lucidata

White Paper

Impact of High-quality, Deeply Curated Data on Biomedical Data Discoverability

Contents

Abstract

- In recent years, there has been notable progress in LLMs, sparking enthusiasm for integrating these models with extensive databases for **natural language-based information retrieval.**
- Discussions on effective search typically spotlight the model's reasoning abilities, but it's vital to **highlight metadata quality's role** in enabling effective search.
- This whitepaper presents a case study on **data discoverability in a large corpus of gene expression data**, and the **impact of metadata annotation** quality on search outcomes.
- Elucidata's meticulous metadata curation yields precise responses to intricate user queries, significantly shaping AI models' search capabilities through curated data accessibility.

Authors: Gaurang Mahajan, *Ph.D., ML Researcher*; Rajdeep Mondal, *Data Scientist*; Nobal Dhruw, *Senior Manager, ML,* Aqsa Aleem*, *M.A., Associate Marketing Manager*

Acknowledgments: Abhishek Jha, *Ph.D., CEO and Co-Founder*; Swetabh Pathak, *M. Tech., CTO and Co-Founder*; Neychelle Fernandes, *Ph.D., MBA, Director of Technical Sales and Product Marketing*

Design: Mantrala Satya Shiva Sai Sriram, *Senior Visual and Information Designer*

All contributors are affiliated with **Elucidata Corporation**

Introduction

Highly effective use of large biomedical corpora for research needs depends on the quality of data search and retrieval supported in response to user queries. Where the data in the corpus is intricately structured at multiple levels, programmatic searches can enable precise responses to complex user queries. However, **this code-first approach to data search presents a barrier in terms of requiring the user to know a database query language, like SQL, to perform effective searches.**

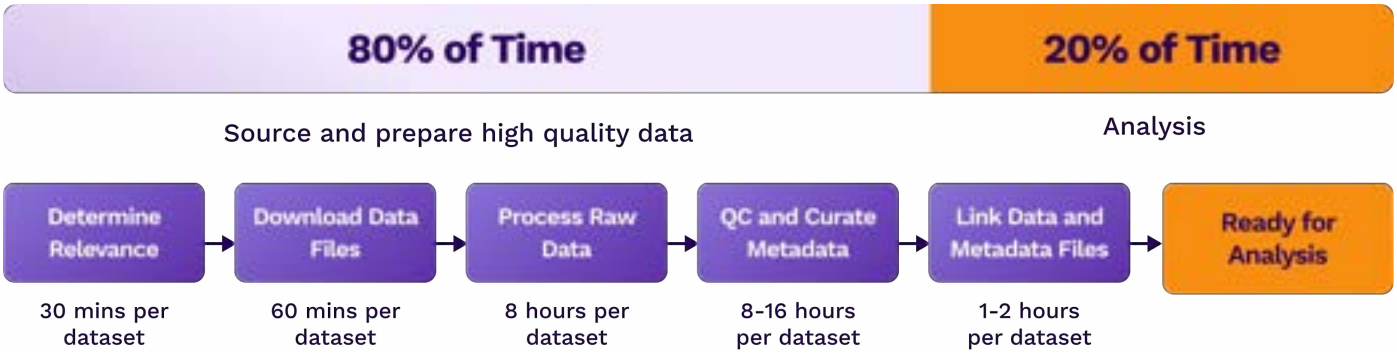
The recent emergence of generative large language models (LLMs), with their demonstrated abilities to convert natural language instructions to code, offers a promising way to address this gap. LLMs connected to a database can potentially act as a bridge, empowering the user to pose queries in natural language, and have precise results from the data corpus seamlessly delivered, in principle. Most discussions around making this work in practice tend to focus on the LLM end.

Efforts are usually devoted to improving the ability of the LLM to correctly interpret nuanced user queries, build in domain awareness to identify biological entities in the query, fine-tune the model to reliably generate structured queries encoding the user question, etc. But placing excess emphasis on the **model alone** is a one-sided view of the complexity of the search problem. It ignores a crucial pre-requisite to enable effective and precise information retrieval, which is the highlight of this whitepaper - the **metadata should be suitably structured and annotated in the database in the first place.**

Building on this theme, we present the results of an in-house case study at Elucidata. It showcases the indispensable role of metadata quality in enabling LLMs to produce meaningful, knowledge-augmented responses to user queries.

Importance of Data Quality for Faster Discovery and the Related Challenges

Data is the cornerstone of discovery in bioinformatics. However, messy and unclean data can pose a significant challenge on the path from raw data to meaningful insights and compromise the integrity of research outcomes. Researchers are estimated to spend roughly 80% of their time preparing data to make it suitable for bioinformatic analysis:



The effort that bioinformatics teams need to devote towards sourcing and cleaning relevant data can translate into significant sunk costs and more importantly, time.

Data available in scattered public sources lacks consistent formats, and its usability may be compromised by missing metadata annotations (meaning missing context). Further, the discoverability of relevant data within a large data corpus via text-based searches is impeded by a lack of standard terminology of experimental design attributes and biological context across studies. Different datasets that essentially refer to the same experimental context and biological attributes, may be labeled in varied and ambiguous ways.

This can make it extremely difficult to establish commonalities across scattered but related studies in a straightforward manner. Data findability relies not only on global dataset annotations but also extends to detailed sample-level attributes. Each sample requires accurate labeling of all relevant biomedical entities, such as tissue types, diseases, cell types, cell lines, drugs, genes, and their genetic perturbations. The demands on curation increase further with more complex queries that go beyond the metadata alone, and involve retrieval of information at the level of the biological measurements themselves, like e.g., finding all bulk RNASeq datasets with neuroendocrine samples where the expression of the gene CHGA exceeds that of YAP1.

This is where Elucidata steps in, harnessing the power of state-of-the-art AI models to mitigate the problem of poor data quality and empower researchers to unlock the full potential of available public multi-omics data for their research needs.

Elucidata's Solution: Polly by Elucidata

Drug discovery research relies heavily on access to high-quality biomedical data. Unreliable, poorly annotated data carries the risk of getting sidetracked by red herrings or missing out on genuine leads, on the road to discovering or confirming novel hypotheses.

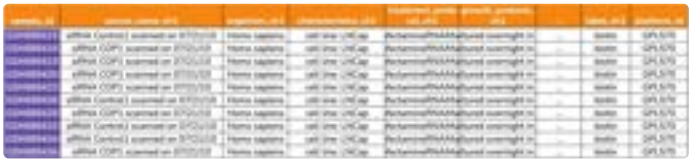
Polly by Elucidata optimizes data quality for pre clinical drug discovery by harmonizing multi omics and assay data into ML-ready formats. Polly's powerful harmonization engine is utilized by trained experts to curate diverse data types, annotate metadata, and ensure consistent processing at affordable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

Polly's Harmonization Engine addresses this crucial need by deeply curating publicly available data from sources like GEO, PRIDE, CPTAC, and various publications. This pipeline involves advanced AI models for automated curation, supplemented by manual review by a team of expert curators, to guarantee the highest data quality.

To illustrate the value-add from the deep curation performed by Polly's harmonization engine, here is a sample dataset sourced from the GEO corpus (dataset id = GSE27914).

To address these challenges and foster data reproducibility in AI models, several strategies are recommended:

Raw Source Data



- The raw data from GEO is akin to a first draft; it's comprehensive but unrefined. It includes a wide array of metadata, but this information is often unstructured and inconsistent.
- The raw GEO data may contain redundancies, ambiguities, and a lack of standardization, which can make it challenging to parse and utilize effectively.
- For researchers, using this raw data requires significant effort to clean, standardize, and interpret before it can be reliably used for analysis.

Data Curated by Polly's Harmonization Engine

- Polly's harmonized data is the result of meticulous curation and standardization into any required ontology for all biological entities. It transforms the raw, unstructured information into a coherent, structured format.
- The curation process involves extracting relevant biomedical information from the GEO dataset pages, their corresponding sample pages, all associated publications, and sample-level metadata tables, ensuring that every piece of data is accurate and standardized.
- By employing advanced natural language processing, Polly aligns the extracted entities with established ontologies at sample level, providing a high-quality annotated dataset that is ready for immediate use, enabling researchers to focus on the insight generation, rather than laboring through the painstaking process of making the data ready for analysis in the first place.

The value of harmonized ML-ready data cannot be overstated.

AI readiness is becoming increasingly relevant with the advent of conversational AI agents that exploit the reasoning capabilities of LLMs. The following case study intends to highlight a particular aspect of this "AI readiness" – the discoverability and retrieval of data stored in biomedical corpora via the adoption of such AI agents. The reasoning abilities of the LLM are often assumed to determine search effectiveness. However, the importance of annotating and making data accessible to the AI model is often overlooked. This plays a crucial role in enabling retrieval-augmented generation (RAG).

With powerful AI agents, how crucial is high-quality harmonized data for data discovery?

Case-study: Role of Metadata Harmonization Quality in AI-assisted Database Search

Context

To determine the necessity of highly curated data for effective responses by Large Language Models (LLMs), we conducted a structured controlled experiment.

The CREEDS project, launched in 2016, focused on crowdsourcing annotations and re-analyzing a substantial number of gene expression datasets from the GEO database. This project encompassed 2460 single gene perturbations, 839 comparisons of diseased versus normal states, and 906 drug perturbation signatures, covering a wide array of datasets that were meticulously curated and quality-checked. We selected these datasets by their unique identifiers and retrieved their corresponding metadata from GEO. This metadata offers a comprehensive overview of each dataset, including essential details like tissue source, disease label, cell type, treatment, genes involved and their genetic modifications.

Typically, such detailed study-level descriptions enable basic searches, like identifying datasets on lung adenocarcinoma. However, for more nuanced queries, such as locating lung adenocarcinoma datasets from primary human patients treated with a specific drug, this level of detail proves insufficient. This scenario underscores the need for in-depth curation at both the sample and data measurement levels.

In our experiment (Figure 1), we assess how LLMs handle data from three distinct sources, each containing the same datasets from CREEDS:

1. Unprocessed data directly from GEO
2. Data manually curated by CREEDS
3. The same datasets but curated through our Polly Harmonization Engine

These sources represent varying levels of data quality, with raw GEO data at the lower end and the data curated by Polly Harmonization Engine at the higher end in terms of quality.

Open schematic-20240117-180856.png

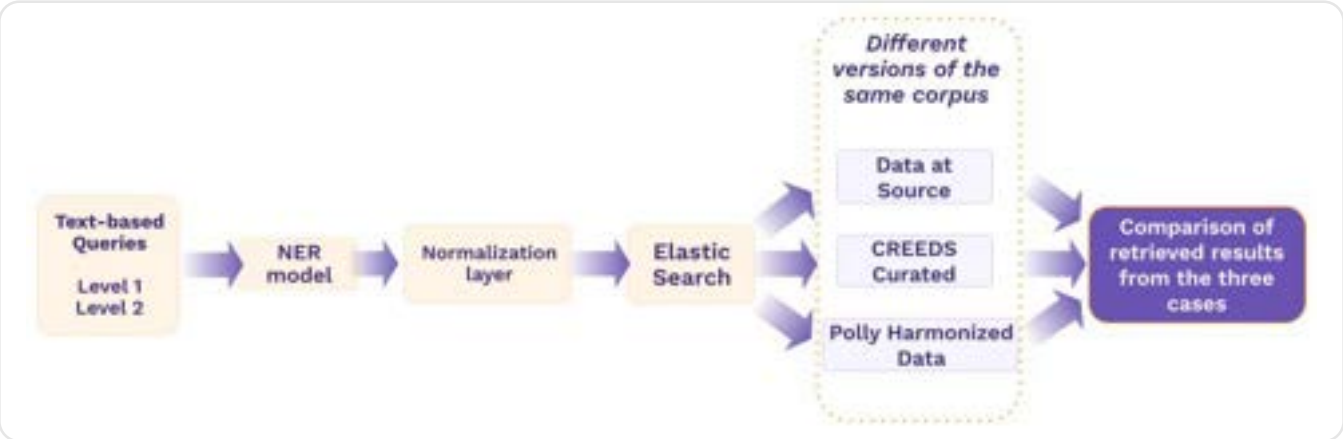


Figure 1: Schematic of our experimental setup to evaluate the quality of retrieval via an AI-enabled RAG pipeline. The key processing steps involved in translating natural language queries into delivery of relevant datasets are highlighted.

Model Architecture and Workflow

Data from the three distinct sources is housed within an ElasticSearch database. For each dataset, including the sample-specific information in the CREEDS and Polly curated datasets, the data is encoded into JSON documents and indexed in this database. It is important to note that for the raw data obtained directly from the source, there is no pre-existing schema for sample-level attributes.

Consequently, the entire sample-level dataset is transformed into a dictionary format and incorporated into the JSON document, lacking any structured schema. In contrast, the CREEDS and Polly curated datasets adhere to a well-defined schema for sample-level data.

Following a natural language query posed by the user, the LLM is:

- primed to interpret the user's request,
- identify the biological entities/terms present in the query,
- and construct a JSON request encapsulating these identified terms.

This request is then dispatched as a POST to the ElasticSearch API endpoint to all three data sources mentioned above. The response from the API is a curated list of datasets, each selected for its relevance to the initial query.

In the model's workflow, when handling curated datasets like CREEDS and Polly harmonized data, it adds an extra vital step. This involves taking the entities identified by the biomedical named entity recognizer in the query, and carefully aligning them with precise terms from established ontologies. For diseases, we use the Human Disease Ontology; for cell lines, the Cellosaurus; for tissues, the Brenda Tissue Ontology; for cell types, the Cell Ontology; and for other entities, our proprietary normalizer. This precise mapping uses the same base technology that supports Polly's curation process.

For example, a general term like 'lung cancer' identified in the query is refined to a more precise ontological counterpart such as 'lung adenocarcinoma', following the Human Disease Ontology. This process allows the model to benefit from Polly's extensive harmonization of the underlying curated data. Our case study summarized below will demonstrate the significant impact of this approach on the overall search outcome.

Experiment Setup

We evaluated the efficacy of AI-aided search and the role of the underlying data quality in facilitating data discoverability. To begin with, we generated natural language queries. These queries mimic the typical searches that would be performed by a biomedical scientist interested in finding gene expression studies relevant to their research question in a database. The AI-enabled agent is primed to interpret every supplied query and translate it into a corresponding ElasticSearch query. We assessed whether the model is able to retrieve the "correct" datasets from the data source, in response to each query. The search experiment was performed across each of the three above-described data sources in turn, using a common pool of queries, ensuring a fair comparative assessment of the search efficacy.

The following steps were involved in the experiment.

Framing Natural Language Queries:

A diverse pool of ~450 natural language search queries, each conditioned on a random combination of biological terms, was first generated in an unbiased manner. Each query can be answered by returning a set of relevant datasets from the corpus. These datasets define the "ground truth", i.e. the ideal response to every query.

For example: given embryo as tissue of interest and gene perturbation Rex1, a valid query is: "I'm looking for datasets where the gene Rex1 knockout was studied in embryo tissue."

Further, we used the following rules to pre-select the queries to be included in the "test set":

1. To keep things realistic, the queries only involved combinations of biological terms represented in the corpus, i.e. we ensure each query has at least one ground-truth dataset present in the corpus
2. The biological terms occurring in the query were replaced by synonymous terms, where possible. This was done to mimic user queries in a more realistic sense, as user queries are likely to involve semi-technical or non-standard terms. This also helps evaluate the efficacy of the metadata harmonization layer to map the user-supplied terms to the right intended concepts. [Note: The list of synonyms which were used as alternatives to the standardized terms, were obtained directly from the source .obo files for the respective ontologies.]
3. The queries were partitioned by complexity into type 1 and type 2 queries. Type 1 queries are high-level and can be typically answered based on the aggregated metadata at the dataset level, such as finding studies involving a tissue or disease of interest. Type 2 queries are more nuanced, and rely on sample-level metadata availability; they can be about finding studies wherein a case versus control comparison has been performed for the user's experimental condition of interest (for instance: "Find datasets where effect of Dabrafenib relative to untreated control was profiled in skin tissue").

Comparing the Results Across the Data Sources: Evaluation criteria

Every query returns a list of datasets from the corpus, ranked by the relevance score. As the ideal response to each query is already known in this controlled setting, we can objectively benchmark the efficacy of the search across the three data sources compared.

The appropriate metrics to evaluate the quality of the search results depend in general on the overall context and expectations from the search. We scored the quality of retrieval in terms of the following criteria (see Figure 2 for an explanation):

1. Recall - For a given query, does the search return all the correct/relevant datasets present in the corpus?
2. Relevance - Given a query, are the correctly identified datasets ranked near the top of the returned list, or does the search often return “false positive” matches too?

(1) Aggregated Reciprocal Rank (ARR): To compute this scoring metric for each query, the reciprocal ranks (1/rank) of those ground truth datasets for the query, which are present in the returned list, are summed up. The ground-truth datasets that fail to show up in the search result are assigned zero weight. This sum of reciprocal ranks is then divided by the corresponding "best-case scenario", i.e. the case wherein all the ground truth datasets corresponding to the query are retrieved, AND show up at the very top of the returned list. This yields a normalized score for every query.

The ARR metric takes into account both the recall and the relevance of the search outcome to the user’s query. A high ARR is obtained when the search returns as many of the ground truth items as possible (high recall rate), and when the correctly retrieved items are ranked very high up in the list (i.e. they are assigned the highest relevance). The particular form of the ARR we use here extends a similar score (mean reciprocal rank) which is sometimes used to evaluate recommendations. However, we are interested here in the rank of not just the first correctly retrieved ground truth dataset. Instead, our metric aggregates and thus takes into account the ranks across all the relevant datasets in assessing the accuracy of the response to the query.

(2) Maximum F1 Score: The F1 Score is another oft-used performance metric that balances recall and precision of a model's outputs. Given our ranked list of returned search results for a query, we first evaluate the precision and recall at different score thresholds (or rank thresholds) K. The F1 score combines precision and recall in a single composite metric (Reference).The maximum F1 Score across all the tested thresholds is then selected to represent the overall quality of the search, and this is repeated for all the queries in our test set. The ideal F1 Score (equal to 1) will be attained only when all the ground truth datasets are retrieved by the search, and they are ranked at the top of the search results sorted by the relevance score.

(Minor note: Neither of the above performance metrics penalizes the occurrence of off-target items in the returned results, as we did not impose any explicit relevance score threshold to restrict the search result.)

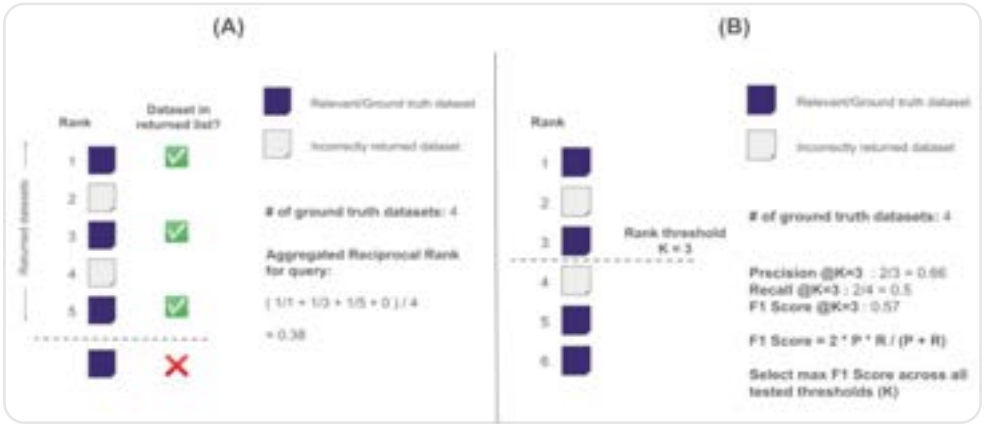


Figure 2: Calculating the search performance metrics on the ranked list of datasets, illustrated with simple examples. (A) Aggregated Reciprocal Rank (B) Maximum F1 Score, which is a combination of precision and recall

Results

Based on our evaluation framework, we compared the results of AI-enabled dataset search across the three data sources representing different levels of curation.

The following example illustrates the differences in search outcomes across the different data sources.

Question - Fetch studies comparing gene expression profiles in dorsal root ganglion.

Ground Truth Datasets

- `GSE15041_GPL1355`
- `GSE2636_GPL85`
- `GSE59727_GPL6101`

Dataset ID	Dataset is in ground truth?	Rank of dataset in returned results (in different versions of the corpus)		
		Polly Harmonized	CREEDS	Raw Source (GEO)
`GSE15041_GPL1355`	Y	1	Not returned	2
`GSE2636_GPL85`	Y	1	Not returned	Not returned

Dataset ID	Dataset is in ground truth?	Rank of dataset in returned results (in different versions of the corpus)		
		Polly Harmonized	CREEDS	Raw Source (GEO)
GSE59727_GPL6101	Y	1	Not returned	Not returned
GSE11208_GPL570	N	2	Not returned	Not returned
GSE1371_GPL85	N	3	Not returned	Not returned
GSE15293_GPL1261	N	3	Not returned	Not returned
GSE594_GPL85	N	3	Not returned	Not returned
GSE16710_GPL1355	N	Not returned	1	Not returned
GSE633_GPL479	N	Not returned	1	Not returned
GSE2869_GPL1261	N	Not returned	Not returned	1
GSE27028_GPL6246	N	Not returned	Not returned	3
GSE1839_GPL81	N	Not returned	Not returned	4
GSE23910_GPL6480	N	Not returned	Not returned	4
GSE31453_GPL1261	N	Not returned	Not returned	5
GSE48989_GPL6244	N	Not returned	Not returned	6

In the above example, all three ground truth datasets are correctly retrieved upon querying the **Polly Harmonized** database, and they are all ranked in first position in the returned list. Both the performance metrics (ARR and F1 Score) are equal to 1 in this case.

In the case of the response based on the **CREEDS curated** source, none of the ground truth datasets were correctly retrieved (ARR = 0, F1 Score = 0). Instead, two false positive results are returned.

The **Raw source** contains the same corpus of GEO datasets, but in a relatively unstructured format and lacks harmonization. Upon querying the raw source, 1 out of the 3 ground truth datasets is returned, and it is ranked at 2nd position in the returned list. This gives an ARR score of 1/6 and a maximum F1 score of 0.4.

Thus, in this particular instance, the meticulously curated Polly Harmonized corpus yields a more accurate response with 100% recall compared to the responses based on the alternate data sources. Let's see how this example generalizes to the full set of tested queries.

Figures 3 and 4 display the distribution of values of the selected performance metrics across all the tested queries. The results show a striking dependence on the data source.

They demonstrate a significant improvement in the search responses when using the Polly Harmonized version of the data corpus, compared to the other two data sources. The LLM-enabled search against the Polly Harmonized corpus accurately retrieves the ground truth datasets for a larger proportion of the tested queries. On the other hand, there is considerably more scatter in the metrics across the queries, and poorer outcomes (lower scores), in the case of the raw source (GEO) and CREEDS.

The Polly Harmonized data ensures accurate responses to the questions, while minimizing the chances of the relevant datasets being missed out. This is made possible by the following factors working together in sync:

- the richly annotated and structured metadata,
- the entity extraction and normalization layers involved in preprocessing the user input, and
- the ElasticSearch query-generating capabilities of the LLM.

By contrast, some of the metadata labels are missing in the manually curated CREEDS version of the corpus. This impacts the retrieval quality of the AI-enabled search due to unavailability of sufficient information needed to correctly identify the right datasets, given the biological context contained in the question. Finally, in the case of the Raw Source version of the corpus, although all the requisite metadata information is present in the data source, it is in a relatively unstructured format, lacking harmonization, and this compromises the reliable retrieval of the right datasets in response to the LLM-generated search queries.

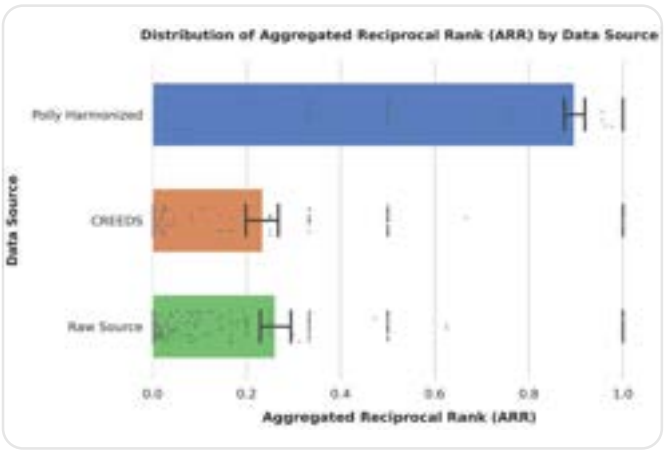


Figure 3: Distribution of the pooled Aggregated Reciprocal Rank (ARR) scores across the results of the tested queries, comparing the three data sources evaluated. Bars represent the mean values and error bars represent standard deviation.

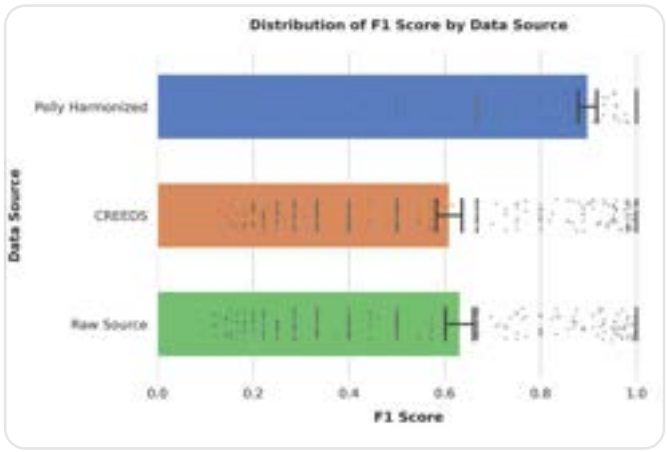


Figure 4: Distribution of the pooled F1 Scores across the results of the tested queries, comparing the three data sources evaluated. Bars represent mean values and standard deviation is depicted by error bars.

Impact: Table 1 summarizes the data distribution displayed in Figures 3 & 4, in terms of the median values of the retrieval quality scores across the different versions of the data corpus. These aggregated scores demonstrate the significant impact of data harmonization on search performance. The F1 Score of search results from the Polly Harmonized corpus is **75% better than CREEDS** and nearly **50% better than the raw (uncurated) GEO corpus.** Our results thus reveal a sharp improvement in the knowledge-augmented responses of the LLM when it is paired with the deeply curated version of the corpus from the Polly harmonization engine, in place of the relatively less curated original CREEDS and raw (GEO) versions of the same data corpus.

Table 1: Median of the distribution of values of the search performance metrics across the tested queries, shown by data source. The three sources represent different depths of curation of the same data corpus (high to low).

Performance Metric	Polly Harmonized	CREEDS (manually curated)	Raw Source (GEO)
Aggregated Reciprocal Rank	1.0	0.0	0.025
Max F1 Score	1.0	0.57	0.67

Polly by Elucidata: Enhancing Discoverability Beyond the Metadata

The above results summarize our experiments around improving responses to user queries conditioned on the metadata. That is, they are about finding datasets in the corpus based on the biological entities or experimental conditions mentioned in the query posed by the user. But the data discoverability enabled by Polly's Harmonization Engine is not limited to just dataset or sample/cohort level metadata. The deeply curated corpus is also geared to enable data-centric queries, where the LLM acts as an interface to extract meaningful information from the stored data, and return relevant datasets conditioned on the data as well. We showcase this ability of the LLM to interpret and perform basic data analytics according to the query posed by the user, with some illustrative examples below. This is supported by the structured format in which the data is stored in the Polly Harmonized corpus. The following example searches directly rely on parsing the gene-level information contained in the underlying data matrix.

Query: Find breast cancer datasets where the gene CASP9 shows significant down-regulation of expression in tumor compared to healthy samples.

Response:

dataset_id	Log-fold change of CASP9 expression (tumor v/s normal)
GSE9574_GPL96	-8.69
GSE3744_GPL570	-8.8
GSE26910_GPL570	-8.47

Query: In which datasets studying gene expression changes induced by estradiol treatment, does the estrogen-related signaling pathway gene EEF1A1 show differential regulation?

Response:

dataset_id	Log-fold change of EEF1A1 expression (estradiol-treated v/s control)
GSE12446_GPL570	-13.78
GSE26834_GPL571	-14.9

Conclusion

Our case study summarized above, based on an unbiased sample of realistic queries, underscores the crucial role played by data quality in enabling superior retrieval of the most relevant information from a data corpus. It is not enough that a language-understanding AI agent reasons and correctly interprets the intent behind a user's question, unless the knowledge source also has been systematically curated, annotated, and structured to facilitate the discoverability of the relevant data. These two facets of the search pipeline thus need to work synergistically to enable efficient translation of, and accurate context-aware responses to, user queries, as demonstrated by our analysis.

It is a cliche to say that an AI model is only as good as the data it has been trained on. The results of the above case study emphasize that this sentiment is not limited to the model training alone, but that the value of data quality extends further, to facilitating the actual real-time operation of the model as well. Our findings highlight that the high-quality deeply curated metadata, created by the Polly harmonization engine, is future-ready to meet the challenge of leveraging AI-enabled tools to find one's way to the right needles (relevant data) hidden within the haystack of large-scale biomedical corpora.

Supplementary Files

1. **Metrics For Discoverability for measuring Data Quality** - This contains all the queries sorted by relevance scores for three data sources: raw_source, CREEDS, and Polly Harmonized, along with their corresponding datasets for ground truth in a separate sheet.
2. **Polly Harmonized Sample Level Metadata for all the datasets in the corpus - Link**
3. **CREEDS Sample Level Metadata for all the datasets in the corpus- Link**
4. **Raw Source Sample Level Metadata directly from GEO for all the datasets in the corpus- Link**

Note :- Each sheet name in the spreadsheet corresponds to a dataset id in the format GSE123_GPL456, and contains sample level metadata for that specific dataset id.

About Elucidata

Elucidata transforms biological discovery by providing high quality bulk RNA-seq and single-cell data, among other data types. They support discovery programs at top pharma companies and have 35+ research partners from premier biopharma companies and research labs.

Their FAIR biomedical data platform, Polly, makes data easily findable and more reusable. Elucidata has helped R&D teams scale up and has enabled 10x faster identification of therapeutic assets with high odds of success in the clinic. Having aided the detection of multiple validated drug targets across immunology, oncology, and metabolic disorders, Elucidata looks forward to helping more teams reach their R&D goals quicker!

For more information, visit **Elucidata | Home** or reach out to us at **info@elucidata.io**

Locations/Offices

Book a Demo