scRAG: Hybrid Retrieval-Augmented Generation for LLM-based Cross-Tissue Single-Cell Annotation

Zhiyin Yu[♡ *], Chao Zheng[♠], Chong Chen[◇ †], Xian-Sheng Hua[◇], Xiao Luo[♣]

[♡] South China University of Technology [♠] University of Southampton
[◇] Terminus Group [♣] University of California, Los Angeles

zhiyinyu89@gmail.com, czly20@soton.ac.uk, chenchong.cz@gmail.com
huaxiansheng@gmail.com, xiaoluo@cs.ucla.edu

Abstract

In recent years, large language models (LLMs) such as GPT-4 have demonstrated impressive potential in a wide range of fields, including biology, genomics and healthcare. Numerous studies have attempted to apply pre-trained LLMs to single-cell data analysis within one tissue. However, when it comes to cross-tissue cell annotation, LLMs often suffer from unsatisfactory performance due to the lack of specialized biological knowledge regarding genes and tissues. In this paper, we introduce scRAG, a novel framework that incorporates advanced LLM-based RAG techniques into cross-tissue single-cell annotation. scRAG utilizes LLMs to retrieve structured triples from knowledge graphs and unstructured similar cell information from the reference cell database, and it generates candidate cell types. The framework further optimizes predictions by retrieving marker genes from both candidate cells and similar cells to refine its results. Extensive experiments on a cross-tissue dataset demonstrate that our scRAG framework outperforms various baselines, including generalist models, domain-specific methods, and trained classifiers. The source code is available at `https://github.com/YuZhiyin/scRAG`.

1 Introduction

Large language models (LLMs) like GPT-4 have highlighted their exceptional capabilities in natural language processing tasks (Naveed et al., 2023; Luo et al., 2025), including text generation and comprehension. Their impact has extended beyond linguistics, playing an increasingly significant role in different research domains (Sandmann et al., 2025; Zhang et al., 2024). In biological sciences, LLMs show tremendous potential in interpreting biological data and following human instructions to solve downstream tasks, such as cell annotation, with efficiency that outperforms human capacity (Hou and Ji, 2024). These advancements suggest that LLMs are poised to open new frontiers in biology, revolutionizing research paradigms, and deepening our understanding of the life sciences.

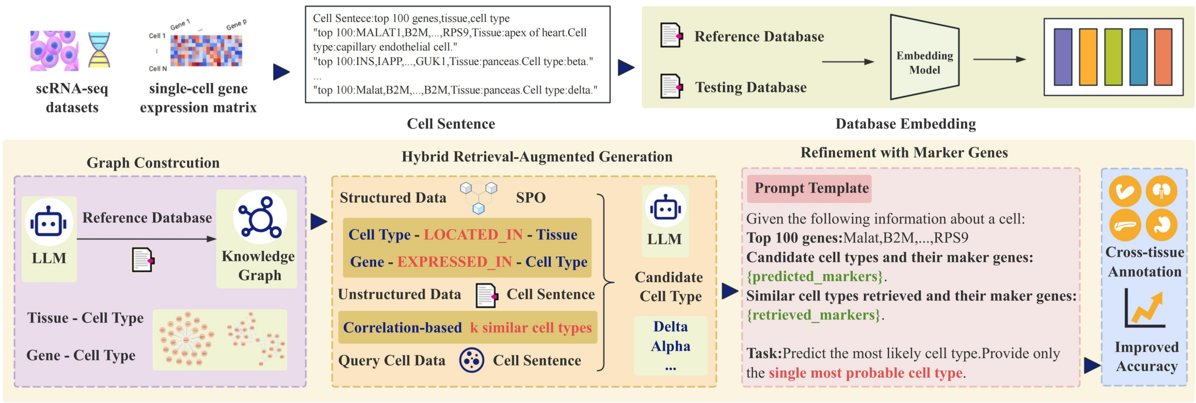
Cell type annotation is crucial in single-cell RNA sequencing (scRNA-seq) analysis, mapping gene expression profiles to specific cell types. While databases like GEO (Barrett et al., 2012) and HCA (Regev et al., 2017) have propelled the field, traditional methods face challenges. Cluster-then-Annotate strategies rely on marker genes, which can introduce bias (Huang et al., 2021), and correlation-based methods often struggle with the high-dimensional and sparse nature of scRNA-seq data (Serra et al., 2018). Supervised and semi-supervised methods are sensitive to hyperparameter tuning and often fail to capture complex gene-gene interactions (Ma and Pellegrini, 2020; Cao et al., 2020b; Li et al., 2020; Yang et al., 2022). These shortcomings assert the need for more robust and accurate cell annotation approaches.

The emerging field of foundational single-cell models aims to address these challenges. At the data representation level, natural language processing techniques are applied to transform gene expression matrices from scRNA-seq data into computable formats, such as cell sentences (Fang et al., 2024; Levine et al., 2023) or embeddings (Chen and Zou, 2023). These representations establish a bridge between raw biological data and large language models. At the model application level, large language models leverage these processed representations through extensive pre-training and fine-tuning (Cui et al., 2024; Yang et al., 2022; Theodoris et al., 2023), thereby enhancing their performance in cell annotation tasks.

However, despite their potential, current pre-trained large language models often struggle to capture the complex relationships between tissues

[†] Corresponding author.
[*] The work is done when Zhiyin Yu interned in Terminus.

Findings of the Association for Computational Linguistics: ACL 2025, pages 954–970
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics



Cell Sentece:top 100 genes,tissue,cell type
"top 100:MALAT1,B2M,...,RPS9,Tissue:apex of heart.Cell type:capillary endothelial cell."
"top 100:INS,IAPP,...,GUK1,Tissue:panceas.Cell type:beta."
...
"top 100:Malat,B2M,...,B2M,Tissue:panceas.Cell type:delta."

Given the following information about a cell:
Top 100 genes: Malat,B2M,...,RPS9
Candidate cell types and their maker genes:
{predicted_markers}.
Similar cell types retrieved and their maker genes:
{retrieved_markers}.

Task: Predict the most likely cell type.Provide only the **single most probable cell type**.

Figure 1: **Framework of scRAG.** The framework consists of four stages: (1) Sentence Construction: Transforming scRNA-seq data into cell sentences, including gene and tissue information. (2) Graph Construction: Building two knowledge graphs, Tissue-Cell Type and Gene-Cell Type. (3) Hybrid Retrieval-Augmented Generation: The LLM retrieves structured triples from the knowledge graphs and similar cells from the reference database to predict candidate cell types. (4) Refinement with marker genes: Retrieving marker genes of candidate and similar cell types, where the LLM refines and confirms the final cell type annotation.

and cell types, primarily because they lack specialized biological knowledge about genes and tissues (Hou and Ji, 2024; Fischer et al., 2024). They rely solely on internal knowledge for reasoning, which limits their ability to generalize cell annotations across diverse tissue contexts. Additionally, existing domain-specific methods cannot effectively incorporate cell-cell similarity as supplementary information (Fang et al., 2024), resulting in unsatisfactory performance.

In this study, we introduce scRAG to address these limitations. To capture tissue-cell type relationships, we construct knowledge graphs linking tissues, cell types, and genes, to generate structured data triples that represent their biological relationships. To capture cell-cell similarity, we retrieve unstructured data on similar cells from a reference dataset. The LLM integrates the above two types of data, i.e.,structured triples and unstructured similar cell information, to predict candidate cell types. Finally, by retrieving and analyzing marker genes from both candidate and similar cell types, scRAG determines the final annotations. The methodology of scRAG closely aligns with expert workflows: LLMs leverage external knowledge bases as prior information and validate preliminary predictions using marker genes, improving the accuracy of cross-tissue annotation in single-cell analysis.

The main contributions are as follows:

- We introduce scRAG, a novel hybrid Retrieval-Augmented Generation framework tailored for single-cell data, bridging large language models to advance cell annotation across tissues.

- Our scRAG leverages structured triples, unstructured information from similar cells, and refines predictions with marker genes for more accurate cross-tissue cell annotations.

- We carry out extensive numeric experiments, showing that scRAG significantly outperforms all baselines in cross-tissue annotation, including generalist models, domain-specific methods, and trained classifiers.

2 Related Work

2.1 Traditional Cell Type Annotation

Accurate cell type annotation is essential in single-cell RNA-sequencing (scRNA-seq) studies for understanding complex tissues at the single-cell level. Marker gene-based approaches rely on clustering cells and assigning annotations based on marker gene databases including CellMarker (Zhang et al., 2019) and PanglaoDB (Franzén et al., 2019), with tools like MACA (Xu et al., 2022) and SCSA (Cao et al., 2020a) automating this process. In contrast, correlation-based methods focus on aligning gene expression profiles of query cells with reference datasets using similarity metrics. Representative tools include CIPR (Ekiz et al., 2020) and ClustifyR (Fu et al., 2020) for cluster-level comparisons, as well as scmap (Kiselev et al., 2018) and scMatch (Hou et al., 2019) for single-cell-level correlations. Machine learning models have been employed to annotation tasks as well, including weighted KNN in scClassify (Lin et al., 2020), Random Forest in SingleCellNet (Tan and Cahan, 2019), and a neural network in ACTINN (Ma and Pellegrini, 2019). Deep learning methods like scTab (Fischer et al., 2024) use TabNet and data augmentation for cross-tissue annotation. However, traditional methods often suffer from limited generalization, as their performance often depends on the quality and composition of training data, prompting the need for more integrative solutions.

2.2 LLMs in Cell Type Annotation

Large language models (LLMs) have recently shown advancements in improving accuracy and efficiency in cell type annotation. Methods such as scBERT (Yang et al., 2022), Geneformer (Theodoris et al., 2023), and scGPT (Cui et al., 2024) pre-train LLMs on large scRNA-seq datasets to learn gene-gene interactions and fine-tune them for annotation tasks. Multi-agent frameworks like CellAgent (Xiao et al., 2024) introduce hierarchical decision-making and role coordination for efficient cell analysis. Embedding-based methods, such as GenePT (Chen and Zou, 2023), leverage OpenAI's embedding model to represent genes and cells for downstream tasks. Sentence generation techniques, including ChatCell (Fang et al., 2024) and Cell2Sentence (Levine et al., 2023), convert single-cell gene expression profiles into natural language sentences for further analysis. Additionally, GPTCelltype (Hou and Ji, 2024) evaluates GPT-4's natural language capabilities in cell type annotation, further highlighting the potential of LLMs in this field. However, LLMs often lack specialized biological knowledge, and pre-training and fine-tuning require substantial computational resources. On the contrary, the proposed method is training-free yet effective.

2.3 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) enhances the accuracy and contextual relevance of responses by incorporating relevant information from external knowledge sources during the generation process (Lewis et al., 2021). For graph-based RAG, KG-RAG (Soman et al., 2024) uses biomedical knowledge graphs to enhance LLMs for biomedical QA, while GraphRAG (Edge et al., 2025) builds knowledge graphs and community summaries to effectively handle questions requiring global understanding. For document-based RAG, GeneRAG (Lin et al., 2024) leverages NCBI gene data to improve genetics-related responses, but only retrieves similar genes rather than similar cells. In contrast, our proposed scRAG extends the above paradigms through a hybrid retrieval strategy: it retrieves structured triples from knowledge graphs and unstructured similar cells, followed by marker gene-based refinement. This significantly improves performance in cross-tissue cell type annotation.

3 Methodology

3.1 Framework Overview

As illustrated in Figure 1, our scRAG workflow is a multi-step process that enhances cross-tissue cell annotation accuracy. First, scRNA-seq data from single cells are transformed into cell sentences. Then, tissue-cell type and gene-cell type knowledge graphs are constructed using a reference database. Upon receiving a query cell sentence from the user, LLM extracts entities and retrieves relevant triples from the knowledge graphs. Next, we build embeddings and use cosine similarity to search for similar cells within the reference cell database. The hybrid retrieval-augmented generation mechanism then integrates information from both knowledge graphs and the reference database, enabling the LLM to generate candidate cell types. Finally, LLM retrieves marker gene information for both candidate and similar cell types, and determines the most accurate cell type annotation.

3.2 Sentence Construction

To facilitate interaction with large language models and streamline the representation of scRNA-seq data, we adopt the CelltoSentence approach (Levine et al., 2023), which transforms h5ad files into a natural language format. The h5ad file consists of three primary components: the X matrix, which is a sparse matrix representing gene expression levels for each cell; the obs metadata, which contains information such as cell type and tissue; and the var component, which stores gene information including gene names and gene IDs. Given the large scale and complexity of the sparse matrix, we simplify the representation by generating a sequence of gene names, structured like a sentence in natural language. Each cell is represented by its top 100 genes with the highest normalized expression levels, ranked from highest to lowest, while genes with zero expression are omitted. The resulting cell sentence is composed of the top 100 expressed genes along with the corresponding tissue and cell

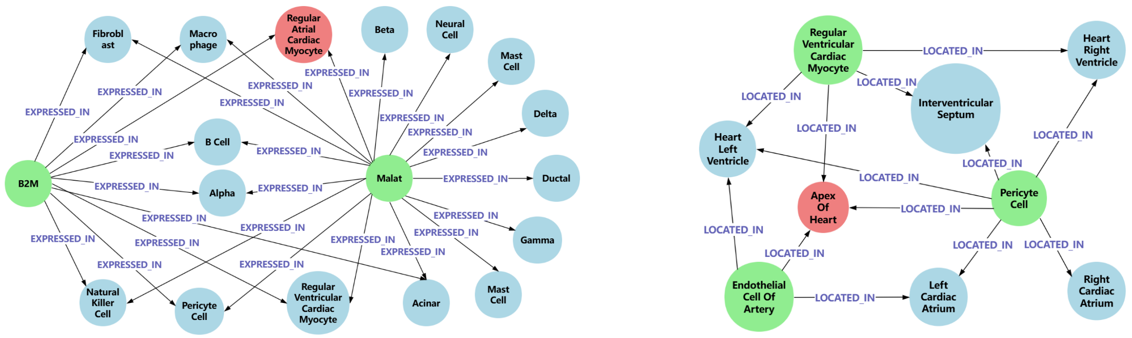


Figure 2: **Two Generated Knowledge Graphs.** The left panel shows the gene-cell type graph, with green nodes representing genes, blue nodes representing cell types, and the red node highlighting one of the cell types in which two genes are co-expressed. The right panel shows the tissue-cell type graph, where green nodes represent cell types, blue nodes represent tissues, and the red node highlights one of the tissues in which two cell types are co-expressed. The edges represent the "EXPRESSED_IN" and "LOCATED_IN" relationships, respectively.

type. This transformation into a natural language sentence enables subsequent graph construction and enhances compatibility with LLM-driven tasks.

3.3 Graph Construction

To systematically model the relationships among tissues, genes, and cell types, we construct knowledge graphs using a reference dataset of 1,350 cell sentences. The LLM extracts entities from the cell sentences, which are then structured and stored within a Neo4j database. Utilizing GPT-4-Turbo as the foundational model, LLM identifies entities (such as genes, tissues, and cell types) and relationships, representing them as nodes and edges in the graph database. Two knowledge graphs are generated: a tissue-cell type graph and a gene-cell type graph, as shown in Figure 2. The tissue-cell type graph captures the associations between tissues and their possible resident cell types, while the gene-cell type graph illustrates the high-expression relationships between genes and specific cell types. These graphs provide structured priors, which are used by the LLM to improve the downstream cell annotation performance.

3.4 Hybrid Retrieval-Augmented Generation

We propose a hybrid Retrieval-Augmented Generation mechanism that combines structured data retrieval from knowledge graphs with unstructured data retrieval from a reference cell database to predict candidate cell types. By leveraging multiple data sources, this approach substantially strengthens the model's ability to infer the most likely cell type.

Structured Data Retrieval. For structured data, we construct two knowledge graphs from a reference dataset that contains entities such as tissues, genes, and cell types. The retrieval process focuses on extracting specific relationships between these entities, represented as triples. There are two main types of relationships used for cell type annotation:

- Cell type - LOCATED_IN → Tissue
- Gene - EXPRESSED_IN → Cell type

These triples represent key biological relationships among genes, cell types, and tissues. The retrieval process begins by querying the knowledge graph to obtain relevant triples, followed by merging similar entries to reduce redundancy. If multiple triples indicate that a specific cell type is expressed in different tissues, or a particular gene is expressed in various cell types, we consolidate this information into a single, non-redundant entry:

$$T_{\text{struct}} = \{E_{i,1} - R_i \rightarrow \{E_{i,2,1}, E_{i,2,2}, \ldots, E_{i,2,m}\}\}, \tag{1}$$

where $E_{i,1}$ and the set $\{E_{i,2}\}$ represent relationships between cell types and tissues or genes and cell types, and R_i denotes the relationship type (such as LOCATED_IN or EXPRESSED_IN).

Unstructured Data Retrieval. For unstructured data, we perform document retrieval to extract similar cell sentences from the reference cell database. A vector-based retrieval mechanism is utilized to perform similarity searches, retrieving the four most relevant sentences for each query. Each retrieved sentence contains a list of the top 100 genes expressed in a particular cell, along with the corresponding tissue and cell type. From these retrieved sentences, we extract the cell types, denoted as **similar cell types**. Formally, the unstructured data retrieval process can be represented as:

$$T_{\text{unstruct}} = \{C_1, C_2, C_3, C_4\}, \tag{2}$$

where T_{unstruct} represents the set of similar cell types C_i (with $i = 1, 2, 3, 4$) retrieved from the reference cell database.

Candidate Cell Type Generation. After retrieving both structured and unstructured data, we integrate these sources to improve the prediction of candidate cell types. The LLM processes the structured triples and similar cell types obtained from unstructured data. Based on the combined information, GPT-3.5-Turbo generates a list of potential candidate cell types. The task of generating candidate cell types can be represented as:

$$\text{Candidate cell types} = \text{LLM}(T_{\text{struct}}, T_{\text{unstruct}}), \tag{3}$$

where LLM processes the structured data T_{struct} and unstructured data T_{unstruct} to predict the most probable candidate cell types.

By employing hybrid Retrieval-Augmented Generation, we effectively combine knowledge from both structured knowledge graphs and unstructured reference cell database. This hybrid approach enables a more robust and contextually aware generation of candidate cell types, improving the overall accuracy and reliability of cell annotation tasks.

3.5 Refinement with Marker Genes

We further enhance the accuracy and robustness of the final cell type prediction by utilizing large language models and integrating marker gene information. Following the generation of both candidate and similar cell types, the refinement process begins by retrieving the associated marker genes for these cell types to further improve prediction power.

This refinement process is guided by three critical inputs provided to the LLM: (1) the top 100 expressed genes of the query cell, (2) the candidate cell types and their corresponding marker genes, and (3) the similar cell types and their corresponding marker genes. With these inputs, the LLM assesses the overlap and biological relevance of the marker genes, in the context of the query cell's gene expression profile, to ensure the prediction aligns with established biological patterns.

By refining with marker genes, we substantially increase the reliability of cross-tissue cell type annotation. This approach not only leverages the full potential of hybrid Retrieval-Augmented Generation, which integrates both structured and unstructured biological data, but also draws on the critical knowledge of biologically relevant markers. Through this process, the LLM is guided to identify the most accurate cell type for annotation, delivering precise and biologically grounded results.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate the proposed method on three publicly available human single-cell datasets: Human heart atlas, Human pancreas data, and Human PBMC. The Human heart atlas dataset contains cells from 1 general tissue, 6 sub-tissues, and 27 cell types (Litviňuková et al., 2020). The Human pancreas dataset includes 1 tissue and 7 cell types (Baron et al., 2016). The Human PBMC dataset consists of 1 tissue and 16 cell types (Oetjen et al., 2018). From each of these datasets, we randomly select 450 samples for constructing the reference cell dataset and 300 samples for the testing cell dataset. Ultimately, we generate two cross-tissue datasets, and test our model's capability in cross-tissue cell annotation. Additionally, the marker gene database is compiled by aggregating data from (Xu et al., 2022). More details on the dataset can be found in the Appendix A.

Baselines. We compare our method to three types of baselines:

- **Generalist Models:** We compare with GPT-3.5-Turbo, GPT-4-Turbo, and GPT-4o-Mini (OpenAI, 2023), as well as their chain-of-thought (CoT) (Kojima et al., 2023) variants, e.g., GPT-3.5-Turbo (CoT) and GPT-4-Turbo (CoT).

- **Domain-Specific Methods:** We compare with GenePT (Chen and Zou, 2023) using cosine similarity, Cell2Sentence (Levine et al., 2023), and Chatcell (Fang et al., 2024), all specifically tailored for single-cell analysis.

- **Trained Classifiers:** We combine GenePT (Chen and Zou, 2023) as an embedding generator with trained classifiers, including Logistic Regression (Cramer, 2002), Support Vector Machines (SVM) (Hearst et al., 1998), K-Nearest

Types	Models	Accuracy		F1 score	Precision	Recall
		FM	FM+PM			
Generalist Models	GPT-3.5-Turbo (OpenAI, 2023)	0.2167	0.2567	0.0991	0.1109	0.1308
	GPT-4o-Mini (OpenAI, 2023)	0.2522	0.2989	0.1015	0.1158	0.1502
	GPT-4-Turbo (OpenAI, 2023)	0.4011	0.4678	0.2189	0.3040	0.2599
	GPT-3.5-Turbo (CoT) (Kojima et al., 2023)	0.2200	0.2533	0.0829	0.1026	0.1182
	GPT-4o-Mini (CoT) (Kojima et al., 2023)	0.2611	0.3067	0.1070	0.1258	0.1676
	GPT-4-Turbo (CoT) (Kojima et al., 2023)	0.3711	0.4278	0.1799	0.2379	0.2111
Domain-Specific Methods	GenePT (Chen and Zou, 2023)	0.6711	0.6756	0.4266	0.4723	0.4346
	Cell2Sentence (Levine et al., 2023)	0.3744	0.5178	0.1187	0.1445	0.1245
	ChatCell (Fang et al., 2024)	0.4156	0.4867	0.1861	0.2518	0.2146
Trained Classifiers	Logistic Regression (Cramer, 2002)	0.6589	0.6633	0.4165	0.4786	0.4259
	Support Vector Machines (Hearst et al., 1998)	0.6689	0.6733	0.4181	0.4391	0.4370
	K-Nearest Neighbors (Peterson, 2009)	**0.8067**	0.8111	0.5369	0.5713	0.5497
	Transformers (Vaswani et al., 2023)	0.6689	0.6722	0.4058	0.4038	0.4252
scRAG	GPT-3.5-Turbo (scRAG)	0.6611	0.6700	0.4538	0.5212	0.4704
	GPT-4o-Mini (scRAG)	0.7667	0.7789	**0.5744**	**0.6117**	**0.5895**
	GPT-4-Turbo (scRAG)	**0.8056**	**0.8122**	**0.5998**	**0.6429**	**0.6045**
	GPT-3.5-Turbo (scRAG+CoT)	0.6422	0.6522	0.4393	0.5172	0.4522
	GPT-4o-Mini (scRAG+CoT)	0.7744	0.7911	0.5431	0.5728	0.5674
	GPT-4-Turbo (scRAG+CoT)	0.8011	**0.8200**	0.5590	0.6113	0.5735

Table 1: Performance comparison with different baselines on the cross-tissue dataset. The **boldfaced** scores represent the **best** and **second-best** results.

Neighbors (KNN) (Peterson, 2009) and Transformers (Vaswani et al., 2023).

Implementation Details. We implement the workflow using the LangChain framework (Mavroudis, 2024). GPT-4-Turbo is employed for graph construction, and GPT-3.5-Turbo is used for the hybrid Retrieval-Augmented Generation mechanism. For prediction generation and refinement with marker genes, we test all baseline models to evaluate the results they produce. Zero-shot prompting (Kojima et al., 2022) and add "Let's think step by step" (Kojima et al., 2023) are applied to the prompt for CoT baselines. We select the top 4 similar cells and top 2 candidate cells to balance efficiency and accuracy. The temperature parameter is set to be 0 for each step. Further implementation details are deferred to Appendix B and Appendix E.

Metrics. We evaluate the model's performance using two sets of metrics: (i) *Accuracy* (including *FM* and *FM+PM*) and *F1 Score*; (ii) *Precision* and *Recall* (all calculated as macro averages). The definitions of *Fully Match* (FM), *Partially Match* (PM), and *Mismatch* (M) are as follows:

- **Fully Match (FM):** Prediction matches the ground truth or is a subtype of ground truth.
- **Partially Match (PM):** Prediction represents a parent type of the ground truth.
- **Mismatch (M):** Prediction does not correspond to the ground truth.

4.2 Main Results

The results on the cross-tissue benchmark are shown in Table 1. From these results, we obtain the following conclusions:

Comparison with Generalist Models. As shown in Table 1, scRAG leads to a substantial 44.44% improvement in fully matched accuracy and a 41.33% improvement in fully+partially matched accuracy for GPT-3.5-Turbo compared to its baseline. GPT-4o-Mini achieves greater improvements of 51.45% in fully matched accuracy and 48.00% in fully+partially matched accuracy over its baseline. GPT-4-Turbo also demonstrates significant improvements, with fully matched accuracy increasing by 40.45% and fully+partially matched accuracy improving by 34.44%. These improvements highlight the fact that LLMs often lack specialized biological knowledge, particularly regarding genes and tissues. The incorporation of external information from knowledge graphs and reference cell databases helps fill these gaps, enhancing the LLM's ability to process and generate more accurate results in cross-tissue cell annotation tasks.

Comparison with Domain-Specific Models. The scRAG-optimized GPT-4-Turbo achieves a 13.45% increase in fully matched accuracy compared to

Ablation Study	w/o structured data	w/o unstructured data	w/o marker genes	scRAG
GPT-3.5-Turbo	0.6380	0.6380	0.6580	**0.6611**
GPT-4o-Mini	0.7390	0.7280	0.7456	**0.7667**
GPT-4-Turbo	0.7580	0.7490	0.7360	**0.8056**

Table 2: Ablation study via accuracy performance comparison of different variants on scRAG.

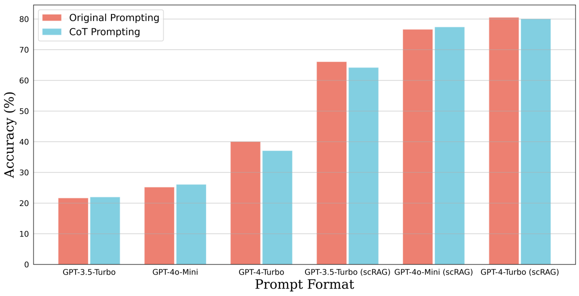


Figure 3: Comparison of CoT Prompting.

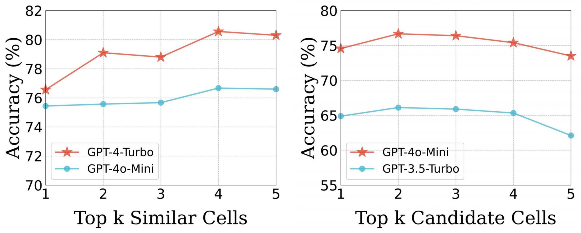


Figure 4: Analysis of scRAG's performance on top-k similar cell retrieval and top-k candidate cell generation.

GenePT, a 43.12% increase over Cell2Sentence, and a 39.00% increase over ChatCell. These results underscore the robustness and efficacy of the scRAG method in enhancing the performance of large language models. Moreover, scRAG offers greater scalability and flexibility. While other domain-specific models are typically constrained by fixed architecture or dataset, scRAG enables adaptation to evolving models and new datasets, ensuring continuous performance improvement.

Comparison with Trained Classifiers. Results in Table 1 show that scRAG achieves comparable performance to GenePT combined with KNN, surpassing KNN in FM+PM accuracy by 0.11%, whilst significantly outperforming LR, SVM, and Transformers in FM accuracy by 14.67%, 13.67%, and 13.67%, respectively. In addition to its superior performance, scRAG provides enhanced stability and operates as an end-to-end framework without the need for classifier training, hyper-parameter tuning, or optimization. This training-free approach simplifies deployment, improves scalability, and makes scRAG a more practical and efficient solution for cross-tissue cell annotation tasks.

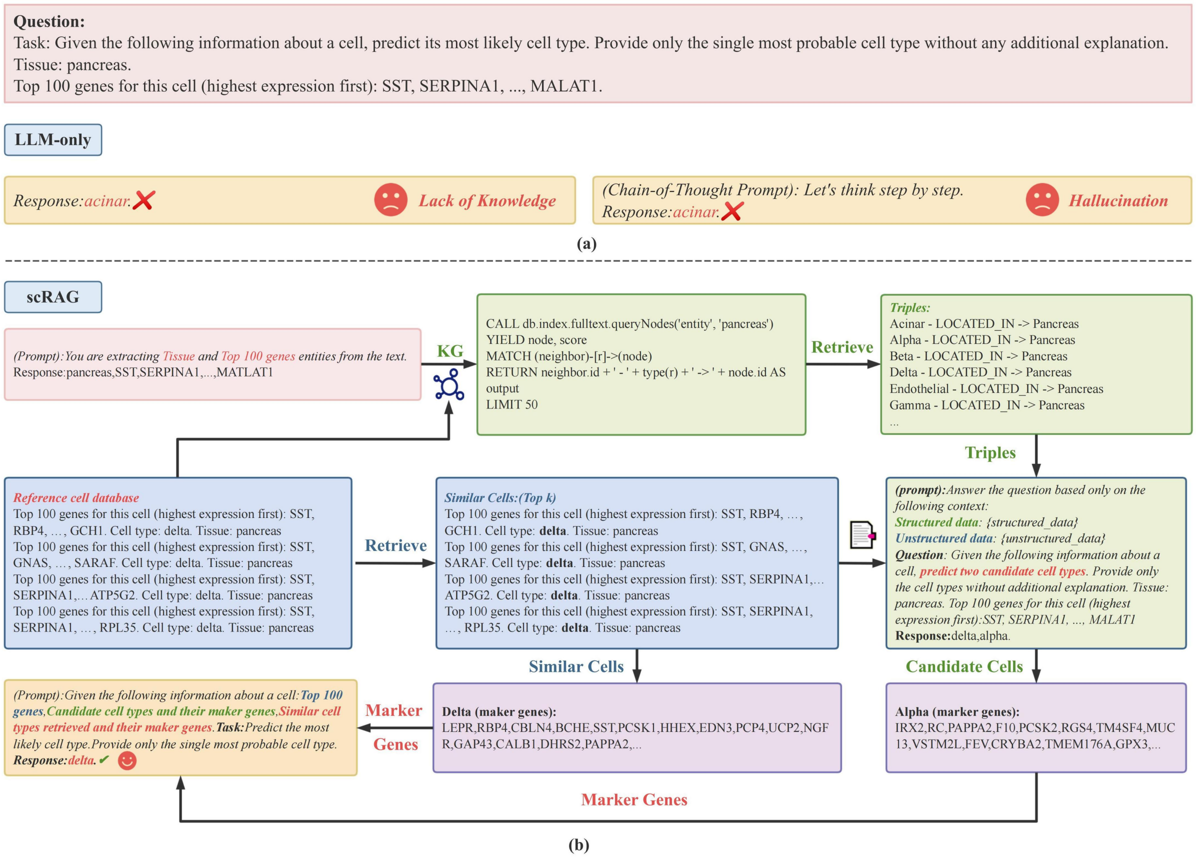
4.3 Further Analysis

In this section, we discuss several factors that would affect the performance of scRAG: CoT prompting, top-k similar cell types, top-k candidate cell types, and the result of the ablation study. More discussion is provided in Appendix C.

CoT Prompting. Figure 3 illustrates the impact of CoT prompting in both the baseline and scRAG settings. In the baseline setting, CoT prompting effectively helps weaker models like GPT-3.5-Turbo improve their performance by guiding them to think step-by-step and reason towards accurate annotation results. However, in the scRAG framework, which combines hybrid Retrieval-Augmented Generation and marker gene information, the addition of CoT creates different dynamics. For GPT-3.5-Turbo, the integration of CoT appears to overwhelm the model, leading to diminished performance. In contrast, GPT-4o-Mini benefits from this combination, as their stronger reasoning capabilities allow them to better utilize both scRAG and CoT, resulting in improved outcomes.

Top-k Similar Cell Types. As shown in Figure 4, we analyze the effect of retrieving top-k similar cell types on the performance of GPT-4-Turbo and GPT-4o-Mini. Our results show that accuracy improves as k increases, reaching its peak at $k = 4$, before slightly decreasing when k becomes larger. This indicates that too few references limit the model's access to contextual knowledge, while too many references introduce noise and increase the likelihood of hallucination. The improvement with higher k highlights scRAG's ability to leverage additional context effectively, but the diminishing in larger k emphasizes the importance of selecting the most relevant references.

Top-k Candidate Cell Types. As shown in Figure 4, generating the top-k candidate cell types for GPT-4o-Mini and GPT-3.5-Turbo reveals that performance peaks at $k = 2$ and declines at $k = 5$. This implies that using two candidates achieves an optimal balance between focus and diversity, enhancing the model's decisions. Too few candidates



Question:
Task: Given the following information about a cell, predict its most likely cell type. Provide only the single most probable cell type without any additional explanation.
Tissue: pancreas.
Top 100 genes for this cell (highest expression first): SST, SERPINA1, ..., MALAT1.

Response:acinar. ✗

(Chain-of-Thought Prompt): Let's think step by step.
Response:acinar. ✗

(Prompt):You are extracting Tissue and Top 100 genes entities from the text.
Response:pancreas,SST,SERPINA1,...,MATLAT1

```
CALL db.index.fulltext.queryNodes('entity', 'pancreas')
YIELD node, score
MATCH (neighbor)-[r]->(node)
RETURN neighbor.id + ' - ' + type(r) + ' -> ' + node.id AS
output
LIMIT 50
```

Triples:
Acinar - LOCATED_IN -> Pancreas
Alpha - LOCATED_IN -> Pancreas
Beta - LOCATED_IN -> Pancreas
Delta - LOCATED_IN -> Pancreas
Endothelial - LOCATED_IN -> Pancreas
Gamma - LOCATED_IN -> Pancreas
...

Reference cell database
Top 100 genes for this cell (highest expression first): SST, RBP4, …, GCH1. Cell type: delta. Tissue: pancreas
Top 100 genes for this cell (highest expression first): SST, GNAS, …, SARAF. Cell type: delta. Tissue: pancreas
Top 100 genes for this cell (highest expression first): SST, SERPINA1,… ATP5G2. Cell type: delta. Tissue: pancreas
Top 100 genes for this cell (highest expression first): SST, SERPINA1, …, RPL35. Cell type: delta. Tissue: pancreas

Similar Cells:(Top k)
Top 100 genes for this cell (highest expression first): SST, RBP4, …, GCH1. Cell type: **delta**. Tissue: pancreas
Top 100 genes for this cell (highest expression first): SST, GNAS, …, SARAF. Cell type: **delta**. Tissue: pancreas
Top 100 genes for this cell (highest expression first): SST, SERPINA1,… ATP5G2. Cell type: **delta**. Tissue: pancreas
Top 100 genes for this cell (highest expression first): SST, SERPINA1, …, RPL35. Cell type: **delta**. Tissue: pancreas

(prompt):*Answer the question based only on the following context:*
Structured data*: {structured_data}*
Unstructured data*: {unstructured_data}*
Question*: Given the following information about a cell,* ***predict two candidate cell types****. Provide only the cell types without additional explanation. Tissue: pancreas. Top 100 genes for this cell (highest expression first):SST, SERPINA1, ..., MALAT1*
Response:delta,alpha.

Delta (maker genes):
LEPR,RBP4,CBLN4,BCHE,SST,PCSK1,HHEX,EDN3,PCP4,UCP2,NGFR,GAP43,CALB1,DHRS2,PAPPA2,...

Alpha (marker genes):
IRX2,RC,PAPPA2,F10,PCSK2,RGS4,TM4SF4,MUC13,VSTM2L,FEV,CRYBA2,TMEM176A,GPX3,...

(Prompt):Given the following information about a cell: ***Top 100 genes****,* ***Candidate cell types and their maker genes****,* ***Similar cell types retrieved and their maker genes****.* ***Task:****Predict the most likely cell type.Provide only the single most probable cell type.*
Response:delta. ✓

Figure 5: **Case Study.** The upper panel illustrates the limitations of LLM-only approaches, while the lower panel showcases the effectiveness of the scRAG framework.

restrict diversity, while too many introduce excessive information therefore lowering the accuracy.

Ablation Study. In Table 2, we evaluate the individual components of scRAG through an ablation study. Specifically, we compare three variants: (1) **w/o structured data**: excludes the retrieval of knowledge graph triples for generating candidate cell types; (2) **w/o unstructured data**: excludes the retrieval of similar cell sentences and similar cell types; (3) **w/o marker genes**: instead of using marker genes to refine the results, generating only one candidate cell type as the final answer. Results show that removing structured data, unstructured data, or marker genes leads to drops in average performance of 3.28%, 3.95%, and 3.13%, respectively, highlighting their contributions. The full scRAG framework achieves the best results, demonstrating the value of integrating hybrid Retrieval-Augmented Generation and refinement for cross-tissue cell annotations.

4.4 Case Study

We design a case study where GPT-3.5-Turbo struggles to generate the correct cell type due to insufficient gene and tissue knowledge, given the top 100 genes and tissue. Even with CoT prompting for step-by-step reasoning, it falls into hallucinations, leading to flawed reasoning and incorrect answers. As shown in Figure 5, the scRAG framework addresses this issue by extracting gene and tissue entities, retrieving knowledge graph triples, and integrating this structured information. It also queries the reference cell database to retrieve top-k similar cell types, generating candidate cell types (e.g., delta and alpha). To ensure accuracy, the framework retrieves marker genes from both similar cells and candidate cells, enabling GPT-3.5-Turbo to refine its reasoning and confidently identify the correct answer "delta". In Appendix E, we give more details about this case study.

5 Conclusion

In this paper, we introduced scRAG, a framework designed to enhance large language models (LLMs) in cross-tissue single cell annotation tasks using RAG. By leveraging the hybrid Retrieval-Augmented Generation method and refining predictions with marker genes, scRAG provides more reliable and accurate cell annotation responses. Notably, scRAG's methodology aligns with expert workflows, where LLMs retrieve external knowledge bases as prior information and validate preliminary answers using marker genes. Our evaluations reveal that scRAG outperforms GPT-3.5, GPT-4, and traditional cell annotations. These findings underscore the potential of scRAG to enhance the application of LLMs in single-cell analysis.

6 Limitations

While our study has yielded encouraging results, there are several limitations that need to be acknowledged. A primary concern is the reliance on external knowledge bases. The accuracy of the scRAG output is likely to be affected by any shortcomings in external databases. Additionally, although scRAG shows substantial improvement in cell annotation tasks, its effectiveness for gene-related tasks, such as gene interaction prediction and biomedical question-answering, requires further investigation. Third, the computational cost associated with large models for graph construction and retrieval can be considerable, which may limit its feasibility in certain real-world applications. Finally, this work is primarily concerned with general cross-tissue single-cell annotation, and the handling of rare cell types remains to be explored.

7 Ethical Statement

We have ensured that this research is conducted in an ethical and responsible manner. A brief summary of the ethical considerations is provided below.

Public Dataset. We ensure that all data sources were cited accurately and appropriately crediting the original authors.

Licensed API Usage. We strictly adhere to the terms and conditions of all licensed APIs used in this research. Our use of existing tools and the artifacts we create aligns with their intended purposes. The estimated computational cost of this research is approximately 300 US dollars.

Transparency. The code and datasets will be appropriately released to ensure the transparency and reproducibility of our work.

References

Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. 2016. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360.

Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. 2012. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995.

Yinghao Cao, Xiaoyue Wang, and Gongxin Peng. 2020a. Scsa: a cell type annotation tool for single-cell rna-seq data. *Frontiers in genetics*, 11:490.

Zhi-Jie Cao, Lin Wei, Shen Lu, De-Chang Yang, and Ge Gao. 2020b. Searching large-scale scrna-seq databases via unbiased cell embedding with cell blast. *Nature communications*, 11(1):3458.

Yiqun T Chen and James Zou. 2023. Genept: A simple but hard-to-beat foundation model for genes and cells built from chatgpt. *bioRxiv*, 2023–10.

Jan Salomon Cramer. 2002. The origins of logistic regression. Technical report, Tinbergen Institute discussion paper.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 2024. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 1–11.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

H Atakan Ekiz, Christopher J Conley, W Zac Stephens, and Ryan M O'Connell. 2020. Cipr: a web-based r/shiny app and r package to annotate cell clusters in single cell rna sequencing experiments. *BMC bioinformatics*, 21:1–15.

Yin Fang, Kangwei Liu, Ningyu Zhang, Xinle Deng, Penghui Yang, Zhuo Chen, Xiangru Tang, Mark Gerstein, Xiaohui Fan, and Huajun Chen. 2024. Chatcell: Facilitating single-cell analysis with natural language. *arXiv preprint arXiv:2402.08303*.

Felix Fischer, David S Fischer, Roman Mukhin, Andrey Isaev, Evan Biederstedt, Alexandra-Chloé Villani, and Fabian J Theis. 2024. sctab: Scaling cross-tissue single-cell annotation models. *Nature Communications*, 15(1):6611.

Oscar Franzén, Li-Ming Gan, and Johan LM Björkegren. 2019. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019:baz046.

Rui Fu, Austin E Gillen, Ryan M Sheridan, Chengzhe Tian, Michelle Daya, Yue Hao, Jay R Hesselberth, and Kent A Riemondy. 2020. clustifyr: an r package for automated single-cell rna sequencing cluster classification. *F1000Research*, 9.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Rui Hou, Elena Denisenko, and Alistair RR Forrest. 2019. scmatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics*, 35(22):4688–4695.

Wenpin Hou and Zhicheng Ji. 2024. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, 1–4.

Qianhui Huang, Yu Liu, Yuheng Du, and Lana X Garmire. 2021. Evaluation of cell type annotation r packages on single-cell rna-seq data. *Genomics, Proteomics & Bioinformatics*, 19(2):267–281.

Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. 2018. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359–362.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Chenwei Li, Baolin Liu, Boxi Kang, Zedao Liu, Yedan Liu, Changya Chen, Xianwen Ren, and Zemin Zhang. 2020. Scibet as a portable and fast single cell type identifier. *Nature communications*, 11(1):1818.

Xinyi Lin, Gelei Deng, Yuekang Li, Jingquan Ge, Joshua Wing Kei Ho, and Yi Liu. 2024. Generag: Enhancing large language models with gene-related task by retrieval-augmented generation. *bioRxiv*, 2024–06.

Yingxin Lin, Yue Cao, Hani Jieun Kim, Agus Salim, Terence P Speed, David M Lin, Pengyi Yang, and Jean Yee Hwa Yang. 2020. scclassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Molecular systems biology*, 16(6):e9389.

Monika Litviňuková, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Catherine L Worth, Eric L Lindberg, Masatoshi Kanda, Krzysztof Polanski, Matthias Heinig, Michael Lee, et al. 2020. Cells of the adult human heart. *Nature*, 588(7838):466–472.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. 2025. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

Feiyang Ma and Matteo Pellegrini. 2019. Actinn: automated identification of cell types in single cell rna sequencing. *Bioinformatics*, 36(2):533–538.

Feiyang Ma and Matteo Pellegrini. 2020. Actinn: automated identification of cell types in single cell rna sequencing. *Bioinformatics*, 36(2):533–538.

Vasilios Mavroudis. 2024. Langchain.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Karolyn A Oetjen, Katherine E Lindblad, Meghali Goswami, Gege Gui, Pradeep K Dagur, Catherine Lai, Laura W Dillon, J Philip McCoy, and Christopher S Hourigan. 2018. Human bone marrow assessment by single-cell rna sequencing, mass cytometry, and flow cytometry. *JCI insight*, 3(23).

OpenAI. 2023. OpenAI GPT-4 Model Overview. `https://platform.openai.com/docs/models/overview`.

Leif E Peterson. 2009. K-nearest neighbor. *Scholarpedia*, 4(2):1883.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. 2017. Science forum: The human cell atlas. *elife*, 6:e27041.

Sarah Sandmann, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils, and Julian Varghese. 2025. Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, pages 1–1.

Angela Serra, Pietro Coretto, Michele Fratello, and Roberto Tagliaferri. 2018. Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data. *Bioinformatics*, 34(4):625–634.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. 2024. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9):btae560.

Yuqi Tan and Patrick Cahan. 2019. Singlecellnet: a computational tool to classify single cell rna-seq data across platforms and across species. *Cell systems*, 9(2):207–213.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, et al. 2024. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *bioRxiv*, 2024–05.

Yang Xu, Simon J Baumgart, Christian M Stegmann, and Sikander Hayat. 2022. Maca: marker-based automatic cell-type annotation for single-cell expression data. *Bioinformatics*, 38(6):1756–1760.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. 2022. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. 2024. Chemllm: A chemical large language model. *Preprint*, arXiv:2402.06852.

Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, et al. 2019. Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research*, 47(D1):D721–D728.

A Description of Dataset

Dataset1. The dataset utilized droplet-based single-cell RNA sequencing (scRNA-Seq), flow cytometry, and mass cytometry to comprehensively assess human bone marrow mononuclear cells (BMMCs) from 20 healthy donors across a wide age range (24–84 years). It includes over 76,000 single cells analyzed, with a mean of 880 genes detected per cell at a sequencing depth of approximately 50,000 reads per cell. All major bone marrow cell populations were identified, including hematopoietic stem/progenitor cells, lymphoid cells, and myeloid lineages, using dimensionality reduction techniques such as t-SNE and UMAP. The dataset is publicly available in the Gene Expression Omnibus under accession numbers GSE120221 and GSE120446 (Oetjen et al., 2018).

Dataset2. The dataset includes transcriptomes of more than 12,000 individual pancreatic cells, encompassing 8,629 human cells from four cadaveric donors as well as 1,886 mouse cells from two strains. The analysis identified 15 major cell clusters corresponding to known cell types, such as alpha, beta, gamma, delta, and epsilon endocrine cells, as well as acinar, ductal, stellate, vascular, Schwann, and immune cells. Substructure within ductal and beta cell populations was observed, revealing novel functional insights into these cell types. t-SNE visualization highlighted clear cell type-specific clusters, and rigorous deconvolution methods facilitated the comparison of bulk RNA-seq datasets with single-cell resolution. The study advances the understanding of pancreatic cell heterogeneity and provides a valuable resource for disease research and therapeutic development. The dataset is publicly available under *NCBI GEO*: GSE84133 (Baron et al., 2016).

Dataset3. The dataset employed single-cell and single-nucleus RNA sequencing to analyze the cellular composition of six anatomical regions of the adult human heart, encompassing left and right atria, left and right ventricles, the interventricular septum, and the left ventricular apex. It includes 487,106 cells and nuclei, representing 11 major cardiac cell types, such as cardiomyocytes, fibroblasts, endothelial cells, immune cells, and adipocytes, with high-resolution subcluster analysis. This dataset is publicly available at `www.heartcellatlas.org` (Litviňuková et al., 2020).

B Competitor Method

We compare our scRAG framework with baseline models and methods for cross-tissue cell annotation. Additionally, we evaluate individual baseline models, including GPT-3.5-Turbo, GPT-4o-Mini, and GPT-4-Turbo. The details of some of these models are elaborated below.

CoT (Kojima et al., 2023): This approach concatenates a trigger sentence, "Let's think step by step," to the test question.

GenePT (Chen and Zou, 2023): This approach leverages large language model (LLM) embeddings to represent genes and cells, derived from NCBI text descriptions. It generates gene embeddings using GPT-3.5 on textual gene summaries and creates single-cell embeddings either by averaging gene embeddings weighted by expression levels or by forming sentence embeddings from gene names ordered by expression.

ChatCell (Fang et al., 2024): This approach streamlines single-cell analysis through natural language interaction. It leverages vocabulary adaptation and unified sequence generation, empowering large language models with expertise in single-cell biology. By converting scRNA-seq data into a cell sentence compatible with LLMs, ChatCell enables diverse tasks, including random cell generation, pseudo-cell generation, cell type annotation, and drug sensitivity prediction.

Cell2Sentence (Levine et al., 2023): In this approach, each cell's expression profile is reformulated as a ranked list of gene names, generating a structured textual representation termed a "cell sentence". It allows the fine-tuning of causal language models like GPT-2 on cell sentences, enabling the generation of biologically valid cells from prompts and the accurate prediction of cell types from cell sentence inputs.

C Discussions on Prompt Formats

Figure 6 illustrates the accuracy comparison between two different prompt formats: **Triples** and **Sentences**. The Triples format presents the retrieved knowledge as concise entity relationships

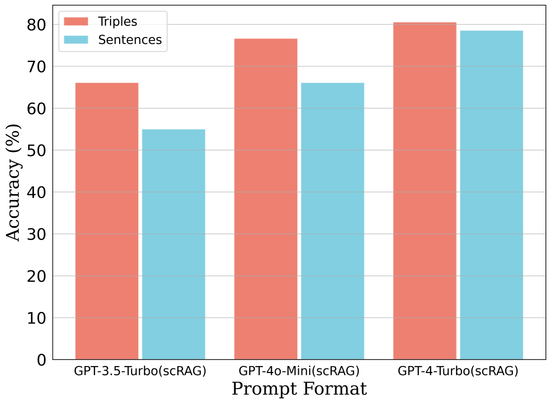


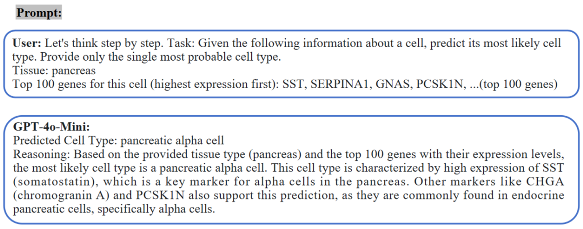
Figure 6: Comparison of different prompt formats on structured data.

(e.g., "X cell type -`LOCATED_IN` Y tissue, Z gene -`EXPRESSED_IN` X cell type"), while the Sentences format transforms these relationships into more verbose natural language statements (e.g., "The X cell type is located in the Y tissue, and Z gene is expressed in X cell type"). Across LLMs—GPT-3.5-Turbo, GPT-4-Mini, and GPT-4-Turbo, the Triples format consistently outperforms the Sentences format.

One possible explanation for this observation is that the "Triples" format is more concise and structured, which allows the language model to focus more effectively on the essential relationships without being distracted by additional linguistic complexity. In contrast, the "Sentences" format, while being more natural, introduces redundancy and potentially increases the cognitive load on the model, resulting in slightly lower accuracy.

D Discussions on Misclassifications

This case study further investigates the model's misclassifications behavior in a cross-tissue setting and analyzes potential causes of the errors. The input prompt and the prediction generated by GPT-4o-Mini are shown in Figure 7. However, the correct answer should be delta cell.



Prompt:

User: Let's think step by step. Task: Given the following information about a cell, predict its most likely cell type. Provide only the single most probable cell type.
Tissue: pancreas
Top 100 genes for this cell (highest expression first): SST, SERPINA1, GNAS, PCSK1N, ...(top 100 genes)

GPT-4o-Mini:
Predicted Cell Type: pancreatic alpha cell
Reasoning: Based on the provided tissue type (pancreas) and the top 100 genes with their expression levels, the most likely cell type is a pancreatic alpha cell. This cell type is characterized by high expression of SST (somatostatin), which is a key marker for alpha cells in the pancreas. Other markers like CHGA (chromogranin A) and PCSK1N also support this prediction, as they are commonly found in endocrine pancreatic cells, specifically alpha cells.

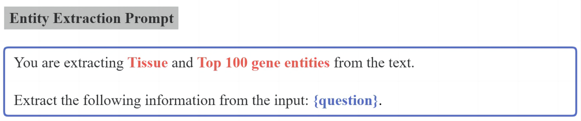
Figure 7: In-depth Analysis of Misclassification by GPT-4o-Mini.

The following analyzes the possible reasons behind this misclassification. (1) *Model Training Reasons*: Overreliance on a Few Specific Markers: If the model has been trained on a dataset where certain markers (e.g., SST for alpha cells) are very strong and well-defined, it may overfit these markers. This overfitting could make the model less robust in cases where these markers are not as dominant, leading to misclassifications in less well-defined cases. (2) *Single Tissue Context*: Many genes, such as SST, CHGA, and GNAS, are expressed across multiple cell types within the same tissue. The presence of these genes in both alpha and delta cells increases the likelihood of misclassification, as the model may struggle to distinguish between these closely related cell types. (3) *Cross-Tissue Context*: In different tissues, similar gene expression patterns can be observed across various cell types, further complicating the classification process. For example, genes like CHGA and PCSK1N, which are typically found in endocrine pancreatic cells, may also appear in other tissues, leading to incorrect associations. The model might not have sufficient contextual knowledge to differentiate between gene expression patterns that are cell-type-specific and tissue-specific. This cross-tissue similarity can create ambiguity, making it difficult to accurately predict the most probable cell type without additional contextual information.

E Examples and Prompts

We construct reference cell sentences to build a reference cell database and generate knowledge graphs, as shown in Figure 9. Testing cell sentences form a testing cell database to evaluate the model's cross-tissue cell annotation capability, illustrated in Figure 10.

Figures 8, 11, and 12 show the prompts used for entity extraction, candidate cell type generation, and refinement with marker genes, respectively. Examples of entity extraction, unstructured retrieval, and structured retrieval are provided in Figures 13, 14, and 15. Figure 16 further demonstrates the final process of candidate cell type generation and refinement with marker genes.



Entity Extraction Prompt

You are extracting **Tissue** and **Top 100 gene entities** from the text.

Extract the following information from the input: **{question}**.

Figure 8: Prompt for entity extraction.

Example cell sentence from reference cell database

Top 100 genes for this cell (highest expression first):MALAT1, B2M, MT-ATP6, TMSB10, IFITM3, MT-CO3, MT2A, MT-CO1, NFKBIA, HLA-C, MT-ND4, FABP5, FABP4, IL6, HSPA1A, MT-CO2, TMSB4X, EEF1A1, CAV1, MT-CYB, CXCL2, JUNB, CEBPD, HLA-A, HLA-B, CD59, SPARCL1, IFI27, H3-3B, STOM, TM4SF1, CALM1, LGALS1, MT-ND3, MT-ND1, MT-ND2, GNG11, RHOB, SRSF3, RPL32, VWF, ACTB, RPL29, ZFP36, PDLIM1, RPL10, HYAL2, FKBP1A, MT1E, SLC9A3R2, ITM2B, S100A11, MT1M, RNASE1, RPL11, MYL6, TAGLN2, CD9, IFITM2, ETS2, TIMP3, HSP90AB1, BST2, RPS2, ID1, TXNIP, FTH1, WARS1, PTMA, ICAM2, EIF1, KLF4, RPL13, SRGN, RPS8, EGR1, CAV2, RPL28, RPS28, TUBA1B, RPS27A, RPL9, MGLL, NBL1, TPM1, HLA-DRA, RPL39, SOX18, KLF2, POMP, RPL12, FOS, NACA, CXCL1, GADD45B, PPP1R15A, RPS6, ID3, TPT1, RPS9.
Cell type:capillary endothelial cell.
Tissue:apex of heart.

Figure 9: Reference cell sentence format.

Example cell sentence from testing cell database

Top 100 genes for this cell (highest expression first):SST, SERPINA1, GNAS, PCSK1N, RBP4, CHGA, RPL3, ACTG1, EEF1A1, TPT1, RPL19, CHGB, HLA-A, HSPA1A, CPE, RPL41, SCG5, EDN3, RPS4X, RPL8, RPL37A, TUBA1B, DYNLL1, RPL7A, GAD2, RPS8, RPL27A, RPS11, B2M, TIMP1, PTPRN, RPS2, RPL15, CD63, RPS15, TTR, RPL13A, SCG2, AQP3, IDS, PCSK2, RPS3A, RPL23A, GPX3, RPL10, TUBA1A, FOS, H3F3A, SEC11C, SERF2, RPS27A, EMC10, SCGN, RPS12, GAPDH, H3F3B, TAGLN2, NLRP1, RPL13, RPL14, PEG10, RPS14, RPS9, RPL24, ZFP36, RPS24, JUNB, RPS23, RPS28, EIF1, FAU, RPL11, FTH1, CLU, ATP5E, CALY, TMSB4X, RPL18, RPS29, RPL35A, FTL, PSAP, ENO1, RPL23, RPS18, DHRS2, RPLP2, RPS19, S100A6, MIF, RPLP1, HSP90AA1, RNASEK, CHCHD2, SSR4, RPL6, RPL28, HSPA5, HINT1, MALAT1.
Tissue:pancreas.

Figure 10: Testing cell sentence format.

Candidate Cell Types Generation Prompt

You are an assistant for a cross-tissue cell annotation task. Answer the question based only on the following context:
Structured data: {triples}
Unstructured data: {similar_cell_sentences}

Question: Given the following cell information, predict two candidate cell types. Provide only the cell types, without any additional explanation:
Tissue: {tissue}
Top 100 genes (highest expression first): {top_100_genes}

Use natural language and be concise.
Answer:

Figure 11: Prompt for candidate cell types generation.

Refinement with Marker Genes Prompt

Given the following information about a cell:
Top 100 genes: {top100_genes}.
Candidate cell types and their marker genes: {predicted_markers}.
Similar cell types retrieved and their marker genes: {retrieved_markers}.

Task: Given the following information about a cell, predict its most likely cell type. Provide only the single most probable cell type without any additional explanation.
Tissue: {tissue}
Top 100 genes (highest expression first): {top_100_genes}

From the following cell types, select the most probable:{cell_types}.

Figure 12: Prompt for refinement with marker genes.

Entity Extraction

Input:
Task: Given the following information about a cell, predict its most likely cell type. Provide only the single most probable cell type without any additional explanation.
Tissue: pancreas.
Top 100 genes for this cell (highest expression first): SST, SERPINA1, GNAS, PCSK1N, RBP4, CHGA, RPL3, ACTG1, EEF1A1, TPT1, RPL19, CHGB, HLA-A, HSPA1A, CPE, RPL41, SCG5, EDN3, RPS4X, RPL8, RPL37A, TUBA1B, DYNLL1, RPL7A, GAD2, RPS8, RPL27A, RPS11, B2M, TIMP1, PTPRN, RPS2, RPL15, CD63, RPS15, TTR, RPL13A, SCG2, AQP3, IDS, PCSK2, RPS3A, RPL23A, GPX3, RPL10, TUBA1A, FOS, H3F3A, SEC11C, SERF2, RPS27A, EMC10, SCGN, RPS12, GAPDH, H3F3B, TAGLN2, NLRP1, RPL13, RPL14, PEG10, RPS14, RPS9, RPL24, ZFP36, RPS24, JUNB, RPS23, RPS28, EIF1, FAU, RPL11, FTH1, CLU, ATP5E, CALY, TMSB4X, RPL18, RPS29, RPL35A, FTL, PSAP, ENO1, RPL23, RPS18, DHRS2, RPLP2, RPS19, S100A6, MIF, RPLP1, HSP90AA1, RNASEK, CHCHD2, SSR4, RPL6, RPL28, HSPA5, HINT1, MALAT1.
Output: pancreas, SST, SERPINA1, GNAS, PCSK1N, RBP4, CHGA, RPL3, ACTG1, EEF1A1, TPT1, RPL19, CHGB, HLA-A, HSPA1A, CPE, RPL41, SCG5, EDN3, RPS4X, RPL8, RPL37A, TUBA1B, DYNLL1, RPL7A, GAD2, RPS8, RPL27A, RPS11, B2M, TIMP1, PTPRN, RPS2, RPL15, CD63, RPS15, TTR, RPL13A, SCG2, AQP3, IDS, PCSK2, RPS3A, RPL23A, GPX3, RPL10, TUBA1A, FOS, H3F3A, SEC11C, SERF2, RPS27A, EMC10, SCGN, RPS12, GAPDH, H3F3B, TAGLN2, NLRP1, RPL13, RPL14, PEG10, RPS14, RPS9, RPL24, ZFP36, RPS24, JUNB, RPS23, RPS28, EIF1, FAU, RPL11, FTH1, CLU, ATP5E, CALY, TMSB4X, RPL18, RPS29, RPL35A, FTL, PSAP, ENO1, RPL23, RPS18, DHRS2, RPLP2, RPS19, S100A6, MIF, RPLP1, HSP90AA1, RNASEK, CHCHD2, SSR4, RPL6, RPL28, HSPA5, HINT1, MALAT1

Figure 13: Example for entity extraction.

Unstructured Retriever

Input:
Task: Given the following information about a cell, predict its most likely cell type. Provide only the single most probable cell type without any additional explanation.
Tissue: pancreas.
Top 100 genes for this cell (highest expression first): SST, SERPINA1, GNAS, PCSK1N, RBP4, CHGA, RPL3, ACTG1, EEF1A1, TPT1, RPL19, CHGB, HLA-A, HSPA1A, CPE, RPL41, SCG5, EDN3, RPS4X, RPL8, RPL37A, TUBA1B, DYNLL1, RPL7A, GAD2, RPS8, RPL27A, RPS11, B2M, TIMP1, PTPRN, RPS2, RPL15, CD63, RPS15, TTR, RPL13A, SCG2, AQP3, IDS, PCSK2, RPS3A, RPL23A, GPX3, RPL10, TUBA1A, FOS, H3F3A, SEC11C, SERF2, RPS27A, EMC10, SCGN, RPS12, GAPDH, H3F3B, TAGLN2, NLRP1, RPL13, RPL14, PEG10, RPS14, RPS9, RPL24, ZFP36, RPS24, JUNB, RPS23, RPS28, EIF1, FAU, RPL11, FTH1, CLU, ATP5E, CALY, TMSB4X, RPL18, RPS29, RPL35A, FTL, PSAP, ENO1, RPL23, RPS18, DHRS2, RPLP2, RPS19, S100A6, MIF, RPLP1, HSP90AA1, RNASEK, CHCHD2, SSR4, RPL6, RPL28, HSPA5, HINT1, MALAT1.
Output: text: Top 100 genes for this cell (highest expression first): SST, RBP4, SERPINA1, GNAS, CHGA, CPE, EEF1A1, RPL27A, RPL3, B2M, PCSK1N, NLRP1, RPL7A, CHGB, CD63, GAPDH, PCSK1, SCG5, GPX3, HSP90AB1, RPL37A, RPL15, TUBA1A, PTPRN, HLA-A, RPL11, RPL41, RPS15, RPL8, RPL13, RPS28, RPS4X, RPLP0, RPL19, RPS8, RPS2, H3F3A, RPL14, ABCC8, PDIA3, TUBA1B, RPLP1, SEC11C, HLA-B, RPS11, RPL6, UBB, RGS2, IDS, RPLP2, RPS27A, TIMP1, RPL17, RPS29, RPL26, IGFBP5, RPL39, RPS9, EIF1, CST3, RPL18A, TPT1, RPS3, ECEL1, RPS20, SERF2, RPS24, CALM2, MYL6, TMEM59, MIF, MALAT1, RNASEK, RPS18, RPL23A, RPL36, RPL18, FOS, RPL32, RPL23, GPX4, DHRS2, PKM, S100A6, RPS17, EDN3, FTH1, CHCHD2, HSPA1A, RPL10A, DSP, PARK7, RPL10, UBC, RPS3A, RPS7, RPS19, PCSK2, RPS5, GCH1. Cell type: delta. Tissue: pancreas.#Document
text: Top 100 genes for this cell (highest expression first): SST, GNAS, B2M, RBP4, MALAT1, EEF1A1, SERPINA1, MIF, GAPDH, CHGB, CPE, HLA-A, RPL37A, CHGA, SCG5, TTR, PCSK1N, PTPRN, RPS27A, RPS29, RPLP1, TMSB4X, RPL7A, RPL8, TUBA1B, HLA-B, ACTB, FTH1, SERF2, RPS15, RPL3, RPS28, NLRP1, RPL14, RPL38, RPS14, RPS3A, RPL11, IDS, TPT1, RPL15, RPLP2, RPS11, RPS4X, H3F3A, SCG2, RPS8, ACTG1, RPL13A, CD63, RPS2, RPL10, SCARB2, PEG10, RPL23A, HSPA1A, RPS9, TMBIM6, RPL13, RPL41, RPL27A, DYNLL1, PDIA3, RPL32, S100A6, RPL6, DHRS2, MYL6, RPS25, RPL7, RPS18, TUBB4B, FTL, RPS6, CALY, EMC10, RPL19, CCNI, RHOA, RPL23, KHDRBS1, ATP5E, TPI1, RPL21, RPSA, RPS24, HSP90AA1, PCP4, HSPA5, PCSK2, RPS5, EIF1, ARFGEF3, CALM2, RPS19, PALLD, RPS13, CANX, RPS20, SARAF. Cell type: delta. Tissue: pancreas.#Document
text: Top 100 genes for this cell (highest expression first): SST, SERPINA1, GNAS, CHGA, EEF1A1, PCSK1N, RPL37A, RPL7A, RPS29, PEG10, NLRP1, RPS28, RPS11, RPL15, RPL41, RPS19, RPS14, RPL8, RPL10A, HLA-A, HSPA1A, TIMP1, GPX3, HES6, RPL32, RPL26, GCG, RPL23, RPL31, RPL27A, KRT8, TPT1, JUNB, RPLP1, RPLP0, FAM102A, RPS2, S100A6, UBB, GAPDH, CPE, RPL13, RPL14, RPL13A, MIF, RPL30, SCG5, RPS4X, CHGB, TIMP2, RPS27, FXYD6, RPL6, TMED3, MYL6, RPLP2, RPL37, RPL3, FOS, RPL12, UBA52, UBC, RPS18, SCG2, RPS23, RPS3, ZFP36, RPS27A, RHBDD2, HSP90AB1, MCHR1, CRIP2, RPS9, IDS, RPL35, PSAP, C17orf89, SNRPN, RPL4, RPS24, RPL10, CDKN1C, DAP, RPS20, RPL18, EIF1, PCBP1, EDN3, HSPA8, GPX1, BAIAP3, TTR, CD81, CD99, PEBP1, RBP4, ACTB, SLC22A17, TRMT112, ATP5G2. Cell type: delta. Tissue: pancreas.#Document
text: Top 100 genes for this cell (highest expression first): SST, CHGB, PCSK1N, EEF1A1, CPE, GNAS, SERPINA1, HLA-A, INS, RPS23, RPS18, RPL7A, RPS4X, RPL3, PCSK1, RPL37A, RPL8, RPLP1, CD63, RPL27A, TMSB4X, RPL11, RPS2, ATP5E, RPL41, ID2, CHGA, RPL15, CALM2, RPS27A, RPL4, SCGN, SEC11C, RPS13, RPL26, RPL35A, RPL34, RPL13A, RPL5, EMC10, PSAP, H3F3A, RPS29, SCG5, RPS12, RPS3, IDS, RBP4, RPS7, B2M, RPS11, PTPRN, RPL30, FTH1, RPS25, MAP1B, RPL10, MIF, RPL38, SPINT2, RBP1, SSR4, RPL22, UBB, SCG2, PARK7, MALAT1, GABARAP, RPLP2, EIF3K, ERP29, TAGLN2, HLA-B, DHRS2, CALM1, EIF4B, RPS8, HSPA1A, CRIP2, S100A6, C10orf10, RPS9, RPL13, GCG, TMEM176A, TMBIM6, RPL23, TMEM205, SNRPN, RPS28, RPL9, DAP, RPS17, LPPR2, RPS24, ATP5L, CAMK2N1, RPS21, YBX1, RHOA. Cell type: delta. Tissue: pancreas.

Figure 14: Example for unstructured retrieval.

Structured Retriever

Input: pancreas, SST, SERPINA1, GNAS, PCSK1N, RBP4, CHGA, RPL3, ACTG1, EEF1A1, TPT1, RPL19, CHGB, HLA-A, HSPA1A, CPE, RPL41, SCG5, EDN3, RPS4X, RPL8, RPL37A, TUBA1B, DYNLL1, RPL7A, GAD2, RPS8, RPL27A, RPS11, B2M, TIMP1, PTPRN, RPS2, RPL15, CD63, RPS15, TTR, RPL13A, SCG2, AQP3, IDS, PCSK2, RPS3A, RPL23A, GPX3, RPL10, TUBA1A, FOS, H3F3A, SEC11C, SERF2, RPS27A, EMC10, SCGN, RPS12, GAPDH, H3F3B, TAGLN2, NLRP1, RPL13, RPL14, PEG10, RPS14, RPS9, RPL24, ZFP36, RPS24, JUNB, RPS23, RPS28, EIF1, FAU, RPL11, FTH1, CLU, ATP5E, CALY, TMSB4X, RPL18, RPS29, RPL35A, FTL, PSAP, ENO1, RPL23, RPS18, DHRS2, RPLP2, RPS19, S100A6, MIF, RPLP1, HSP90AA1, RNASEK, CHCHD2, SSR4, RPL6, RPL28, HSPA5, HINT1, MALAT1

Output: Endothelial, Acinar, Ductal, Beta, Alpha, Gamma, Delta - LOCATED_IN -> Pancreas

Sst - EXPRESSED_IN -> Endothelial, Acinar, Ductal, Beta, Alpha, Gamma, Delta

Serpina1 - EXPRESSED_IN -> Acinar, Ductal, Cd14+ Monocytes, Gamma, Alpha, Delta

Gnas - EXPRESSED_IN -> Delta, Endothelial, Vein Endothelial Cell, Ductal, Capillary Endothelial Cell, Beta, Endothelial Cell Of Artery, Alpha, Gamma, Acinar, Endothelial Cell

Pcsk1N - EXPRESSED_IN -> Endothelial, Ductal, Beta, Gamma, Alpha, Delta

Rbp4 - EXPRESSED_IN -> Delta, Beta, Ductal

Chga - EXPRESSED_IN -> Ductal, Beta, Gamma, Alpha, Delta

Rpl3 - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Alpha, Acinar, Delta, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, B Cell, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Ductal, Natural Killer Cell, Fibroblast

Actg1 - EXPRESSED_IN -> Vein Endothelial Cell, Smooth Muscle Cell, Beta, Regular Ventricular Cardiac Myocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Alpha, Delta, Acinar, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, Ductal, Cd14-Positive, Cd16-Positive Monocyte, Capillary Endothelial Cell, Natural Killer Cell, Fibroblast

Eef1A1 - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Mesothelial Cell, Alpha, Acinar, Delta, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, B Cell, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Ductal, Natural Killer Cell, Fibroblast

Tpt1 - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Alpha, Delta, Acinar, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, Ductal, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Natural Killer Cell, Fibroblast

Rpl19 - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Alpha, Delta, Acinar, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, B Cell, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Ductal, Natural Killer Cell

Chgb - EXPRESSED_IN -> Ductal, Beta, Gamma, Alpha, Delta

Hla-A - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Beta, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Alpha, Delta, Acinar, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, Ductal, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Natural Killer Cell, Fibroblast

Hspa1A - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Regular Ventricular Cardiac Myocyte, Gamma, Endothelial Cell Of Artery, Endothelial, Native Cell, Cd10+ B Cells, Alpha, Delta, Acinar, Macrophage, Pericyte Cell, Hspcs, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, Ductal, Cd14-Positive, Cd16-Positive Monocyte, Capillary Endothelial Cell, Natural Killer Cell, Fibroblast

Cpe - EXPRESSED_IN -> Endothelial, Ductal, Beta, Pericyte Cell, Alpha, Gamma, Delta

Rpl41 - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Alpha, Acinar, Delta, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, B Cell, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Ductal, Natural Killer Cell, Fibroblast

Scg5 - EXPRESSED_IN -> Endothelial, Ductal, Beta, Gamma, Alpha, Delta

Edn3 - EXPRESSED_IN -> Delta, Beta, Gamma

Rps4X - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Alpha, Acinar, Delta, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, B Cell, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Ductal, Natural Killer Cell, Fibroblast

Rpl8 - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Alpha, Delta, Acinar, Macrophage, Pericyte Cell, Ductal, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Natural Killer Cell, Fibroblast

Rpl37A - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial, Native Cell, Alpha, Delta, Acinar, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, Ductal, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Natural Killer Cell, Fibroblast

Tuba1B - EXPRESSED_IN -> Erythrocytes, Vein Endothelial Cell, Smooth Muscle Cell, Cd4+ T Cells, Beta, Gamma, Cd14+ Monocytes, Endothelial Cell, Endothelial, Cd8+ T Cells, Native Cell, Cd10+ B Cells, Alpha, Delta, Acinar, Plasmacytoid Dendritic Cells, Nkt Cells, Pericyte Cell, Hspcs, Nk Cells, Monocyte-Derived Dendritic Cells, Monocyte Progenitors, Megakaryocyte Progenitors, Ductal, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Cd20+ B Cells, Plasma Cells, Erythroid Progenitors

Dynll1 - EXPRESSED_IN -> Erythrocytes, Smooth Muscle Cell, Cd4+ T Cells, Beta, Regular Ventricular Cardiac Myocyte, Gamma, Cd14+ Monocytes, Endothelial, Cd8+ T Cells, Cd10+ B Cells, Alpha, Delta, Plasmacytoid Dendritic Cells, Nkt Cells, Hspcs, Nk Cells, Monocyte-Derived Dendritic Cells, Monocyte Progenitors, Megakaryocyte Progenitors, Ductal, Capillary Endothelial Cell, Cd20+ B Cells, Plasma Cells, Erythroid Progenitors

Rpl7A - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Alpha, Delta, Acinar, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, B Cell, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Ductal, Natural Killer Cell, Fibroblast

Gad2 - EXPRESSED_IN -> Ductal, Beta, Gamma, Alpha, Delta

Rps8 - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Alpha, Delta, Acinar, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, Ductal, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Natural Killer Cell, Fibroblast

Rpl27A - EXPRESSED_IN -> Vein Endothelial Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial, Alpha, Delta, Acinar, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, B Cell, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Ductal, Natural Killer Cell, Fibroblast

Rps11 - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Alpha, Delta, Acinar, Macrophage, Pericyte Cell, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, Ductal, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Natural Killer Cell, Fibroblast

B2M - EXPRESSED_IN -> Vein Endothelial Cell, Cd4-Positive, Alpha-Beta Cytotoxic T Cell, Smooth Muscle Cell, Beta, Regular Ventricular Cardiac Myocyte, Cd14-Positive Monocyte, Gamma, Endothelial Cell Of Artery, Endothelial Cell, Endothelial, Native Cell, Mesothelial Cell, Alpha, Acinar, Delta, Macrophage, Pericyte Cell, Regular Atrial Cardiac Myocyte, Cd8-Positive, Alpha-Beta Cytotoxic T Cell, B Cell, Activated Cd8-Positive, Alpha-Beta T Cell, Capillary Endothelial Cell, Cd14-Positive, Cd16-Positive Monocyte, Ductal, Natural Killer Cell, Fibroblast

...

Figure 15: Example for structured retrieval.

Candidate Cell Types Generation

Input:
Answer the question based only on the following context:
Structured data: {structured_data}
Unstructured data: {unstructured_data}
Question: Given the following information about a cell, predict two candidate cell types. Provide only the cell types without additional explanation. Tissue: pancreas. Top 100 genes for this cell (highest expression first):SST, SERPINA1, GNAS, PCSK1N, RBP4, CHGA, RPL3, ACTG1, EEF1A1, TPT1, RPL19, CHGB, HLA-A, HSPA1A, CPE, RPL41, SCG5, EDN3, RPS4X, RPL8, RPL37A, TUBA1B, DYNLL1, RPL7A, GAD2, RPS8, RPL27A, RPS11, B2M, TIMP1, PTPRN, RPS2, RPL15, CD63, RPS15, TTR, RPL13A, SCG2, AQP3, IDS, PCSK2, RPS3A, RPL23A, GPX3, RPL10, TUBA1A, FOS, H3F3A, SEC11C, SERF2, RPS27A, EMC10, SCGN, RPS12, GAPDH, H3F3B, TAGLN2, NLRP1, RPL13, RPL14, PEG10, RPS14, RPS9, RPL24, ZFP36, RPS24, JUNB, RPS23, RPS28, EIF1, FAU, RPL11, FTH1, CLU, ATP5E, CALY, TMSB4X, RPL18, RPS29, RPL35A, FTL, PSAP, ENO1, RPL23, RPS18, DHRS2, RPLP2, RPS19, S100A6, MIF, RPLP1, HSP90AA1, RNASEK, CHCHD2, SSR4, RPL6, RPL28, HSPA5, HINT1, MALAT1.
Use natural language and be concise.
Answer:
Output: delta,alpha.

Refinement with Marker Genes

Input:
Given the following information about a cell:
Top 100 genes:SST, SERPINA1, GNAS, PCSK1N, RBP4, CHGA, RPL3, ACTG1, EEF1A1, TPT1, RPL19, CHGB, HLA-A, HSPA1A, CPE, RPL41, SCG5, EDN3, RPS4X, RPL8, RPL37A, TUBA1B, DYNLL1, RPL7A, GAD2, RPS8, RPL27A, RPS11, B2M, TIMP1, PTPRN, RPS2, RPL15, CD63, RPS15, TTR, RPL13A, SCG2, AQP3, IDS, PCSK2, RPS3A, RPL23A, GPX3, RPL10, TUBA1A, FOS, H3F3A, SEC11C, SERF2, RPS27A, EMC10, SCGN, RPS12, GAPDH, H3F3B, TAGLN2, NLRP1, RPL13, RPL14, PEG10, RPS14, RPS9, RPL24, ZFP36, RPS24, JUNB, RPS23, RPS28, EIF1, FAU, RPL11, FTH1, CLU, ATP5E, CALY, TMSB4X, RPL18, RPS29, RPL35A, FTL, PSAP, ENO1, RPL23, RPS18, DHRS2, RPLP2, RPS19, S100A6, MIF, RPLP1, HSP90AA1, RNASEK, CHCHD2, SSR4, RPL6, RPL28, HSPA5, HINT1, MALAT1.
Candidate cell types and their maker genes: delta:LEPR,RBP4,CBLN4,BCHE,SST,PCSK1,HHEX,EDN3,PCP4,UCP2,NGFR,GAP43,CALB1,DHRS2,PAPPA2,CDH19,GPX3,PIPOX,CALCB,TMOD1.
Similar cell types retrieved and their maker genes:
delta:LEPR,RBP4,CBLN4,BCHE,SST,PCSK1,HHEX,EDN3,PCP4,UCP2,NGFR,GAP43,CALB1,DHRS2,PAPPA2,CDH19,GPX3,PIPOX,CALCB,TMOD1;
alpha: IRX2,RC,PAPPA2,F10,PCSK2,RGS4,TM4SF4,MUC13,VSTM2L,FEV,CRYBA2,TMEM176A,GPX3,VGF,LOXL4,KCTD12,PCSK1N,FXYD5,CALY.

Task: Given the following information about a cell, predict its most likely cell type. Provide only the single most probable cell type without any additional explanation.
From the following cell types, select the most probable: 'fibroblast', 'activated CD4-positive, alpha-beta T cell', 'HSPCs', 'ductal', 'regular ventricular cardiac myocyte', 'vein endothelial cell', 'Erythrocytes', 'endothelial cell of artery', 'acinar', 'beta', 'B cell', 'natural killer cell', 'CD4-positive, alpha-beta cytotoxic T cell', 'epicardial adipocyte', 'regular atrial cardiac myocyte', 'macrophage', 'Plasmacytoid dendritic cells', 'gamma', 'native cell', 'smooth muscle cell', 'endothelial', 'CD20+ B cells', 'neural cell', 'Megakaryocyte progenitors', 'Plasma cells', 'endothelial cell', 'CD4+ T cells', 'CD10+ B cells', 'Monocyte-derived dendritic cells', 'CD14+ Monocytes', 'delta', 'Erythroid progenitors', 'activated CD8-positive, alpha-beta T cell', 'NK cells', 'mature NK T cell', 'alpha', 'CD8-positive, alpha-beta cytotoxic T cell', 'monocyte', 'pericyte cell', 'capillary endothelial cell', 'NKT cells', 'CD14-positive, CD16-positive monocyte', 'CD8+ T cells', 'Monocyte progenitors'.

Output: delta.

Figure 16: Candidate cell types generation and refinement with marker genes.