Towards Iterative and LLM-Assisted Dataset Search Interfaces

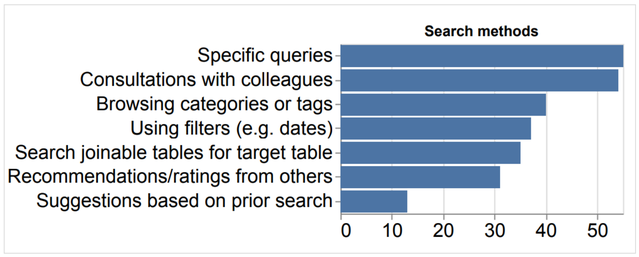
Rachel Lin, Wenjing Lin, Bhavya Chopra, Shreya Shankar, Madelon Hulsebos, Aditya G. Parameswaran

Background

Problem

Why and How We Search: Insights from Survey

(79%) To find the right dataset for analysis



Top search methods:

1. Specific queries
2. Consultations with coworkers
3. Filters/categories

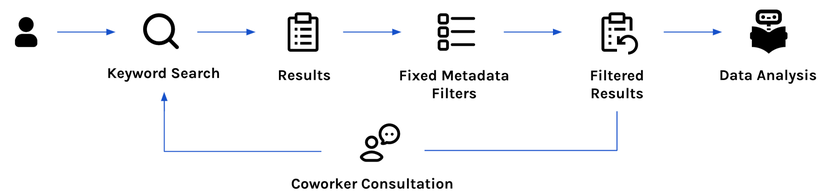


Figure 1: The standard industry dataset search workflow is rigid and offers little support for improving queries, often leading to significant time spent consulting with colleagues.

Objective

Build a **flexible** interface that allow users to **independently** & **iteratively** search for datasets, providing **proactive guidance** to help build effective queries tailored to their specific tasks.

System Design

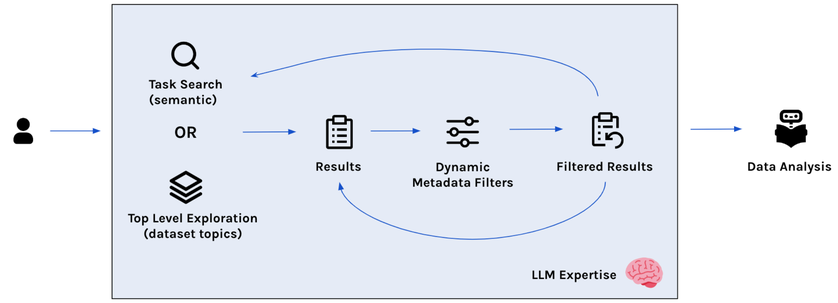


Figure 2: Framework of the proposed interface. Task/topic and metadata searches are performed through a natural language interface. The entire search process is supported by an LLM, which provides query suggestions, dynamically generates task and metadata filters, and answers clarification questions.

Design Principles

D1 LLM Elicitation through Proactive Guidance

Purpose: Prompt users to share more information about their needs, which will be reflected in the query blocks & search interface.

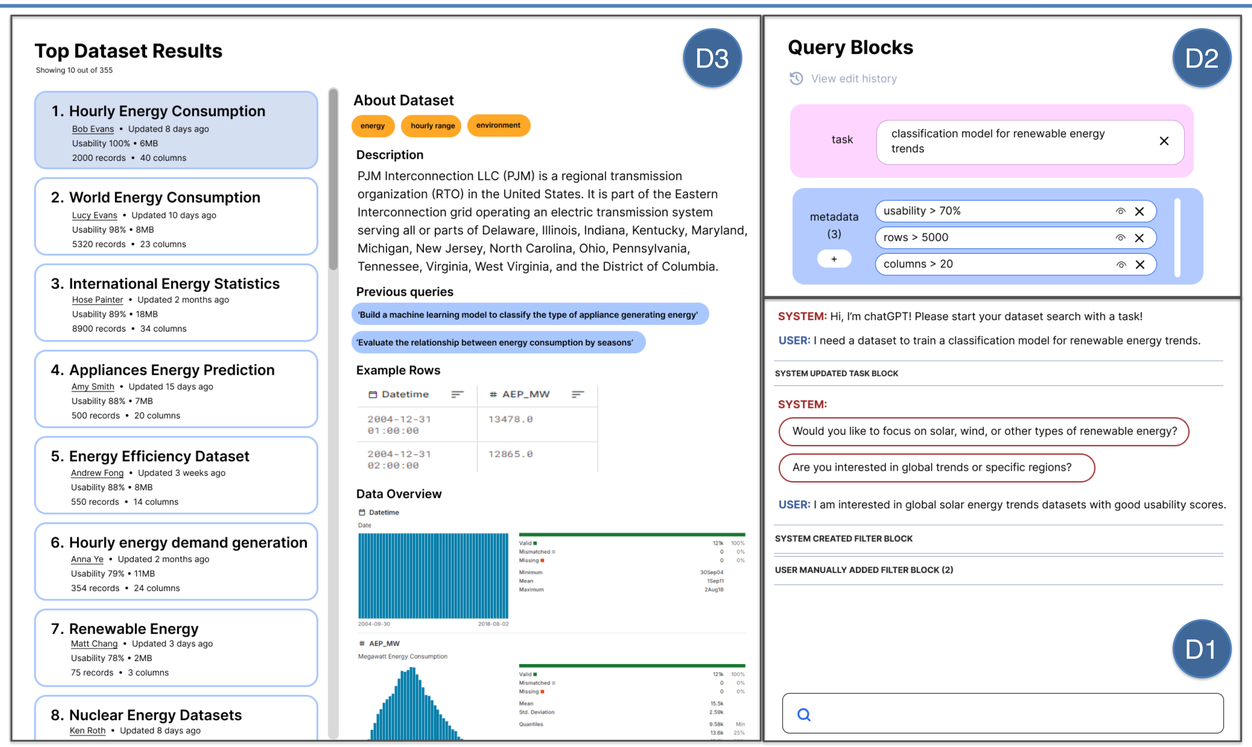
D2 Dynamic Query Decomposition

Purpose: Allow users to see how the LLM is dynamically updating and refining the search space, providing transparency into the search process.

D3 Allowing Users to Compare Datasets Efficiently

Purpose: Facilitate high-level exploration of datasets by organizing them into topics and enable users to delve into metadata details of individual datasets as they iteratively build and refine their queries.

Interface



Next Steps

1. **Providing "better" proactive guidance**
   - How to inform search space for remaining metadata attributes?
   - How to convey which proposed change will result in the greatest amount of disparity?
2. **Creating a baseline interface**
   - Features to include: natural language queries, fixed filters, displaying search results
3. **Evaluation against baseline interface with user studies**

Work Referenced

Madelon Hulsebos, Wenjing Lin, Shreya Shankar, and Aditya Parameswaran. 2024. It Took Longer than I was Expecting: Why is Dataset Search Still so Hard? In Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics (HILDA 24). Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3665939.3665959