Quest2DataAgent: Automating End-to-End Scientific Data Collection

Tianyu Yang[1], **Yuhan Liu**[2], **Ethan Brown**[4], **Sobin Alosious**[3], **Jason Rohr**[4], **Tengfei Luo**[3], **Xiangliang Zhang**[1*]

[1]Department of Computer Science and Engineering, University of Notre Dame
[2]Department of Chemistry and Biochemistry, University of Notre Dame
[3]Department of Aerospace and Mechanical Engineering, University of Notre Dame
[4]Department of Biological Sciences, University of Notre Dame

Abstract

Scientific research often requires constructing high-quality datasets, yet the current workflows remain labor-intensive, and dependent on domain expertise. Existing approaches automate isolated steps such as retrieval or generation, but lack support for the full end-to-end data collection process. We present Quest2DataAgent, a general-purpose multi-agent framework for automating scientific data collection workflows. Given a natural language research question, it decomposes tasks into structured subtasks, retrieves relevant data using hybrid strategies, evaluates dataset quality, and generates visualizations through a conversational interface. We demonstrate its flexibility in two domains: EcoData for ecological research and PolyData for polymer materials. Both systems share the same core architecture but operate over distinct datasets and user needs. Human evaluations show that Quest2DataAgent significantly improves data relevance, usability, and time efficiency compared to manual collection and tool-assisted baselines. The framework is open-source and extensible to other domains.[1]

1 Introduction

Recent studies have explored automated methods to alleviate the labor-intensive and time-consuming manual workflows typical of scientific data collection (Xu et al., 2021; Chen et al., 2022; Chew, 2023; Liu et al., 2024). Tool-based approaches (Liu et al., 2024; Schick et al., 2023; Qin et al., 2023) leverage search engine APIs to automatically label, and clean image data from web sources. In parallel, LLM-driven frameworks (Huang et al., 2024; Zhu et al., 2023; Wang et al., 2024b) utilize large language models (LLMs) to support dataset design, synthetic data generation, and automated annotation. Additionally, multi-agent systems (Borgeaud et al., 2022; Lewis et al., 2020; Yu et al., 2023) have been proposed for synthesizing new datasets (Sengupta et al., 2024; Long et al., 2024; Arif et al., 2024). While these methods can retrieve or generate data that match specific predefined criteria, they typically address only a portion of the broader end-to-end data collection workflow. In scientific research, the full pipeline from formulating a question to acquiring relevant data often involves multiple stages of reasoning, and iterative refinement that go beyond data retrieval or generation alone.

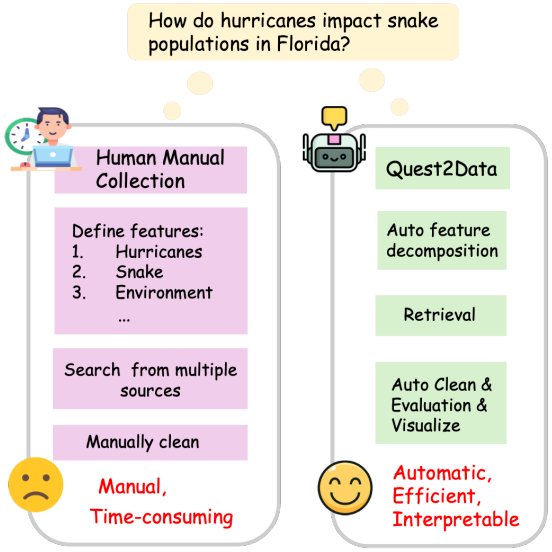

Figure 1: Scientific Data Collection: Labor-intensive manual process vs. automated Quest2Data workflow.

As shown in Fig. 1 (left), for addressing an ecological research question: *"How do hurricanes affect snake populations in Florida?"*, scientists must first decompose the problem into well-defined subtasks and then identify and collect relevant ecological data across multiple dimensions from diverse sources. This workflow requires labor-intensive efforts and relies heavily on domain expertise. Our collaborators in domains such as biology, ecology and material science have expressed signifi-

*Corresponding author: xzhang33@nd.edu

[1]Code is available at https://github.com/Tianyu-yang-anna/Quest2DataAgent. A video demonstration is available at https://youtu.be/7-XEeVpdPZk.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 500–514
November 4-9, 2025

cant challenges in managing this process efficiently, highlighting a critical need for automation in scientific data collection and analysis.

To address these challenges, we introduce **Quest2DataAgent**, the first multi-agent system specially designed to support end-to-end scientific data collection workflows, from problem formulation and data retrieval to evaluation and analysis, as shown in Fig. 1 (right). This system is motivated by the growing capabilities of LLMs to emulate expert-level reasoning across domains (Wu et al., 2025; Qian et al., 2024; Wang et al., 2025). Recent advances, such as OpenAI's Deep Research (OpenAI, 2025), demonstrate that LLMs can effectively navigate large-scale knowledge spaces. However, while such systems excel in reasoning and synthesis, they fall short when it comes to automated, grounded scientific data collection. Quest2DataAgent bridges this gap by simulating expert workflows: decomposing a research question into structured subtasks, retrieving data from heterogeneous sources through a plugin-based architecture, evaluating dataset relevance using automated LLM-based evaluators, and dynamically generating intuitive visualizations.

To demonstrate the feasibility and value of Quest2DataAgent, we developed two system prototypes tailored to ecology and materials science, respectively named as **EcoData** and **PolyData**.

- **EcoData**: A demo system for ecological data collection that supports structured queries across 1,900 species and 172,000 recorded observations, along with environmental events (e.g., hurricanes, wildfires) and climate variables such as temperature, and precipitation.
- **PolyData**: A demo system for polymer data collection that enables fine-grained queries based on chemical structures and functionalities. It supports search over 12,800 polymers, including comprehensive polymer metadata and simulated physical properties.

These two prototype systems demonstrate Quest2DataAgent's capability to support complex scientific queries across domain-specific datasets. Notably, both prototypes are built on curated data repositories that are fully accessible, rather than relying on open-ended web searches or access-controlled data sources. While such scenarios are left for future extensions, or may require alternative strategies for secure and scalable access, the current systems already reduce manual data collection workloads from several days to just 10 minutes, highlighting the efficiency and transformative potential of Quest2DataAgent.

Quest2DataAgent is domain-agnostic, and easily adaptable to new scientific domains by modifying only the user-facing interface and data plugins. The core modules include: (1) Planner Agent: breaks down complex questions into structured subtasks; (2) Retriever Agent: performs hybrid retrieval via a plugin-based architecture; (3) Data Preprocessing Module: cleans, standardizes, and integrates data; (4) Evaluation Agent: uses LLMs to assess dataset relevance and suggest additional resources; (5) Visualization Agent: generates tailored, executable visualization code; (6) Conversational Interface: enables interactive, multi-turn data exploration. Further details are presented in Section 3.

In summary, Quest2DataAgent provides a flexible, modular framework for automating end-to-end scientific data collection. Through qualitative case studies and human evaluations for EcoData and PolyData, we confirm Quest2DataAgent's effectiveness in reducing manual effort while preserving the relevance and usability of the retrieved data.

2 Related Works

LLM-based Scientific Assistants: Large language models (LLMs) have shown significant potential in scientific tasks such as question answering (Dasigi et al., 2021; Lee et al., 2023; Xu et al., 2024), summarization (Ding et al., 2023; Takeshita et al., 2024; Li et al., 2023), and literature recommendation (Zheng et al., 2024; Pinedo et al., 2024). Recent work extends LLM usage to hypothesis generation (Wang et al., 2024a; Si et al., 2024), experimental design (Bran et al., 2023; Li et al., 2024; Lu et al., 2024), and manuscript drafting (Wang et al., 2019; August et al., 2022). However, existing systems typically overlook dataset-centric tasks, such as retrieval, cleaning, and visualization, essential for data-driven research.

Quest2DataAgent addresses this gap by automating dataset retrieval, integration, evaluation, and visualization, thereby providing end-to-end support for complex scientific research workflows.

Scientific Data Retrieval: Retrieval-augmented generation (RAG) enhances LLMs with external knowledge sources (Borgeaud et al., 2022; Lewis et al., 2020; Yu et al., 2023). Early RAG systems focused on unstructured text, but recent work covers more modalities and domains, including finance (Zhao et al., 2025), education (Jia

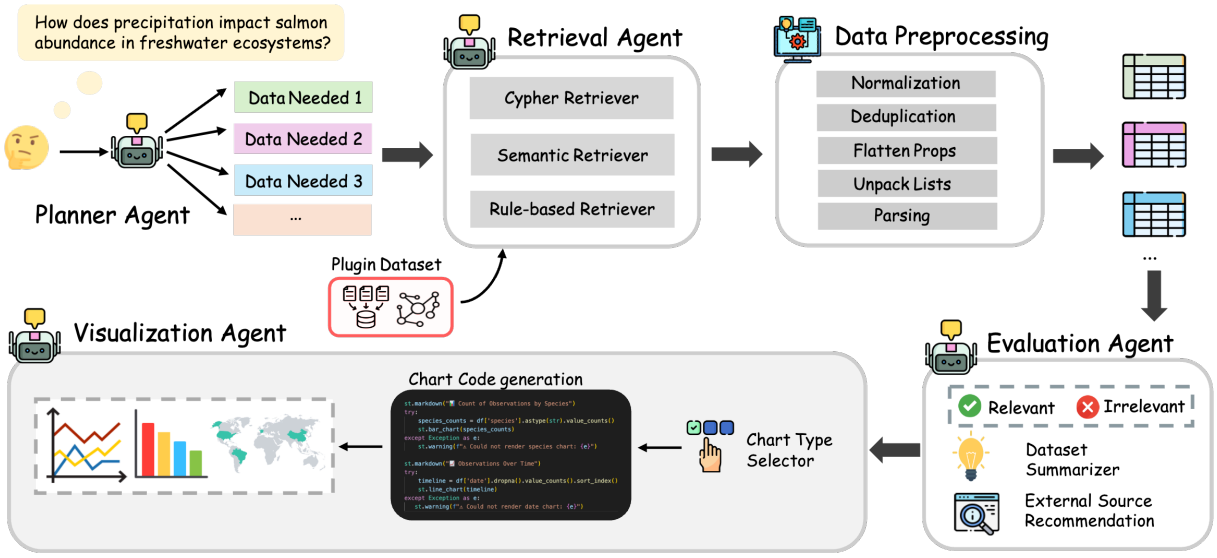

Figure 2: Overview of the Quest2DataAgent framework.

et al., 2025), and multimodal scientific knowledge (Tian et al., 2025; Han et al., 2025). Tools like WildlifeLookup (Wang et al., 2025) and HeteRAG (Yang et al., 2025) enable structured or heterogeneous retrieval for scientific tasks.

However, most of these systems are built for only the retrieval step, which is only a portion of the end-to-end data collection workflow.

LLM-based Multi-Agent Systems: Recent LLM-based multi-agent frameworks (Qin et al., 2023; Hao et al., 2023; Guo et al., 2024) have demonstrated strong capabilities in both complex planning and collaborative problem-solving tasks (Chen et al., 2024; Rasal; Qian et al., 2024; Dibia et al., 2024; Ku et al., 2025). However, most of these systems are not tailored to the needs of scientific data workflows. Quest2DataAgent addresses this gap by introducing a modular multi-agent architecture specifically designed for scientific dataset construction.

3 The Quest2DataAgent System

As illustrated in Fig. 2, the Quest2DataAgent system integrates multiple specialized modules that work together to decompose tasks, retrieve multimodal data, assess relevance, visualize results, and support interactive exploration.

3.1 Planner Agent

A core challenge in scientific data construction lies in the abstract and ambiguous nature of many research questions, which often lack explicit data requirements. These queries frequently span multiple domains and require the integration of heterogeneous data sources. To address this, Quest2DataAgent introduces a Planner Agent that simulates the reasoning process of researchers.

Given a research question, the Planner Agent (1) identifies key semantic dimensions, such as target species, spatial scope, and temporal resolution, and (2) decomposes the question into a set of structured subtasks. Each subtask contains a concise title that indicates the dataset needed and a short description of the corresponding data requirement. This structured decomposition transforms vague research queries into executable data retrieval plans, laying the foundation for efficient and interpretable downstream processing.

To ensure accurate task decomposition, we design a structured prompt template that clearly defines the role of the Planner Agent, the expected output format, and the overall task objective. We incorporate Chain-of-Thought (CoT) prompting (Wei et al., 2022) to guide the agent through step-by-step reasoning. The prompt includes a few carefully selected in-domain examples, each illustrating how to transform a complex research question into structured subtasks with concise titles and descriptions. These few-shot examples enable the agent to generalize effective decomposition strategies to previously unseen queries. The full prompt is shown in Appendix A.3.

3.2 Multi-Strategy Retriever Agent

Scientific data collection faces unique challenges: relevant information is often dispersed across het-

erogeneous sources, with data distributed among structured knowledge graphs, semi-structured tables, and unstructured documents. Coverage is frequently sparse or fragmented, schemas and query patterns are diverse, and access restrictions or data quality issues are common obstacles. These complexities make it difficult to retrieve all relevant data using a single retrieval strategy.

To address this, Quest2DataAgent adopts multiple retrieval strategies (**Cypher**, **semantic**, and **rule-based**) and dynamically selects them based on query type and data modality. Given a subtask query q and a candidate data unit $x \in \mathcal{X}$, where $\mathcal{X}$ is the set of all retrievable data units (e.g., graph triples, table rows, metadata entries), the system computes a relevance score $R(q, x)$ to surface semantically aligned results. Each retrieval strategy is tailored to the specific format of x, enabling effective handling of structured, semi-structured, and unstructured scientific data.

Cypher Retrieval: For structured data in knowledge graphs (e.g., Neo4j), LLM-generated Cypher queries, conditioned on schema descriptions, enable multi-hop reasoning over entity-event-location chains. If the result set is too small or previously seen (based on hash), a fallback is triggered to avoid trivial or redundant outputs.

Semantic Retrieval: For unstructured data, we encode q and each $x \in \mathcal{X}$ into dense vectors via a pretrained encoder, computing $R(q, x) = \langle f(q), f(x) \rangle$. Top-$k$ results are retrieved, optionally re-ranked with metadata filters (e.g., species, location, time). This strategy handles vague queries and diverse schemas.

Rule-Based Retrieval: When the above fail, a rule-based module extracts entities and constraints from q and fills Cypher/SQL templates (e.g., *"species in Florida before 2010"*), ensuring coverage under noisy or under-specified inputs.

Together, these strategies provide robust, interpretable, and flexible retrieval across structured and unstructured sources. When no relevant data can be retrieved due to limited coverage or access restrictions, a fallback module suggests external data sources with direct links, ensuring continuity by guiding users to alternative data pathways.

Plugin Dataset Support: To accommodate diverse domains, users can connect custom Neo4j databases or upload structured text files.

3.3 Data Processing Module

Scientific data collection often poses challenges beyond those in general datasets. For example, polymer science faces limited and fragmented databases, slow and costly data generation, and high sensitivity of properties to experimental conditions. Key fields such as molecular structure or SMILES notation are frequently missing or inconsistently formatted, requiring extensive expert curation. Manual cleaning is tedious, error-prone, and demands domain expertise.

To address these issues, Quest2DataAgent employs a data processing module that performs schema normalization, noise reduction, and format unification. The system flattens nested structures, unpacks singleton lists, standardizes key semantic fields, and removes duplicates via hash-based comparison. Type conversion and missing value imputation are guided by schema heuristics.

The processed output is (1) structurally aligned for downstream evaluation, (2) compatible with automatic visualization generation, and (3) free from trivial redundancy. This automation transforms raw, heterogeneous data into structured and interoperable datasets, reducing manual effort and preserving critical scientific information for downstream analysis. This stage ensures robustness and scalability across scientific domains without requiring manual data processing.

3.4 Evaluation Agent

Assessing dataset relevance and quality is a critical step in scientific workflows. An agent-based evaluation module is thus introduced to use LLMs as an automated judger to streamline this process.

For each retrieved and processed dataset, the LLM is prompted with the subtask description, schema metadata, and representative samples. It evaluates whether the dataset satisfies the subtask's information need. If deemed relevant, the LLM generates a concise summary outlining the dataset's content, structure, and utility. If not, it suggests alternative data sources or follow-up strategies.

By automating judgment and fallback suggestion, this module reduces manual screening effort, enhances interpretability, and improves system robustness in low-coverage scenarios.

3.5 Visualization Module

We include a visualization module that enables users to interactively explore complex scientific

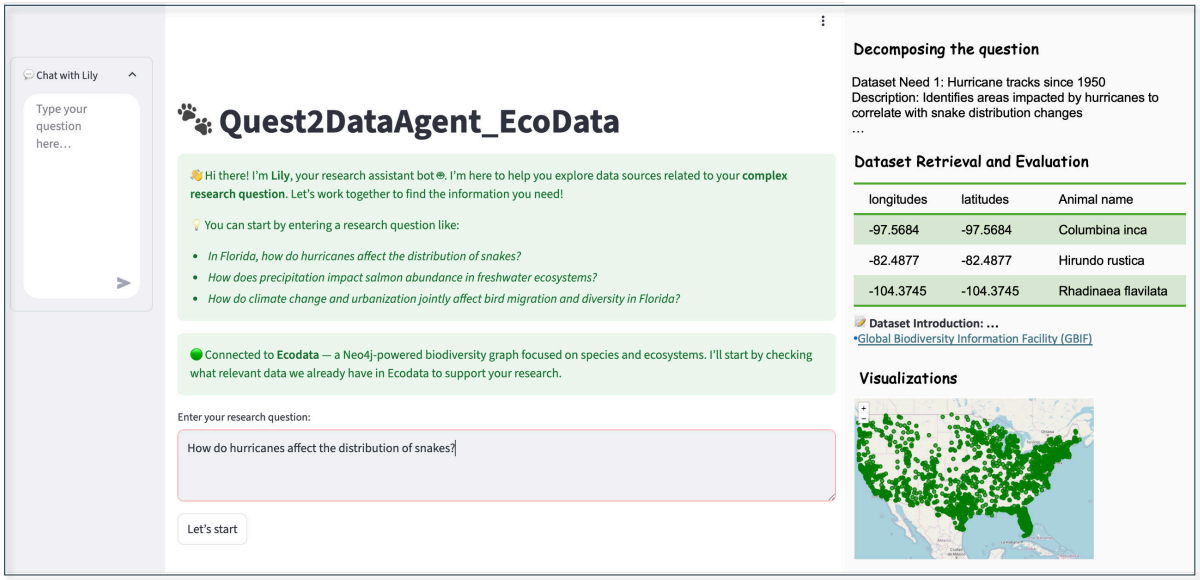

Figure 3: Interface of EcoData.

data. Placing this module in the middle of the pipeline allows for early inspection and validation of intermediate results. This ensures that potential errors, noise, or irrelevant data can be identified and corrected before proceeding further. Together with the conversational interface (introduced next), domain experts can assess data quality and iteratively refine their queries, improving overall efficiency and result accuracy.

For each processed dataset, the system inspects schema metadata and sampled rows, prompting the LLM to recommend an appropriate visualization type and generate Python-based code (e.g., via `Streamlit` and `pydeck`). The module supports diverse formats such as bar charts, line plots, geospatial maps, and molecular structure diagrams, depending on the data characteristics.

Quest2DataAgent features a conversational interface, enabling natural language interaction throughout the data construction workflow (see Fig. 3). It is context-aware, incorporating current subtask states and evaluation outcomes to support multi-turn dialogue. Users can iteratively refine queries, request clarifications, and explore related tasks, with conversation history maintained for coherence.

3.6 System Optimization

To support responsive and scalable interaction, the system applies lightweight optimization, including caching for repeated operations and a modular architecture with fallback logic to ensure robustness.

4 Demos of EcoData and PolyData

To demonstrate the versatility and domain-adaptability of Quest2DataAgent, we build two prototype systems: **EcoData** (Ecology)[2] and **PolyData** (Materials Science)[3]. These systems share the same backend and workflow engine but differ in domain-specific plugin datasets and user interfaces. Together, they showcase the ability of Quest2DataAgent to generalize across scientific domains while maintaining high usability and retrieval effectiveness. They are implemented as web-based systems using Streamlit, deployed on Hugging Face Spaces, and open-sourced under the MIT License. They are powered by GPT-4o (Hurst et al., 2024) and support modular integration of other foundation models (e.g., LLaMA (Touvron et al., 2023), Qwen (Bai et al., 2023)) and FAISS (Douze et al., 2024) for vector-based retrieval.

4.1 Demonstration of EcoData

EcoDatais a domain-specific demonstration of Quest2DataAgent framework for ecological research. It integrates a structured knowledge graph containing over 172,000 species observations (from GBIF (Global Biotic Interactions (GloBI), 2024), USGS, iNaturalist), 1,932 interspecies relationships, and environmental events (e.g., hurricanes, wildfires) from NOAA and KnowWhereGraph. Each entity is enriched with Wikidata and IUCN (International Union for Conservation of Na-

[2] https://huggingface.co/spaces/tyang4/Quest2DataAgent_EcoData
[3] https://huggingface.co/spaces/tyang4/poly_retrivel

ture, 2024) identifiers. The system incorporates both Neo4j-based graphs and structured tabular datasets (e.g., observation logs, climate reports). Fig. 3 shows the interface of the EcoData demo (more demo screenshots in Appendix Figure 7)

4.2 Demonstration of PolyData

PolyData is another domain-specific deployment of Quest2DataAgent framework tailored for polymer research. It comprises around 12,800 polymer entries sourced from PolyInfo (Ishii et al., 2024) and open literature, annotated with IUPAC names, SMILES strings, and polymer classes. Roughly 4,000 entries include polymerization metadata (e.g., monomer names, reaction types), and 1,000 contain molecular dynamics-based property data (Ishii et al., 2024) (e.g., heat capacity, dielectric constant, thermal conductivity). Appendix Fig. 5 shows the interface, and a complete demonstration screenshot is provided in Fig. 6 in the appendix. More details are available in Appendix A.2.

5 Evaluation Settings and Results

5.1 Experimental Setup

We conduct a comprehensive evaluation on both EcoData and PolyData to assess their effectiveness in supporting scientific data collection workflows. We recruited 18 domain researchers, with 9 participants each from ecology and materials science. This group includes 13 PhD students (72%) and 5 postdocs (28%).

Metrics: All participants complete a standardized task, record time spent, and respond to a questionnaire consisting of 12 Likert-scale (1–5) items and open-ended feedback forms. The 12 questions cover the following six evaluation criteria: M1. Task Decomposition Accuracy: Whether the system generates appropriate sub-questions, M2. Data Relevance and Coverage: Usefulness and completeness of retrieved datasets, M3. Evaluation and Recommendation: Effectiveness of automated dataset evaluation and external resource suggestion, M4. Cleaning and Visualization Quality: Clarity of structured output and informativeness of generated charts, M5. User Experience: Ease of use, responsiveness, and satisfaction, M6. Time Efficiency: Task completion time.

Baselines: As there is currently no existing system that supports end-to-end scientific data collection, we compare our framework against two representative baseline methods, and Table 1 summarizes the key capabilities across all approaches.

Capability	Manual	Tool-assisted	Ours
Task Planning	✗	✓	✓
Retrieval	✗	✗	✓
Evaluation	✗	Partial	✓
External Suggestion	✗	Partial	✓
Preprocessing	✗	✗	✓
Visualization	✗	✗	✓
Dialogue Interface	✗	✓	✓
End-to-End Workflow	✗	✗	✓

Table 1: Comparison of capabilities across baseline methods and Quest2DataAgent.

1. Manual workflow: Researchers manually break down the research question, search for datasets across multiple platforms, clean the data in spreadsheets, and create visualizations.

2. Tool-assisted: Participants use tools like LLMs or domain-specific platforms to assist with dataset search. However, they still manually assess relevance, clean data, and generate visualizations.

5.2 Results

Quantitative Results and Analysis: As shown in Figure 4, our system consistently achieves higher scores than both baselines across all aspects.

On PolyData, Quest2DataAgent obtains the highest ratings in nearly every aspect. It performs especially well in task decomposition, achieving an average score of 4.87 compared to 3.17 for Baseline 1 (manual workflow) and 4.36 for Baseline 2 (tool-assisted). It also shows statistically significant improvements in dataset retrieval, recommendation, preprocessing clarity, and user interaction. In particular, its user interaction score is more than three points higher than that of Baseline 1, indicating a much smoother and more intuitive experience. Perceived efficiency is also rated highly by users.

On EcoData, the system maintains strong performance. It achieves the highest score in evaluation and recommendation, scoring 4.47 compared to 2.54 for Baseline 1 and 3.79 for Baseline 2. It also receives consistently higher ratings in preprocessing and user interaction. Although the relative improvements are smaller than those observed on PolyData, Quest2DataAgent remains the top-performing system across all dimensions.

User Feedback Analysis: In addition to quantitative results, we collected user feedback comparing the Manual Workflow, Tool-Assisted Workflow, and Quest2DataAgent. Participants consistently praised Quest2DataAgent for its ability to decom-

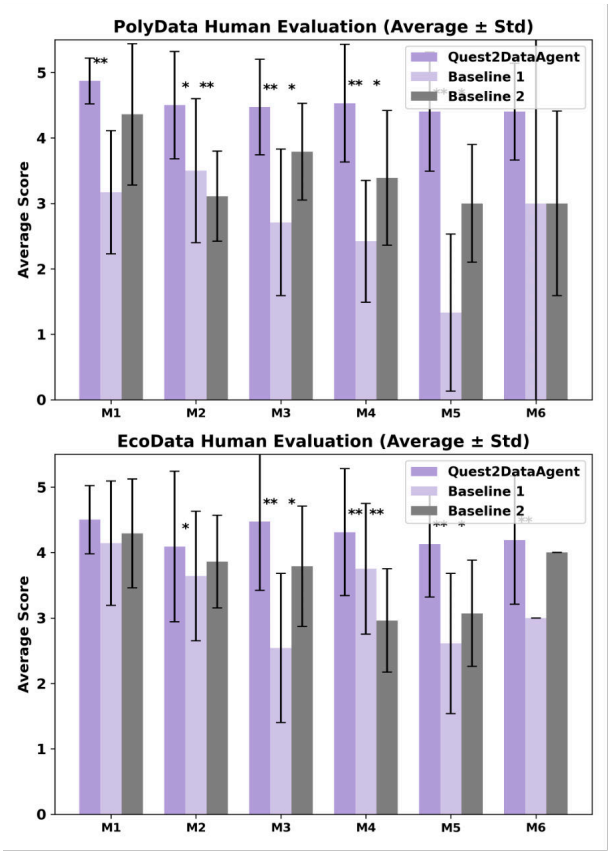

Figure 4: Quantitative human evaluation results on **PolyData** (top) and **EcoData** (bottom) with Baseline 1: manual workflow and Baseline 2: tool-assisted workflow. Bars show mean scores ± standard deviation. Asterisks denote significance based on two-sided paired t-tests ($p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***).

pose complex questions, retrieve relevant datasets, and present intuitive visualizations. Users noted that it "quickly helped obtain and filter datasets" and "enabled efficient screening," particularly benefiting those unfamiliar with domain-specific data collection.

In contrast, manual workflows were described as time-consuming and fragmented, requiring domain expertise and repetitive integration across platforms. Tool-assisted workflows helped with dataset suggestions but still required iterative prompting, manual filtering, and source verification. Users also raised concerns about hallucinations and lack of transparency.

Overall, the feedback highlights that Quest2DataAgent offers a more streamlined and guided experience, significantly reducing user effort while improving clarity and efficiency.

Statistical Significance: Paired two-sided t-tests confirm that the observed improvements are statistically significant in multiple aspects, particularly task decomposition, recommendation, visualization clarity, and user interaction. These results demonstrate the system's effectiveness in improving usability and reducing manual effort.

Method	EcoData Time (min)	PolyData Time (min)
Ours	13.06	7.38
Baseline 1	488.00 (~8.1h)	6330.00 (~105.5h)
Baseline 2	49.80	1724.58 (~28.7h)

Table 2: Average time spent to complete tasks under different methods in Eco and Poly domains.

Time Efficiency: We compare the time required for scientific data collection using Quest2DataAgent, manual workflows, and tool-assisted methods. As shown in Table 2, Quest2DataAgent substantially reduces time cost in both domains. In EcoData, our system averages 13.06 minutes, compared to 488 minutes for manual and 49.8 minutes for tool-assisted workflows. In PolyData, Quest2DataAgent averages 7.38 minutes, while Baseline 1 and Baseline 2 require 105.5 and 28.7 hours, respectively. These results demonstrate that Quest2DataAgent greatly improves time efficiency, reducing researcher workload by orders of magnitude without compromising data quality or task completeness.

6 Conclusion

We present Quest2DataAgent, a modular multi-agent framework for automating end-to-end scientific data collection. By integrating task decomposition, hybrid retrieval, evaluation, and visualization, it significantly reduces manual effort in dataset construction. We demonstrate its versatility through two domain-specific instances: EcoData for ecology and PolyData for materials science. Both share a common framework but operate on domain-adapted datasets and interfaces. Human evaluations confirm its effectiveness in improving data quality, usability, and efficiency.

Acknowledgement

This work is supported by the National Science Foundation (No: 2333795).

Broader Impact

Quest2DataAgent lowers the barrier to scientific data collection by enabling users to generate high-quality datasets from natural language queries. It supports researchers with limited technical expertise and promotes wider access to data-driven research. We demonstrate its applicability in two domains: ecology (EcoData) and materials science (PolyData). These demonstrations highlight the framework's flexibility and potential for broader scientific use.

While the system improves efficiency, it may lead to overreliance on automated outputs. To address this, we provide transparent workflows and encourage human oversight. Quest2DataAgent contributes to more accessible, reproducible, and efficient scientific research.

Looking ahead, we plan to incorporate dynamic tool use and Web API integration, enabling agents to autonomously access external sources and better adapt to evolving data availability.

References

Samee Arif, Sualeha Farid, Abdul Hameed Azeemi, Awais Athar, and Agha Ali Raza. 2024. The fellowship of the llms: Multi-agent workflows for synthetic preference optimization dataset generation. *arXiv preprint arXiv:2408.08688*.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Haihua Chen, Huyen Nguyen, and Asmaa Alghamdi. 2022. Constructing a high-quality dataset for automated creation of summaries of fundamental contributions of research articles. *Scientometrics*, 127(12):7061–7075.

Pei Chen, Boran Han, and Shuai Zhang. 2024. Comm: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. *arXiv preprint arXiv:2404.17729*.

Emily Y Chew. 2023. Publication of datasets, a step toward advancing data science. *Ophthalmology Science*, 3(3):100381.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Victor Dibia, Jingya Chen, Gagan Bansal, Suff Syed, Adam Fourney, Erkang Zhu, Chi Wang, and Saleema Amershi. 2024. Autogen studio: A no-code developer tool for building and debugging multi-agent systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–79.

Yixi Ding, Yanxia Qin, Qian Liu, and Min-Yen Kan. 2023. CocoSciSum: A scientific summarization toolkit with compositional controllability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 518–526, Singapore. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Global Biotic Interactions (GloBI). 2024. About global biotic interactions (globi). Accessed: 2024-09-13.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*.

Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. 2023. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. *Advances in neural information processing systems*, 36:45870–45894.

Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Chaowei Xiao, Jianfeng Gao, Lichao Sun, and 1 others. 2024. Datagen: Unified synthetic dataset generation via large language models. In *The Thirteenth International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

International Union for Conservation of Nature. 2024. The iucn red list of threatened species. Accessed: 2024-09-20.

Masashi Ishii, Takuro Ito, Hiroko Sado, and Isao Kuwajima. 2024. Nims polymer database polyinfo (i): an overarching view of half a million data points. *Science and Technology of Advanced Materials: Methods*, 4(1):2354649.

Yanhao Jia, Xinyi Wu, Hao Li, Qinglin Zhang, Yuxiao Hu, Shuai Zhao, and Wenqi Fan. 2025. Uni-retrieval: A multi-style retrieval framework for stem's education. *arXiv preprint arXiv:2502.05863*.

Max Ku, Thomas Chong, Jonathan Leung, Krish Shah, Alvin Yu, and Wenhu Chen. 2025. Theoremexplainagent: Towards video-based multimodal explanations for llm theorem understanding. *arXiv preprint arXiv:2502.19400*.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. Qasa: advanced question answering on scientific articles. In *International Conference on Machine Learning*, pages 19036–19052. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. 2024. Mlr-copilot: Autonomous machine learning research based on large language models agents. *arXiv preprint arXiv:2408.14033*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Minghao Liu, Zonglin Di, Jiaheng Wei, Zhongruo Wang, Hengxiang Zhang, Ruixuan Xiao, Haoyu Wang, Jinlong Pang, Hao Chen, Ankit Shah, and 1 others. 2024. Automatic dataset construction (adc): Sample collection, data curation, and beyond. *arXiv preprint arXiv:2408.11338*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey, 2024. *URL https://arxiv.org/abs/2406.15126*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

OpenAI. 2025. Deep research system card. Technical Report Version dated February 25, 2025, OpenAI. System card detailing capabilities and safety considerations of the Deep Research agent.

Iratxe Pinedo, Mikel Larrañaga, and Ana Arruarte. 2024. Arzigo: A recommendation system for scientific articles. *Inf. Syst.*, 122(C).

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024. Scaling large-language-model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

S Rasal. Llm harmony: multi-agent communication for problem solving (2024). *arXiv preprint arXiv:2401.01312*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Saptarshi Sengupta, Harsh Vashistha, Kristal Curtis, Akshay Mallipeddi, Abhinav Mathur, Joseph Ross, and Liang Gou. 2024. Mag-v: A multi-agent framework for synthetic data generation and verification. *arXiv preprint arXiv:2412.04494*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Ponzetto. 2024. ACLSum: A new dataset for aspect-based summarization of scientific publications. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6660–6675, Mexico City, Mexico. Association for Computational Linguistics.

Yang Tian, Fan Liu, Jingyuan Zhang, Yupeng Hu, Liqiang Nie, and 1 others. 2025. Core-mmrag: Cross-source knowledge reconciliation for multimodal rag. *arXiv preprint arXiv:2506.02544*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991, Florence, Italy. Association for Computational Linguistics.

Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. 2024b. Benchmark self-evolving: A multi-agent framework for dynamic llm evaluation. *arXiv preprint arXiv:2402.11443*.

Xiangqi Wang, Tianyu Yang, Jason Rohr, Brett Scheffers, Nitesh Chawla, and Xiangliang Zhang. 2025. Wildlifelookup: A chatbot facilitating wildlife management with accessible data and insights. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 1064–1067.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, and 1 others. 2025. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*.

Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. 2024. Kiwi: A dataset of knowledge-intensive writing instructions for answering research questions. *arXiv preprint arXiv:2403.03866*.

Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, and 1 others. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4).

Peiru Yang, Xintian Li, Zhiyang Hu, Jiapeng Wang, Jinhua Yin, Huili Wang, Lizhi He, Shuai Yang, Shangguang Wang, Yongfeng Huang, and 1 others. 2025. Heterag: A heterogeneous retrieval-augmented generation framework with decoupled knowledge representations. *arXiv preprint arXiv:2504.10529*.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023. Augmentation-adapted retriever improves generalization of language models as generic plug-in. *arXiv preprint arXiv:2305.17331*.

Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. 2025. Finragbench-v: A benchmark for multimodal rag with visual citation in the financial domain. *arXiv preprint arXiv:2505.17471*.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, Yang Xu, Qingkai Min, Zizhao Zhang, Yiwen Wang, Wenjie Li, and Pengfei Liu. 2024. OpenResearcher: Unleashing AI for accelerated scientific research. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 209–218, Miami, Florida, USA. Association for Computational Linguistics.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2023. Dyval: Dynamic evaluation of large language models for reasoning tasks. *arXiv preprint arXiv:2309.17167*.

A Appendix

A.1 System Interface

We show the interface of PolyData in Fig. 5, and provide screenshots of both PolyData and EcoData in Fig. 6 and Fig. 7.

A.2 PolyData demo

Fig. 5 shows the interface of the PolyData demo (more demo screenshots in Appendix Figure 6). The workflow begins when the user submits a research question (e.g., *"What functional groups are common in high-Tg polymers?"*). The system confirms the query, initiates automated analysis, and then decomposes the question into structured data needs, such as polymer structures, functional group annotations, and experimentally measured glass transition temperatures (Tg), each accompanied by a concise description. For each need, relevant datasets are retrieved from PolyData, evaluated for coverage and relevance, and presented in tabular form, with external links provided when necessary. Finally, the platform generates tailored visualizations, including SMILES-based dataset previews and chemical visualizations, to help users quickly identify functional group patterns and gain insights into polymer properties.

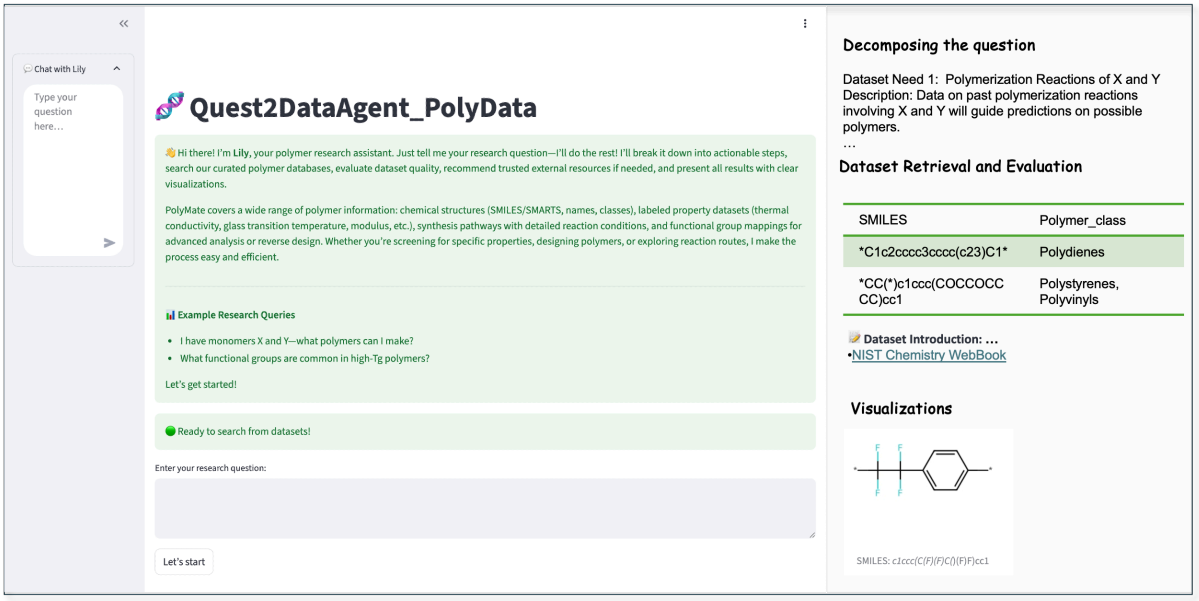

Figure 5: Interface of PolyData, an interactive web platform for polymer data exploration. In the main workspace, users input polymer-related questions to receive structured data decompositions, SMILES-based dataset previews, and chemical visualizations. The left panel enables interactive conversation, while the right panel presents retrieved data, metadata, and molecular structure diagrams.

A.3 Prompt Templates

We incorporate Chain-of-Thought (CoT) (Wei et al., 2022) prompting in our design. We present the full prompt templates for each agent module below.

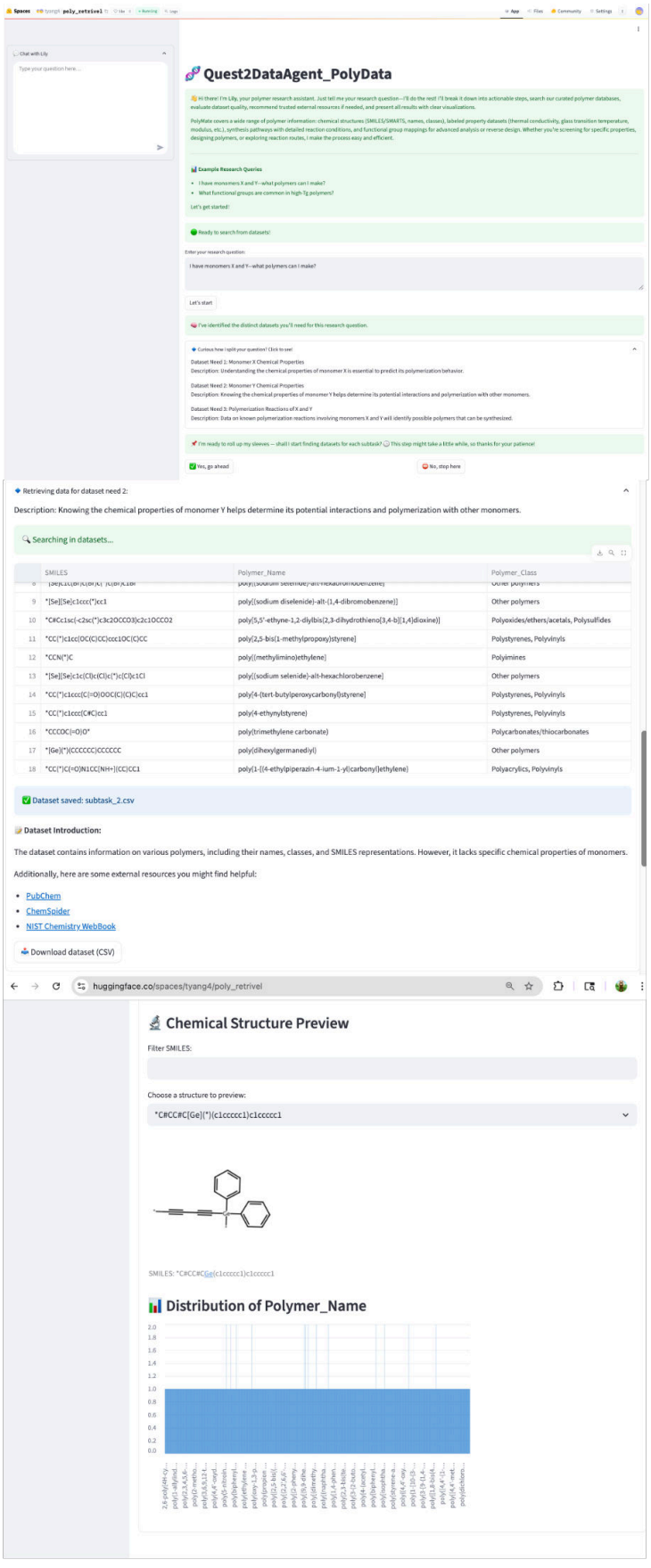

Figure 6: Screenshot of PolyData

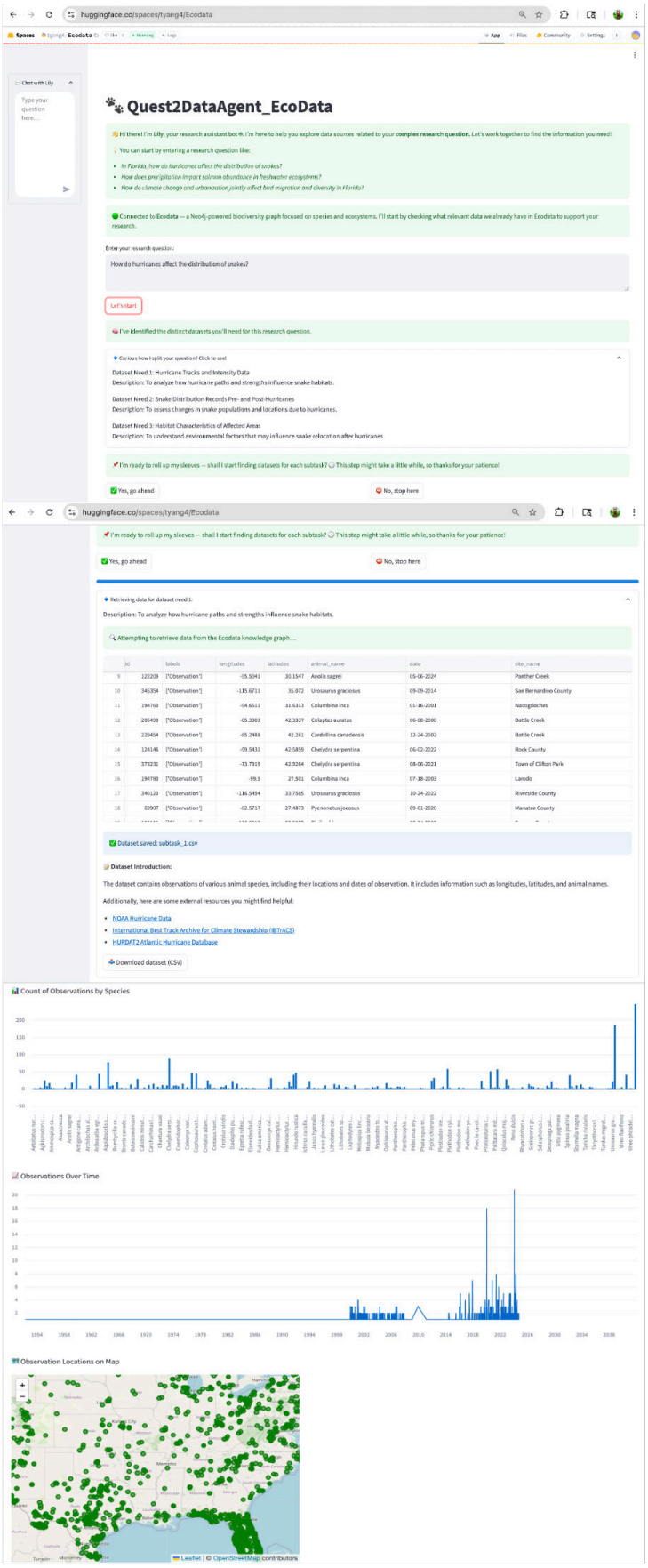

Figure 7: Screenshot of EcoData

Planner Agent Prompt Template

Role: Research-data planning assistant
Task: List the separate datasets a researcher must collect to answer the research question below. Each dataset should focus on one clearly defined entity or phenomenon (e.g., "Tracks of hurricanes affecting Florida since 1950," "Geo-tagged snake observations in Florida 2000–present").
Output Format:

- Write 1–6 blocks. For each block, use both lines exactly:
- `Dataset Need X: <Concise title, ≤10 words>`
- `Description: <Why this data matters – 1 short sentence>`
- **Do NOT** add extra lines or markdown.
- **Keep** variable names short; no code blocks; no quotes.

Research Question: `{question}`

Retriever Agent Prompt

Role: Cypher query generator for a Neo4j biodiversity database
Node Types and Properties:

- Observation: `animal_name`, `date`, `latitude`, `longitude`
- Species: `name`, `species_full_name`
- Site: `name`
- County: `name`
- State: `name`
- Hurricane: `name`
- Policy: `title`, `description`
- ClimateEvent: `event_type`, `date`
- TemperatureReading: `value`, `date`, `location`
- Precipitation: `amount`, `date`, `location`

Relationship Types:

- `(Observation)-[:OBSERVED_IN]->(Site)`
- `(Observation)-[:OBSERVED_ORGANISM]->(Species)`
- `(Site)-[:BELONGS_TO]->(County)`
- `(Observation)-[:IN_COUNTY]->(County)`
- `(County)-[:IN_STATE]->(State)`
- `(Species)-[:interactsWith]->(Species)`
- `(Species)-[:preysOn]->(Species)`

Instructions:

- Generate a precise and efficient Cypher query for the subtask: `{subtask}`
- **Do NOT** return all nodes of a type unless explicitly requested.
- Use location filters (`IN_COUNTY`, `IN_STATE`, `BELONGS_TO`) if mentioned or implied.
- For taxonomic/common groups, filter with `CONTAINS/STARTS WITH` on `Species.name` or `species_full_name` with `toLower(...)`.
- Filter by date if a time range is included.
- Prefer `DISTINCT` to avoid redundant results.
- Return only fields needed for the subtask.

Return a JSON object with:

- `"intent"`: description of query purpose
- `"cypher_query"`: the Cypher query
- `"fields"`: e.g., `["species","county","date"]`

Dataset Evaluation Agent Prompt Template

Role: Data-validation assistant

===== TASK =====
Subtask: `{subtask}`

===== DATASET PREVIEW =====

- Schema (first `{len(selected_cols)}` columns): `{json.dumps(column_info, indent=2)}`
- Sample rows (max 3): `{json.dumps(sample_rows, indent=2)}`

===== OUTPUT INSTRUCTIONS (follow strictly) =====

A Relevant:

- Write exactly two sentences, each no more than 30 words.
- Summarize what the dataset contains and why it helps the subtask.
- Do not mention column names or list individual rows.

B Not relevant:

- Write one or two sentences (max 30 words each) describing only what the dataset contains.
- Do not mention the subtask, relevance, suitability, limitations, or missing information.
- After the sentences, output:
  `Additionally, here are some external resources you might find helpful:`
- Format output in markdown as: `- [Name of Source](URL)`
- List 2–3 bullet points, each on its own line, starting with `-` and a URL likely to contain the needed data.
- No additional commentary.

General rules: Plain text only—no code fences. Markdown link syntax (`[text](url)`) is allowed.

External Resource Recommender Prompt Template

Role: External resource recommender

Please recommend 3 reliable and relevant online datasets or websites that can help with the following subtask: `{subtask}`

Format your output in markdown as:

- `- [Name of Source](URL)`
- `- [Name of Source](URL)`
- `- [Name of Source](URL)`