{nuanced}} = {} & 0.20\, s_{\text{ont}}\left(\hat{\ell}_i^{\text{orch}}, \ell_i^{\text{orch}}\right) + 0.15\, a_i^{\text{route}} + 0.15\, a_i^{\text{abstain}} \\ & + 0.15\, a_i^{\text{evid}} + 0.15\, a_i^{\text{explain}} + 0.10\, a_i^{\text{marker}} + 0.10\, a_i^{\text{ground}}. \end{aligned}$$

The reported nuanced overall score is the mean of a_i^{nuanced} over $\mathcal{I}_{\text{test}}^{\text{orch}}$.

C.3.7 Statistical Reporting

Pairwise model comparisons use paired randomization tests on the per-example metric differences. The reported tests compare scACORN with GPT-5 because GPT-5 is the strongest non-scACORN baseline in the main orchestration comparison.

C.3.8 Checkpoint Reporting

Pre-optimization values were computed using the initial playbook $\mathcal{P}_0$, and the final endpoint was specified before evaluation. Intermediate checkpoints were used to describe the optimization trajectory and were not used to select the reported playbook on the held-out benchmark.

D Qualitative scACORN Interactions

We selected the following held-out Orchestration examples to illustrate the question-answering behavior supported by scACORN. To keep the appendix readable, we omit the complete ranked-gene lists and report only the user question, scACORN response, and reference answer.

D.1 High-Agreement Examples

D.1.1 Single-Cell Endothelial Query

User question

I pasted a ranked gene list above and I am not sure what cell type this profile represents. Could you suggest the most likely cell identity and point to which markers in the list support that assignment? I am trying to decide between an endothelial versus a mesenchymal or smooth-muscle interpretation.

BioAgent

Most likely identity: cardiac endothelial cell (predicted tissue: heart). This assignment is supported by endothelial markers present in the ranked list, including VWF, ENG, and EMCN. Mesenchymal and smooth-muscle alternatives are weakened by the absence of COL1A1, DCN, ACTA2, and TAGLN from the ranked profile. Confidence: high. Plausible alternative labels include generic endothelial cell, vascular endothelial cell, capillary endothelial cell, and vein endothelial cell.

Ground truth

Most likely cell identity: cardiac endothelial cell. This assignment is supported by VWF and EMCN being present in the ranked genes, with ENG further reinforcing an endothelial signature. The lack of COL1A1 provides little support for a fibroblast interpretation.

D.1.2 Multi-Cell Immune Versus Myeloid Query

User question

I pasted two ranked gene lists above corresponding to profile IDs A and B. For each profile ID, provide a single best cell-type annotation with a confidence level and short rationale, and explicitly list the supporting evidence and negative markers used from each list.

BioAgent

Profile A—naive thymus-derived CD4-positive, alpha-beta T cell (confidence: high). The ranked list contains IL7R and LTB, which support a naive CD4 T-cell interpretation, while effector and activation alternatives are weakened by the absence of GZMB, IFNG, IL4, and IL5.
Profile B—mononuclear phagocyte (confidence: moderate). Antigen presentation markers including CD74, HLA-DRA, HLA-DRB1, and CST3 support an antigen-presenting mononuclear phagocyte state. B-cell and activated T-cell alternatives are weakened by the absence of canonical lymphoid and cytotoxic markers.

Ground truth

Profile A: naive thymus-derived CD4-positive, alpha-beta T cell—high confidence. IL7R and LTB are present, while canonical effector or activation markers including IFNG, IL4, IL5, and GZMB are not detected among the top-ranked genes.
Profile B: mononuclear phagocyte—high confidence. AIF1, TYROBP, and FCER1G appear in the ranked list, while lymphoid and endothelial markers including CD3D, MS4A1, PECAM1, and EPCAM are absent or low.

D.2 Biologically Distinctive Examples

D.2.1 Single-Cell Retinal Query

User question

I pasted a ranked gene list from a cluster above. Can you give the most likely cell identity and one or two plausible alternative annotations, with a brief explanation of the evidence?

BioAgent

Most likely identity: eye photoreceptor cell (predicted tissue: eye). This assignment is supported by neuronal genes SNAP25 and MAP2 in the ranked list. S100B, TMEM119, and COL1A1 are absent, weakening glial and fibroblast interpretations. Plausible alternatives include a generic retinal neuron or retinal bipolar neuron. Confidence: moderate because the detected neuronal genes are not fully photoreceptor specific.

Ground truth

Most likely identity: eye photoreceptor cell. MAP2 and SNAP25 support a neuronal retinal identity. The absence of glial or stromal markers such as S100B and COL1A1 argues against a glial or fibroblast interpretation. Plausible alternatives include other retinal neurons such as bipolar or amacrine cells because MAP2 and SNAP25 are neuronal but not fully photoreceptor specific.

D.2.2 Multi-Cell Epithelial Versus Perivascular Query

User question

I pasted two ranked gene lists above for profiles A and B. Could you annotate each profile and determine whether they represent related states within one lineage or clearly distinct cell identities?

BioAgent

Profile A—enterocyte of the epithelium proper of the ileum. The ranked genes include APOA1, APOA4, SLC5A1, FABP2, and ANPEP, supporting an absorptive ileal enterocyte identity. Confidence: high.
Profile B—pericyte. The ranked genes include RGS5, NOTCH3, SPARC, IGFBP7, and CALD1, supporting a perivascular mesenchymal identity. Confidence: moderate.
The two profiles represent clearly distinct cell identities rather than related states within one lineage.

Ground truth

Profile A: enterocyte of the epithelium proper of the ileum. APOA1, APOA4, and SLC5A1 support an absorptive enterocyte identity, whereas MUC2 and CHGA are not enriched and argue against goblet or enteroendocrine interpretations.
Profile B: pericyte. RGS5, NOTCH3, and COL4A1 support a pericyte or perivascular program, whereas PECAM1 and VWF are not enriched and argue against an endothelial identity.
The two profiles are clearly distinct cell identities rather than related states of one lineage.

Data availability

All experiments use the publicly available Tabula Sapiens single-cell atlas. Derived benchmark splits, per-sample model outputs, summary tables and plotting artifacts used for the manuscript are available in the accompanying project repository and will be archived with a persistent identifier upon publication.

Code availability

The code for domain alignment, expert specialization, textual orchestration, benchmark evaluation and figure generation will be available upon publication.

Competing interests

The authors declare no competing interests.

References

Abdelaal T, Michielsen L, Cats D, et al (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biology 20:194. https://doi.org/10.1186/s13059-019-1795-z (cited on pp. 2 and 10)

Ahlmann-Eltze C, Huber W, Anders S (2025) Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods 22:1657–1661. https://doi.org/10.1038/s41592-025-02772-6 (cited on p. 2)

Anthropic (2025) Introducing Claude Sonnet 4.5. URL https://www.anthropic.com/news/claude-sonnet-4-5, accessed: 2026-08-18 (cited on pp. 3 and 4)

Cao J, O'Day DR, Pliner HA, et al (2020) A human cell atlas of fetal gene expression. Science 370(6518):eaba7721. https://doi.org/10.1126/science.aba7721 (cited on p. 2)

Chen T, Kornblith S, Norouzi M, et al (2020) A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 119. PMLR, pp 1597–1607, URL https://proceedings.mlr.press/v119/chen20j.html (cited on p. 12)

Cui H, Wang C, Maan H, et al (2024) scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature Methods 21(8):1470–1480. https://doi.org/10.1038/s41592-024-02201-0 (cited on p. 2)

Diehl AD, Meehan TF, Bradford YM, et al (2016) The cell ontology 2016: enhanced content, modularization, and ontology interoperability. Journal of Biomedical Semantics 7:44. https://doi.org/10.1186/s13326-016-0088-7 (cited on p. 22)

Domínguez Conde C, Xu C, Jarvis LB, et al (2022) Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science 376(6594):eabl5197. https://doi.org/10.1126/science.abl5197 (cited on pp. 2 and 10)

Dubey A, et al (2024) The llama 3 herd of models. arXiv:2407.21783 (cited on p. 5)

Fang Z, et al (2025) Cell-o1: Training llms to solve single-cell reasoning problems via reinforcement learning. arXiv:2506.02911 (cited on pp. 2 and 5)

Gemma Team (2024) Gemma 2: Improving open language models at a practical size. arXiv:2408.00118 (cited on p. 5)

Hao M, Gong J, Zeng X, et al (2024) Large-scale foundation model on single-cell transcriptomics. Nature Methods 21(8):1481–1491. https://doi.org/10.1038/s41592-024-02305-7 (cited on p. 2)

Houlsby N, Giurgiu A, Jastrzebski S, et al (2019) Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning, vol 97. PMLR, pp 2790–2799 (cited on p. 15)

Hrovatin K, Sikkema L, Shitov VA, et al (2025) Considerations for building and using integrated single-cell atlases. Nature Methods 22(1):41–57. https://doi.org/10.1038/s41592-024-02532-y (cited on pp. 2 and 10)

Hu EJ, Shen Y, Wallis P, et al (2022) LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=nZeVKeeFYf9 (cited on p. 12)

Human Cell Atlas (2025) HCA data now freely available on AWS. URL https://www.humancellatlas.org/news/hca-data-now-freely-available-on-aws/, reports data from more than 66 million cells and more than 300 TB made available to the community (cited on p. 2)

Human Cell Atlas Data Portal (2020) HCA Data Portal 2.0 launches; new projects, contributor generated matrices and HCA Data Portal 2.0 infrastructure. URL https://data.humancellatlas.org/dcp-updates, reports raw data and standardized metadata for 5.8 million cells (cited on p. 2)

Jiang AQ, Sablayrolles A, Mensch A, et al (2023) Mistral 7b. arXiv:2310.06825 (cited on p. 5)

Kedzierska KZ, Crawford L, Amini AP, et al (2025) Zero-shot evaluation reveals limitations of single-cell foundation models. Genome Biology 26(1):101. https://doi.org/10.1186/s13059-025-03574-x (cited on p. 2)

Lähnemann D, Köster J, Szczurek E, et al (2020) Eleven grand challenges in single-cell data science. Genome Biology 21(1):31. https://doi.org/10.1186/s13059-020-1926-6 (cited on p. 2)

Levine D, Rizvi SA, Lévy S, et al (2024) Cell2Sentence: Teaching large language models the language of biology. In: Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 235. PMLR, pp 27299–27325, URL https://proceedings.mlr.press/v235/levine24a.html (cited on pp. 2 and 12)

Li Z, Hoiem D (2016) Learning without forgetting. In: European Conference on Computer Vision. Springer, pp 614–629, https://doi.org/10.1007/978-3-319-46493-0_37 (cited on pp. 10 and 13)

Luecken MD, Büttner M, Chaichoompu K, et al (2022) Benchmarking atlas-level data integration in single-cell genomics. Nature Methods 19(1):41–50. https://doi.org/10.1038/s41592-021-01336-8 (cited on p. 2)

McInnes L, Healy J, Saul N, et al (2018) UMAP: Uniform manifold approximation and projection. Journal of Open Source Software 3(29):861. https://doi.org/10.21105/joss.00861, URL https://doi.org/10.21105/joss.00861 (cited on p. 6)

OpenAI (2024) GPT-4o system card. arXiv:2410.21276 (cited on p. 5)

OpenAI (2025) GPT-5 system card. URL https://openai.com/index/gpt-5-system-card/ (cited on p. 5)

OpenAI (2026) Introducing GPT-5.4 mini and nano. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/, accessed: 2026-08-18 (cited on pp. 3 and 4)

Park W, Kim D, Lu Y, et al (2019) Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3967–3976, https://doi.org/10.1109/CVPR.2019.00409 (cited on p. 13)

Pfeiffer J, R"ucklé A, Poth C, et al (2020) AdapterHub: A framework for adapting transformers. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp 46–54, https://doi.org/10.18653/v1/2020.emnlp-demos.7 (cited on p. 15)

Pfeiffer J, Kamath A, R"ucklé A, et al (2021) AdapterFusion: Non-destructive task composition for transfer learning. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp 487–503, https://doi.org/10.18653/v1/2021.eacl-main.39 (cited on p. 15)

Pryzant R, Iter D, Li J, et al (2023) Automatic prompt optimization with "gradient descent" and beam search. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp 7957–7968, https://doi.org/10.18653/v1/2023.emnlp-main.494 (cited on p. 15)

Qwen Team (2025) Qwen2.5 technical report. arXiv:2412.15115 (cited on p. 5)

Regev A, Teichmann SA, Lander ES, et al (2017) The human cell atlas. eLife 6:e27041. https://doi.org/10.7554/eLife.27041 (cited on p. 2)

Rizvi SA, Levine D, Patel A, et al (2025) Scaling large language models for next-generation single-cell analysis. bioRxiv https://doi.org/10.1101/2025.04.14.648850, preprint (cited on pp. 5 and 12)

Rood JE, Wynne S, Robson L, et al (2025) The human cell atlas from a cell census to a unified foundation model. Nature 637:1065–1071. https://doi.org/10.1038/s41586-024-08338-4 (cited on p. 2)

Sade-Feldman M, Yizhak K, Bjorgaard SL, et al (2018) Defining T cell states associated with response to checkpoint immunotherapy in melanoma. Cell 175(4):998–1013.e20. https://doi.org/10.1016/j.cell.2018.10.038 (cited on p. 2)

Schick T, Dwivedi-Yu J, Dessì R, et al (2023) Toolformer: Language models can teach themselves to use tools. In: Advances in Neural Information Processing Systems, URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html (cited on pp. 3 and 14)

Sun B, Saenko K (2016) Deep CORAL: Correlation alignment for deep domain adaptation. In: Computer Vision – ECCV 2016 Workshops. Springer, pp 443–450, https://doi.org/10.1007/978-3-319-49409-8_35 (cited on p. 13)

Svensson V, Vento-Tormo R, Teichmann SA (2018) Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols 13(4):599–604. https://doi.org/10.1038/nprot.2017.149 (cited on p. 2)

Tabula Sapiens Consortium (2022) The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science 376(6594):eabl4896. https://doi.org/10.1126/science.abl4896 (cited on pp. 2, 3, and 4)

Theodoris CV, Xiao L, Chopra A, et al (2023) Transfer learning enables predictions in network biology. Nature 618(7965):616–624. https://doi.org/10.1038/s41586-023-06139-9 (cited on p. 2)

Tirosh I, Izar B, Prakadan SM, et al (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352(6282):189–196. https://doi.org/10.1126/science.aad0501 (cited on p. 2)

Yang A, et al (2025) Qwen3 technical report. arXiv:2505.09388 (cited on p. 5)

Yang C, Wang X, Lu Y, et al (2024) Large language models as optimizers. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=Bb4VGOWELI (cited on p. 15)

Yao S, Zhao J, Yu D, et al (2023) ReAct: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=WE_vluYUL-X (cited on pp. 3 and 14)

Yuksekgonul M, Bianchi F, Boen J, et al (2025) TextGrad: Automatic differentiation via text. Nature 639:609–616. https://doi.org/10.1038/s41586-025-08661-4 (cited on p. 15)

Zhang Y, et al (2025) Agentic context engineering: Evolving contexts for self-improving language models. arXiv:2510.04618 (cited on pp. 3 and 15)