Agentic AI for Exploring and Analyzing Omics Metadata at Scale

Aditya Goel, Shardul Kamble, Navneet Kumar, Lavanya Nemani, Rohan K, Prakash Hiremath, Shekhar Nath, Nihesh Rathod, Tasmia Kausar, Badri Padhukasahasram, Ramesh Hariharan, Radhakrishna Bettadapura, Shrutee Jakhanwal **Strand Life Sciences, Bangalore, India**

VISIT OUR WEBSITE

strand

Contact

Radhakrishna (RK) Bettadapura
VP, Research Informatics

+1 (415) 917-9605 rk@strandls.com

Introduction

Large, dense, and highly multi-dimensional biomedical datasets are increasingly common, making manual exploration slow and inefficient.

- Identifying consistent patterns across large datasets is not scalable through manual review and the data spans multiple formats: numeric, categorical, and free-text requiring complex programmatic queries .
- We present a chatbot-based interface powered by LLMs and Agentic AI to enable fast, intuitive, and scalable exploration and analysis of such datasets.
- The system interprets user requests and maps them to structured analytical operations on underlying metadata.

Strand's scRNA Portal: Snapshot

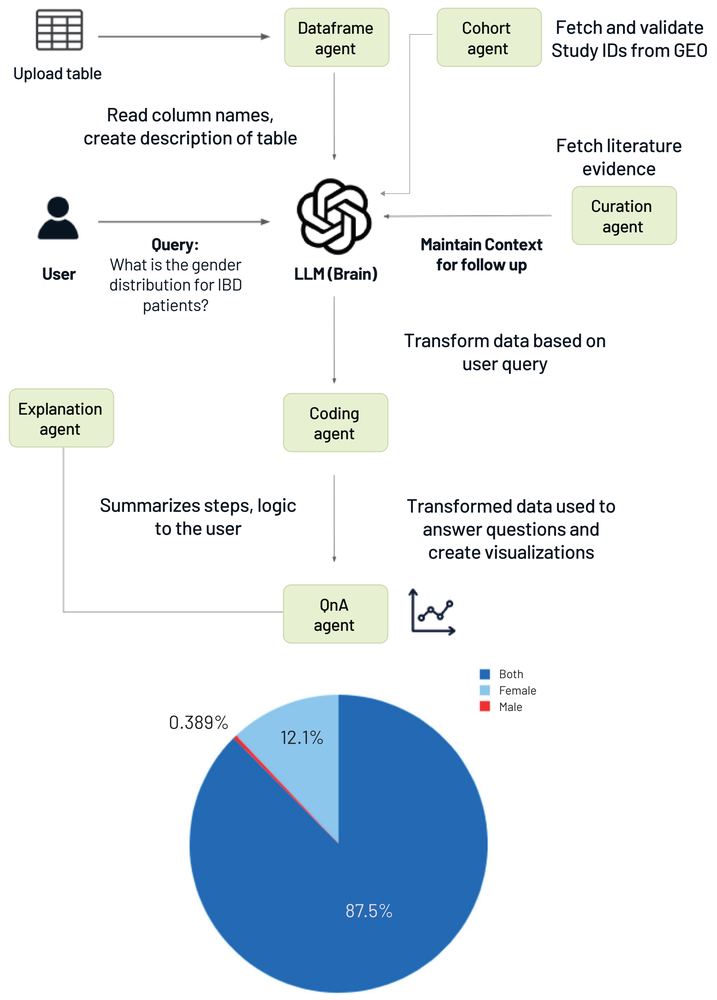
A Disease-Specific Repository

- Our portal hosts deep, curated metadata from specific disease areas like IBD and Neurodegenerative disorders.
- Datasets span a wide range of modalities across single-cell RNA sequencing, bulk RNA and LC-MS/MS proteomics.
- We curated 100+ metadata fields, with 35+ fields involving either ontology mapping or formatting to maintain homogeneity across datasets.

Business value

- Reduce time-to-insight from days to minutes.
- Enable non-technical users to explore complex datasets, curate cohorts for their research goals, fetch relevant literature.
- Support interactive and iterative analysis.
- Ensure data security with local/on-prem deployment.
- Scale across diverse datasets and domains.

Methods



Simple, conversational interface designed for domain scientists without programming or data expertise

LLM + Agentic AI Framework

- LLM interprets user intent and maps it to appropriate analytical operations
- Agent layer selects and executes the appropriate tools based on the interpreted request:
  - Database querying with multi-step reasoning
  - Tool usage (e.g., QnA, visualization,
  - cohort building)

Key Capabilities

- Query and analyze structured experimental metadata (e.g., sample attributes, QC status, study parameters)
- Grounded Responses
  - Outputs are derived directly from the under lying database.
  - Each answer is traceable and verifiable.

  Cohort building based on research goals

Validation Process

- Curated benchmark questions for each dataset.
- Manual validation with Subject Matter Experts (SMEs)
- Evaluation across:
  - Textual responses (25 questions with curated ground truth)
  - Visual outputs (25 questions visually validated by user)
- Cohort agent validated with gold standard manually curated cohorts from GEO.
  - Internal benchmarks: 5 research goals ~30 Study IDs
  - External benchmarks: 1 research goal ~250 Study IDs

Conclusion

- Agentic AI driven exploration transforms how users interact with complex biomedical datasets
- Eliminates the need for manual querying and scripting.
- Maintains accuracy through database-grounded responses.
- A hybrid approach combining LLMs, Agentic AI, and structured databases enables scalable, secure, and efficient data exploration.

Key highlights

- Context-aware, multi-step reasoning via Agentic AI
- Integrated visualization for faster decision-making
- Supports multiple LLM providers: Gemini, OpenAI, Ollama, and OpenRouter for flexible deployment and model choice.
- Support for expanding Cohorts sourced from GEO
- Extensible tool-based framework for integrating additional analysis capabilities

Results

- Accuracy across ~50 questions (textual + visual) ~90%
- Average Latency of the tool is 5.54 secs
- Average Cost with gpt-5.4 is ~$ 0.2
- The estimated reduction is turn around time for getting queries answered is around ~7x
- Cohort agent's precision on internal, external benchmarks: 100%, 92.9%
- Curation agent's TAT reduction in target extraction and mapping to Mechanism of action for a neurological disorder ~2x.

Example Use Cases:

Cohort Discovery Tool

- Gender-based stratification
- Age-based stratification
- Cell/tissue based selection

Biological Insights Assistant

- Summarizing most studied tissue/cell types in different disease states, most commonly used gene markers, cell types involved in treatment response/resistance