VISIT OUR WEBSITE

Automated Metadata Ingestion in Strand's scRNA-seq Portal

Shrutee Jakhanwal, Tasmia Kausar*, Rohan Karthikeyan*, Shardul Kamble*,
Karan Ladha*, Lavanya Nemani*, Radhakrishna Bettadapura, Ramesh Hariharan,
Badri Padhukasahasram; **Strand Life Sciences, Bangalore, India** **Contributed equally*

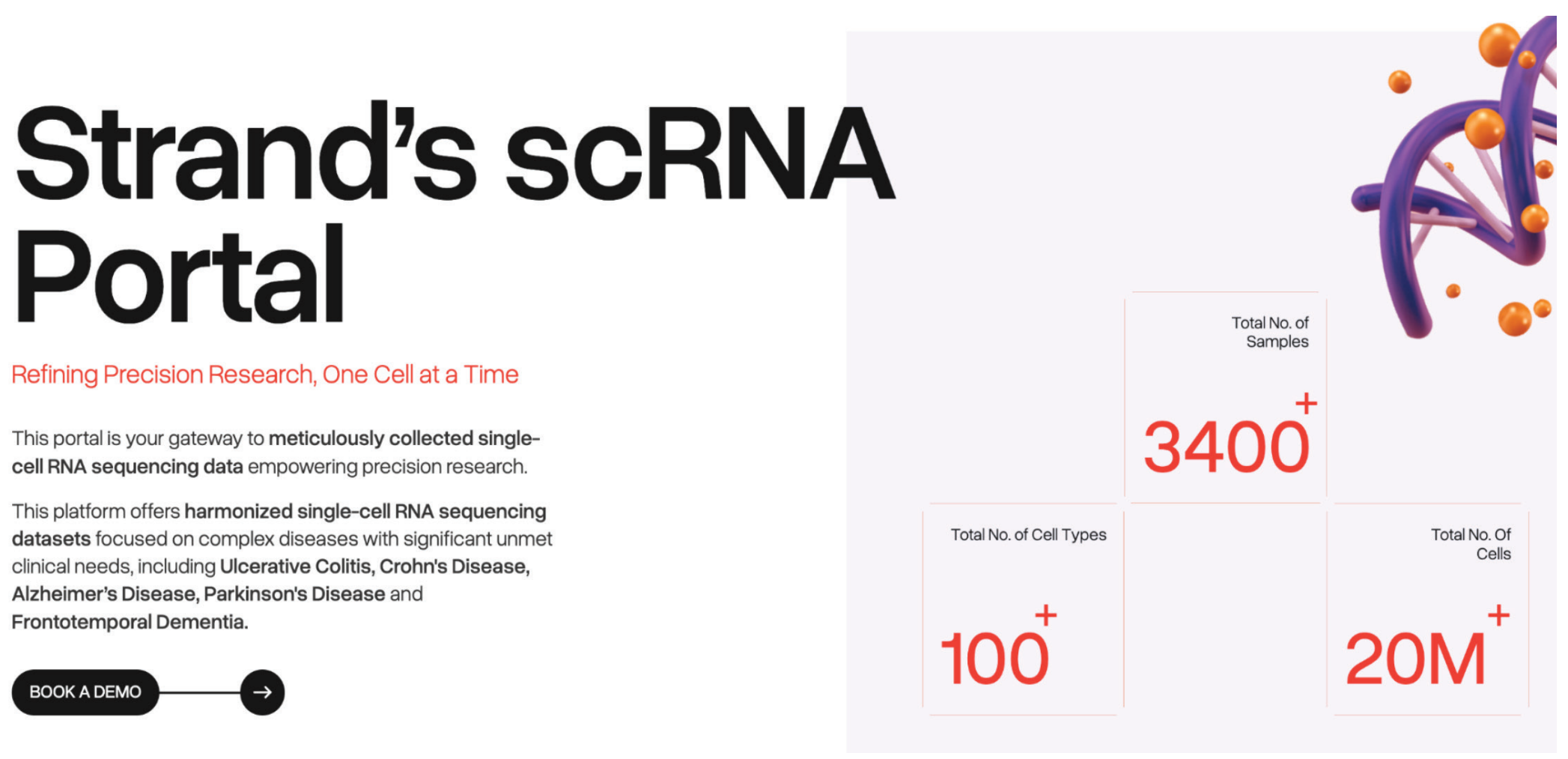
Contact

Radhakrishna (RK) Bettadapura
VP, Research Informatics

☎ +1 (415) 917-9605 ✉ rk@strandls.com

Introduction

- Harnessing and standardizing the metadata from public repositories (like NCBI's GEO) can help unlock its potential for a range of research questions.
- We have developed an automated system based on Retrieval Augmented Generation (RAG) and Large Language Models (LLM) to quickly and accurately standardize biomedical terms.
- Performance of this pipeline is demonstrated on manually curated ground truths from Strand's scRNA-seq portal.



Strand's scRNA Portal

Refining Precision Research, One Cell at a Time

This portal is your gateway to **meticulously collected single-cell RNA sequencing data** empowering precision research.

This platform offers **harmonized single-cell RNA sequencing datasets** focused on complex diseases with significant unmet clinical needs, including **Ulcerative Colitis, Crohn's Disease, Alzheimer's Disease, Parkinson's Disease** and **Frontotemporal Dementia.**

BOOK A DEMO

Total No. of Samples 3400+

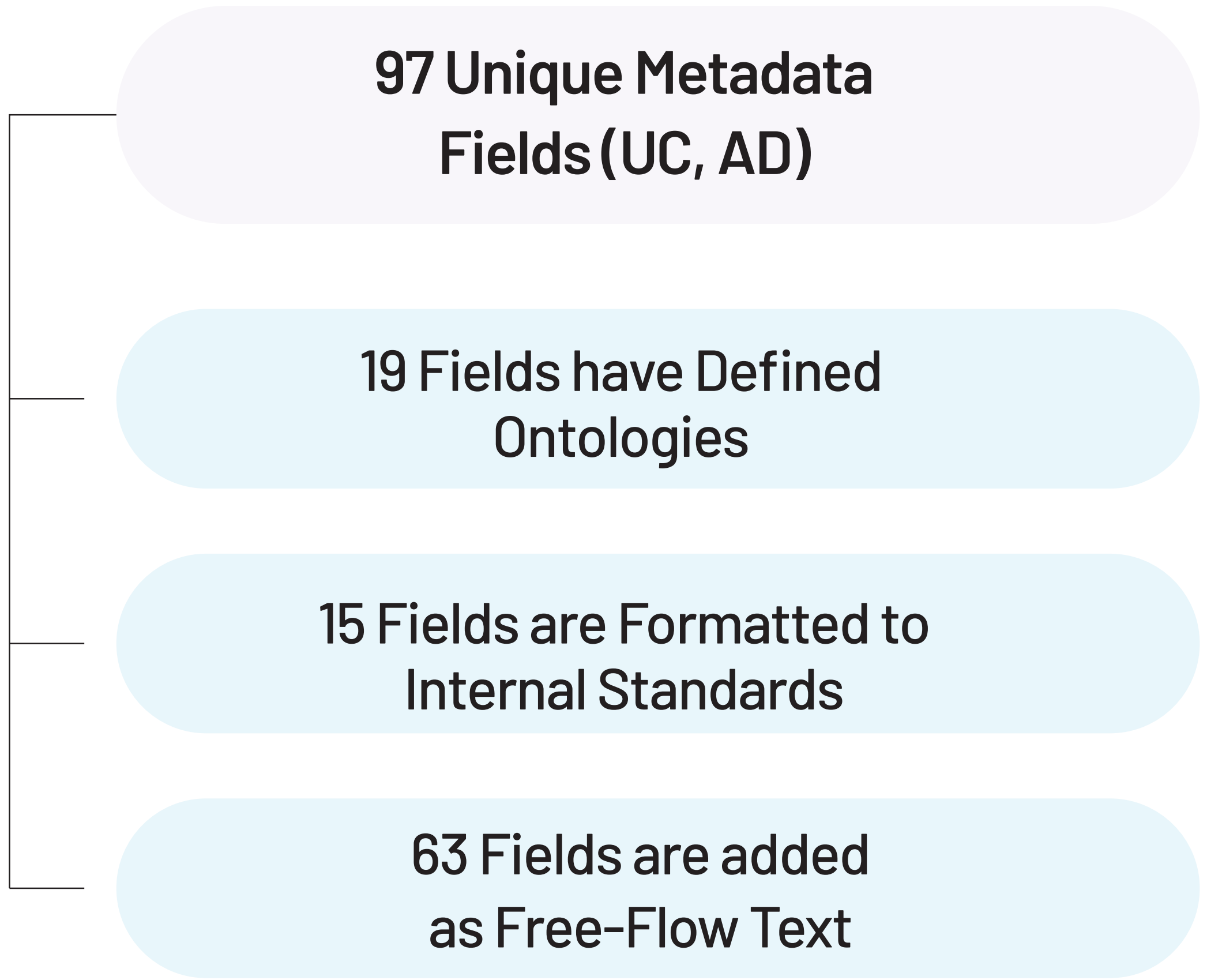
Total No. of Cell Types 100+

Total No. Of Cells 20M+

Key Features:

- 80+ Metadata fields
- 3 levels of curation
- 26 filters for easy navigation
- Search bar to explore "free-flow" texts

Metadata Curation: A Time-Intensive Process



97 Unique Metadata Fields (UC, AD)

19 Fields have Defined Ontologies

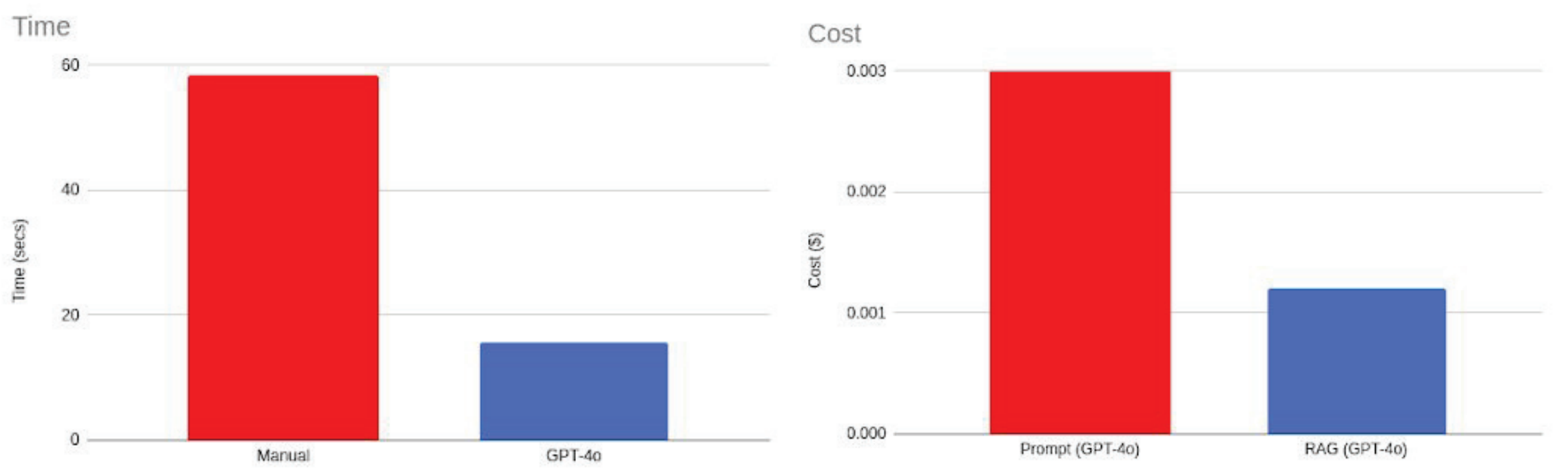
15 Fields are Formatted to Internal Standards

63 Fields are added as Free-Flow Text

Ontology mapping and formatting metadata fields to internal standards are among the most time-consuming and labor-intensive tasks. Utilizing methods like RAG and LLM can greatly improve efficiency and scalability.

Achievements and Highlights

- A high quality single-cell RNA-seq portal with comprehensive and normalized metadata with focus on Inflammatory Bowel Diseases.
- A rapid LLM and RAG pipeline for metadata ingestion of datasets compiled in this in-house portal.
- Accuracy is higher than prior RAG implementations due to the biomedical-specific embedding.
- ~ 3x reduction in turn around time through LLM automation compared to time and labor-intensive manual curation.
- RAG for biomedical normalization helps reduce candidates presented to GPT-4o, achieving ~ 3x reduction in number of tokens and cost compared to purely prompt based methods.



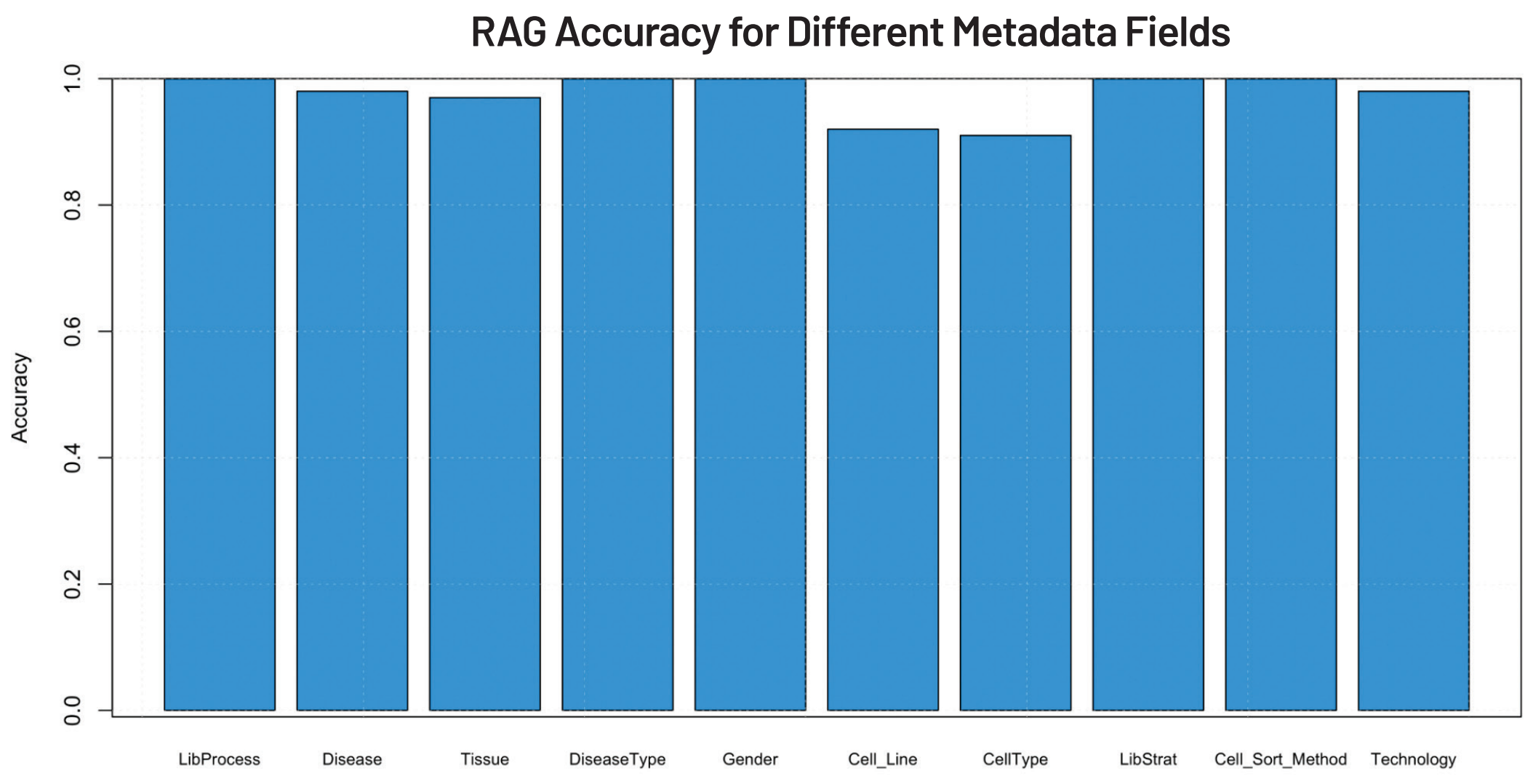
Semantic Embedding and RAG Optimization

Embedding model	Dim	Time taken (local) 2048 terms	Storage (MB) 2048 terms	Accuracy (%) 750 SNOMED CT TERMS
all-MiniLM-L6-v2	384	3.55 sec	3	78.09
bge-large-en-v1.5	1024	3:24 min	8	86.76
ember-v1	1024	3:16 min	8	86.88
GIST-large-Embedding-v0	1024	3:10 min	8	88.11
sf_model_e5	1024	4:08 min|	8	87.72

RAG Implementation Performs Similarly as Prior Methods

Dataset from Berkowitz et al 2024	RAG Berkowitz et al 2024	RAG Strand
106 SNOMED CT terms oncology specific	89.8%	89.81%
750 SNOMED CT terms cross domain	80.0%	84.11%
10 disease term Strand scRNA portal	93.58%	95.38%

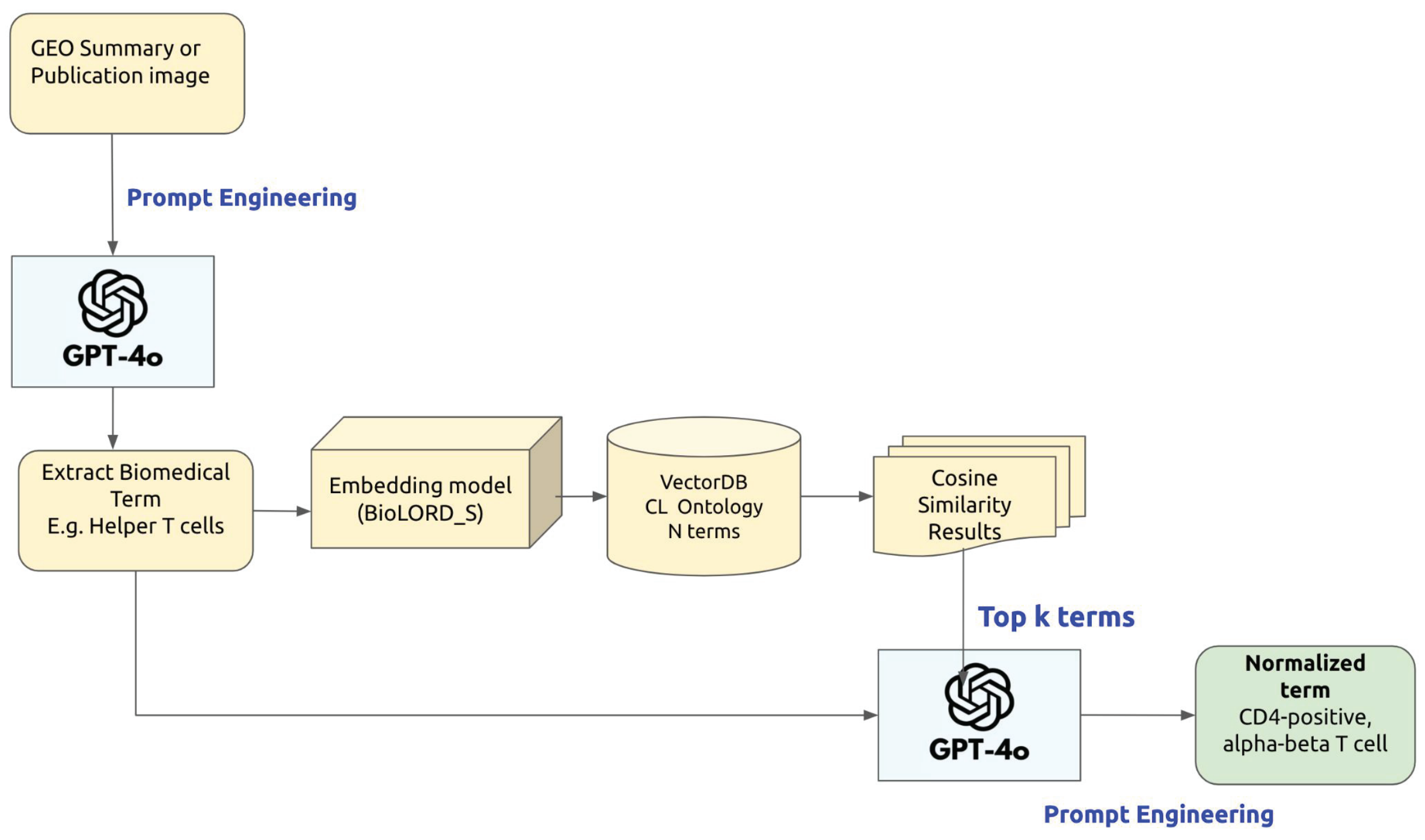
Performance for Metadata Fields in Strand Portal



Conclusion

- Strand's scRNA-seq portal hosts meticulously curated datasets with 80+ metadata fields, 3 levels of curation and 26 filters for easy navigation.
- Automation decreased turnaround time compared to manual curation alone while incurring a modest cost.
- LLM-based normalization was not entirely precise, some manual checks were unavoidable.
- Not all fields require the use of LLMs and a suitable combination of traditional methods as well as RAG-based LLMs can help scale up metadata ingestion for public data.

Summary of RAG Pipeline for Metadata Fields



GEO Summary or Publication image

Prompt Engineering

GPT-4o

Extract Biomedical Term E.g. Helper T cells

Embedding model (BioLORD_S)

VectorDB CL Ontology N terms

Cosine Similarity Results

Top k terms

GPT-4o

Normalized term CD4-positive, alpha-beta T cell

Prompt Engineering

Accuracy was enhanced through an optimal combination of prompt engineering for extracting metadata fields, choice of best semantic embeddings via empirical testing, and refining LLM prompts for choosing best normalized term for a query term.

Methods

RAG is a technique that enhances LLM responses by retrieving relevant information from an external knowledge base and providing that as context to the LLM. Our pipeline first performs entity recognition with GPT-4o, followed by RAG on the extracted entity using a biomedical terms specific embedding. Facebook's AI Similarity Search (FAISS) was used for creating vector store of standard terms and finding top k closest terms from cosine-similarity. Choice of k as well as embedding was guided by empirical testing.