CASE STUDY

AI-driven Chatbot Optimization: Achieving Human-Level Accuracy and Speed in Data Retrieval

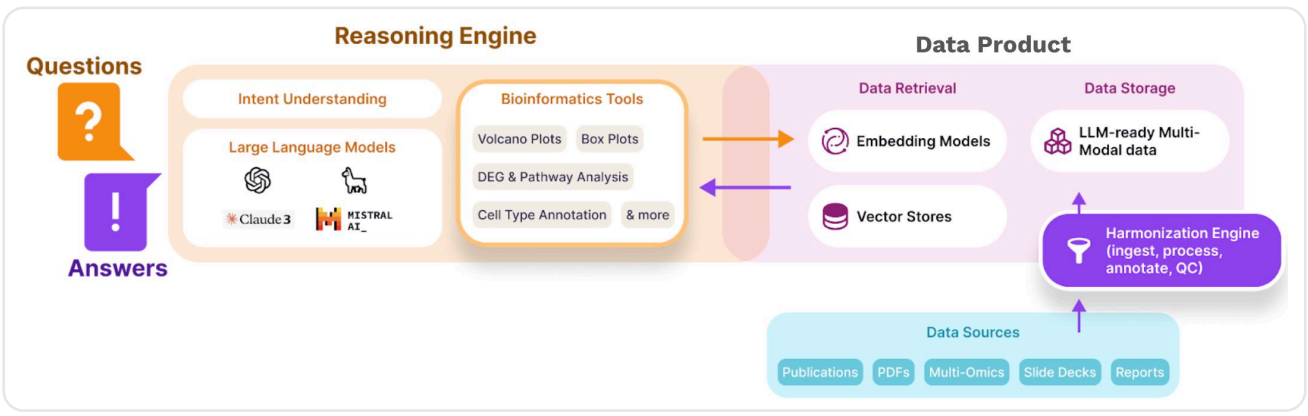
A leading pharmaceutical company needed an **AI-assisted solution for querying large-scale multi-modal data.** Elucidata addressed this by developing an **LLM-powered chatbot** with **optimized RAG** that leverages natural language queries to streamline data retrieval and query execution. This chatbot allows users to retrieve information, answer queries, and perform downstream analysis on muti-modal data and relevant data sources without requiring coding expertise. **Benchmarking against human-expert responses**, Elucidata significantly enhanced the chatbot's performance, delivering **2X better performance** in information retrieval and enabling researchers to obtain insights **5X faster** than manual data querying processes.

Challenges

- **Information Overload and Fragmentation:** Navigating through publications, slide decks, multi-omics tabular data, and reports can be overwhelming. The abundance of scattered, uncurated information significantly hinders efficient data querying, retrieval, and analysis, with the primary challenge being the time-consuming process of sorting and analyzing the data.
- **Lack of Coding Expertise:** Non-bioinformatics experts may struggle with complex queries and downstream analyses, often relying on trained professionals to write code and generate relevant analyses. A user-friendly, no-code solution is needed for efficient data wrangling and generating valuable insights.

The Solution

Elucidata developed an **LLM-powered chatbot** that leverages **harmonized, AI-ready data** and incorporates biological context through **intent understanding** and **retrieval-augmented generation (RAG)** to enhance the efficiency of complex data querying with natural language queries. This approach involved several key components:



Harmonizing Data to Create AI-ready Knowledge Base

Elucidata harmonized diverse multi-modal data sources such as multi-omics data, slide decks, reports publication, and PDFs with corresponding supplementary files into a comprehensive, AI-ready knowledge base. This well-structured knowledge base served as the foundation for efficient and accurate data retrieval.

Using Elucidata's proprietary **harmonization engine**, the data was curated, structured, and made queryable, ensuring consistency in metadata, vocabularies, and terminologies across all datasets. This harmonized data provided the LLM-powered chatbot access to accurate and well-organized information, enabling it to generate precise, contextually relevant responses.

Once harmonized, the data was stored as LLM-ready multi-modal data, ensuring it was curated, annotated, and easily accessible for accurate chatbot interactions. The embedding models and vector stores were integrated for data retrieval into the data product, converting multi-modal datasets into vector representations. The vector-based indexing system, combined with retrieval-augmented generation (RAG), ensured fast and precise data access, further improving the chatbot's ability to deliver high-quality, domain-specific insights.

Intent Understanding and Biological Context

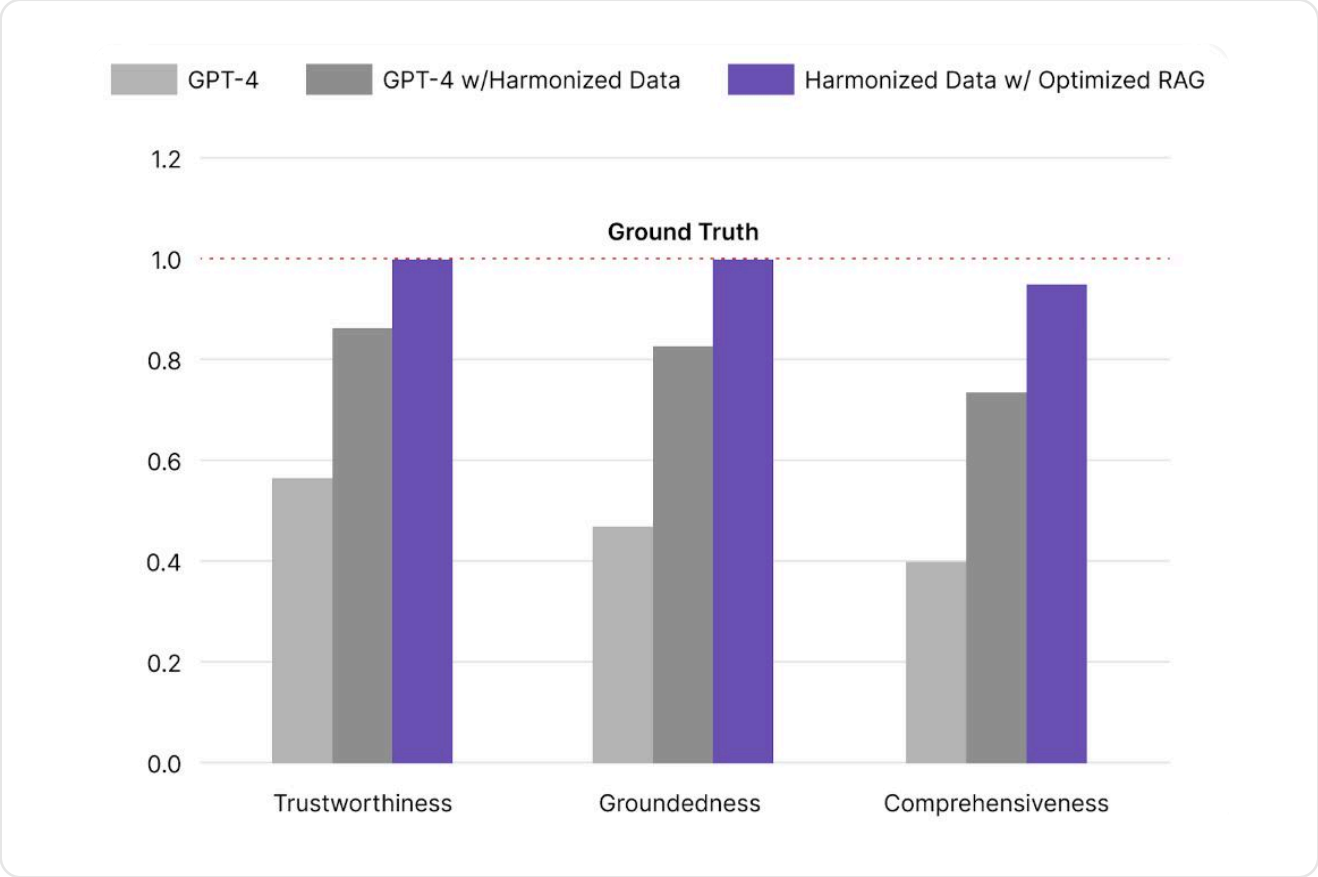
At the core of this solution is a reasoning engine that augments large language models (LLMs) with biological context. The system understands user intent through natural language queries and enriches outputs using bioinformatics tools such as volcano plots, pathway analysis, and cell type annotation. These tools ensure that biological insights are incorporated into the LLM's responses, leading to context-aware, domain-relevant outputs.

Optimizing Retrieval-augmented Generation (RAG)

Elucidata conducted rigorous testing and optimization of the Retrieval-Augmented Generation (RAG) system to ensure precise and efficient information retrieval. This RAG system was designed to extract the most relevant and context-specific data from the harmonized knowledge base in response to natural language queries. By leveraging this knowledge base, the RAG system enhances the specificity and accuracy of information retrieval, ensuring that each user prompt results in precise and reliable outputs.

To evaluate the system's performance, Elucidata conducted a benchmark study comparing the chatbot's outputs to responses provided by human experts. The accompanying graph illustrates these results:

- The **red line** represents the baseline established by human experts, serving as the gold standard for **trustworthiness, groundedness,** and **comprehensiveness.**
- The **bar plots** demonstrate that the RAG-optimized system, in combination with harmonized data, achieves parity with human-level accuracy across all metrics, surpassing the performance of GPT-4 alone or GPT-4 with harmonized data.



Elucidata's optimization efforts enabled the RAG system to achieve human-level accuracy by refining retrieval mechanisms, validating outputs through extensive testing, and leveraging a robust harmonized knowledge base. The RAG system significantly enhances the LLM chatbot's utility for multi-modal data retrieval, improving user satisfaction and decision-making accuracy.

GUI Application Development

To enhance user experience, a graphical user interface (GUI) was developed that features an intuitive chat interface, providing output in text, tables, or charts based on user queries. Each response includes citations that indicate the source of the information, whether from publications, slide decks, multi-omics tabular data, or other resources, enhancing traceability and allowing users to easily verify the data's origins.

This solution simplified data retrieval with natural language queries, improved processing speed, and made unstructured multi-modal data accessible to users without requiring advanced coding expertise.

Impact

5x broader adoption: Building end-user trust in information retrieval through robust evaluation frameworks.

Reduced data retrieval infrastructure setup and maintenance costs by **100K/Yr.**

Turnaround time was potentially reduced from **24 hours to real-time** for user queries to the application chatbot.

About Elucidata

Elucidata's Polly optimizes data quality for pre-clinical drug discovery by harmonizing multi-omics and assay data into ML-ready formats. Polly's powerful harmonization engine is utilized by trained experts to curate diverse data types, annotate metadata, and ensure consistent processing at affordable costs. These ML-ready data are stored on an Atlas on the platform, robust data stores ideal for analysis and management.

Polly's advanced technology accommodates 25+ R&D data types, catering to teams in pre-clinical drug discovery and diagnostics R&D. Polly is trusted by over 25 research organizations, including 4 of the top 10 pharma companies, to accelerate their discovery programs.

Book a Demo