Toward Resource Rational Dataset Search Interfaces

Danni Liu
University of Chicago
Chicago, Illinois, USA
danni6@uchicago.edu

Siyuan Xia
University of Chicago
Chicago, Illinois, USA
stevenxia@uchicago.edu

Raul Castro Fernandez
University of Chicago
Chicago, Illinois, USA
raulcf@uchicago.edu

Alex Kale
University of Chicago
Chicago, Illinois, USA
kalea@uchicago.edu

ABSTRACT

Data workers benefit from reusing existing datasets, but finding candidate data and judging whether it fits a task remain difficult. Because dataset fit depends on use, evaluation during discovery is often provisional. Through a survey (N=41), interviews (N=10), and a secondary analysis of an LLM-assisted discovery study (N=36), we examine how data workers evaluate candidate datasets under uncertainty. We find that evaluation is selective, task-conditioned, and cost-sensitive: workers use accessible signals to screen candidates, escalate effort when uncertainty becomes consequential, and proceed when remaining limitations are acceptable for the task. In the LLM-assisted setting, automation did not remove evaluation work; it changed which checks were easiest to perform and where evaluation effort was directed. These findings motivate discovery interfaces that align the cost of checking with the evidence needed to judge task-specific fit.

VLDB Workshop Reference Format:
Danni Liu, Siyuan Xia, Raul Castro Fernandez, and Alex Kale. Toward Resource Rational Dataset Search Interfaces. VLDB 2026 Workshop: DASHSys: Systems for Data-centric Agents with Human-in-the-loop.

VLDB Workshop Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/6danni/resource-rational-discovery-artifacts.

1 INTRODUCTION

Dataset discovery helps organizations make use of existing data for analysis, tools, and decision-making. Realizing these benefits requires fast and flexible ways of searching for datasets and assessing whether they meet a worker's information need [33]. In data management, an *information need* describes what a worker is seeking during discovery: what the data should contain, how it should be structured, and what conditions it must satisfy to be useful for a particular purpose [14, 27, 31, 33].

Satisfying an information need is difficult for reasons that are not only algorithmic. Data needs are hard to express through the keyword and SQL-style interfaces that discovery tools commonly provide [26]. Documentation is often sparse, inconsistent, or too generic to support task-specific judgments [20]. When search and documentation fall short, workers may contact data providers, colleagues, or other users, which can answer situated questions but introduces additional cost and delay [29]. These difficulties are compounded by the fact that workers often do not know in advance exactly what to ask for [9]. The requirements that make a dataset suitable may become clear only after inspecting candidates, comparing alternatives, or attempting to use the data for the intended purpose. Search and evaluation are therefore interleaved. Search and discovery systems may return datasets relevant to a query and provide documentation or previews for inspection, but workers revise their judgments as they engage with candidates. As a result, any evaluation during discovery remains provisional.

We interpret this provisional evaluation through a resource-rationality lens. Because dataset fit is often settled only in use, the evidence available during discovery is incomplete. Metadata, documentation, samples, social evidence, and LLM-generated summaries can all inform judgment, but each remains a proxy for whether the data will work for a specific task. Dataset evaluation can therefore be understood as the cost-sensitive allocation of limited effort under uncertainty: data seekers screen candidates with cheap signals, escalate to deeper checks when a candidate becomes promising or consequential, and proceed when remaining limitations are manageable for the task.

Recent LLM-assisted systems have shown strong capabilities in dataset discovery and data question answering [9]. These systems offer a timely setting for examining how evaluation changes when parts of discovery are automated. While such systems can lower the cost of finding data and producing answers, they do not eliminate the need for data seekers to judge whether the surfaced data and generated artifacts are appropriate for use. This motivates studying dataset evaluation both in current practice and in LLM-assisted discovery, where the interface can change which evidence is surfaced, which checks are easy to perform, and how workers allocate evaluation effort. Specifically, we ask:

- **RQ1.** How do data workers form and refine judgments of whether a dataset will fit a task during discovery?
- **RQ2.** How does an LLM-assisted discovery interface change how data workers evaluate dataset fit?

We address RQ1 through a survey of 41 data workers and in-depth interviews with 10 of them reflecting on prior dataset discovery experiences (Study 1). We address RQ2 through a secondary analysis

Proceedings of the VLDB Endowment. ISSN 2150-8097.

of a user study of an inspectable interface built on top of Pneuma-Seeker [9], an LLM-assisted system for data discovery and question answering (Study 2). Rather than treating the studies as a controlled comparison, we read them together to characterize dataset evaluation as selective, task-conditioned, and cost-sensitive.

Study 1 showed that data seekers evaluated datasets through a flexible, iterative process, beginning with accessible signals and investing in deeper checks when a candidate appeared promising enough to warrant further evaluation. Which criteria mattered depended on the intended use: process-oriented work emphasized whether the data could reliably support an automated workflow, analytical work emphasized whether it provided sufficiently complete and credible evidence for a conclusion, and communicative work foregrounded whether the dataset could support a clear and credible narrative for an audience. Dataset choice was therefore less a search for the best dataset overall than a judgment about whether the expected benefits justified the costs and whether remaining limitations could be managed.

Study 2 showed that LLM assistance changed where evaluation effort was directed. Participants worked from the datasets, intermediate artifacts, and answers surfaced by the interface, shifting effort from open-ended search toward checking whether the system had faithfully used the surfaced data to produce its answer. Dataset-level properties that mattered in Study 1, such as data collection methods and data completeness, were less consistently examined. Evaluation therefore became more targeted but also more bounded: effort concentrated on what the interface made easiest to check, consistent with the tendency to treat what is visible as the whole picture [28].

This paper makes two contributions. First, it offers a resource-rational account of dataset evaluation as the operationalization of information need under uncertainty. We describe how data seekers apply relevance, usability, and quality criteria unevenly as they allocate limited effort across imperfect signals of dataset fit. Second, we use this account to examine LLM-assisted discovery and identify a design hazard: when system-fidelity checks become cheap and visible, they can draw evaluation away from broader questions of dataset fit. Together, these findings suggest design directions for human-in-the-loop discovery systems: support evolving information needs, condition evaluation on intended use, surface evidence progressively, preserve traceability without narrowing fit judgments, and enable situated comparison across candidates.

2 RELATED WORK

Our study draws on research in data discovery systems, LLM-assisted data tools, and data search and exploration behavior.

2.1 Data Discovery Systems

Data discovery systems help users identify datasets that satisfy an information need [17, 18]. Much work in data management addresses well-defined discovery tasks, such as finding tables for union, join, augmentation, or downstream model improvement [19, 21, 30, 36, 46, 60]. These systems make dataset discovery computationally tractable by assuming that the user's need can be expressed through a known operation, objective function, or utility model. Our findings suggest that this assumption does not always hold in early-stage dataset search, where workers may still be learning what the task requires and which dataset limitations matter.

A second class of systems supports more open-ended discovery through keyword, metadata, or programmatic interfaces. Google Dataset Search, Auctus, Aurum, and Ver allow users to search or specify information needs in different ways [11, 12, 18, 21]. These systems are useful when their interface abstractions match how users can articulate their needs. However, prior work shows that dataset search remains difficult because needs can be hard to express, documentation is uneven, and users often need information beyond what search results expose [14, 26, 27, 33]. Our work complements this literature by examining how workers evaluate candidates once they are surfaced, and why they apply different criteria selectively across tasks.

2.2 LLM-Assisted and Agentic Data Tools

A growing class of systems uses large language models to reduce the cost of expressing and executing data tasks. Natural-language discovery systems retrieve tables from free-form questions, LLM-based preparation systems automate transformation and wrangling, and table question-answering systems answer queries over individual tables [16, 39, 55, 56, 59]. Pneuma-Seeker [9], the system underlying Study 2, integrates discovery and preparation into an agentic pipeline that returns a draft answer together with the tables, operations, and scripts behind it.

These systems lower the cost of producing an answer, but they also shift evaluation work. Users still need to judge whether the surfaced data, intermediate operations, and final answer are appropriate for the task [13]. HCI and data system work has developed interfaces for inspecting text-to-SQL output, grounding queries with explanations, steering agentic analysis, and debugging query errors [45, 47, 52, 57]. Most of this work focuses on verifying the correctness of a query, operation, or answer. Our Study 2 asks how users evaluate multi-step system-mediated discovery, where the issue is not only whether one operation is correct, but whether the surfaced evidence is an appropriate basis for trusting the result.

2.3 Data Exploration and Search Behavior

Visual data exploration systems support interactive inspection and sensemaking once data has been loaded, from early systems such as Polaris/Tableau to recommendation-driven tools such as SeeDB, Voyager, and Draco [40, 43, 51]. These systems offer richer interaction than most discovery interfaces, but they typically assume that users have already selected data to inspect. Some systems blur this boundary by embedding previews, profiling, or visual answers into search workflows [25, 38, 54]. This line of work motivates discovery interfaces that help users move more fluidly between search and lightweight evaluation.

Empirical studies in HCI and data management characterize data discovery and exploration as iterative sensemaking, where analysts refine their understanding through interaction with data and contextual evidence around it [31, 41]. Prior work has also characterized dataset search behaviors, search criteria, and the limits of current discovery tools [14, 22, 26, 27, 32, 33]. We build on this work with a resource-rational account of dataset evaluation. Drawing on resource rationality [23] and information foraging [48], we examine

Table 1: Participant backgrounds. Interviewees are in bold.

Background	IDs
Data analysts, data scientists, consultants, and data architects	P_8, P_{11}, P_{13}, $\mathbf{P}_{15}$, P_{16}, P_{18}, P_{21}, P_{25}, $\mathbf{P}_{34}$, $\mathbf{P}_{36}$, $\mathbf{P}_{38}$
Software engineers	P_1, P_{14}
Research scientists (academia)	P_6, P_7, P_9, P_{10}, P_{29}, P_{30}, $\mathbf{P}_{31}$, $\mathbf{P}_{35}$, P_{37}, P_{39}, P_{40}, P_{41}
Research scientists (industry)	P_2, P_3, P_4, P_5, P_{12}
Entrepreneurs and business leaders	P_{17}, P_{19}, P_{20}, P_{22}, $\mathbf{P}_{23}$, P_{26}
Data journalists, visualization designers, and media professionals	P_{24}, $\mathbf{P}_{27}$, $\mathbf{P}_{28}$, $\mathbf{P}_{32}$, P_{33}

how workers allocate limited verification effort across available signals of fit, including documentation, provenance, documented use, preview, and other contextual cues. Rather than treating relevance, usability, and quality as a fixed checklist, we analyze how these criteria become active differently depending on intended use, uncertainty, and the cost of checking.

3 METHODS

We draw on two qualitative analyses to understand how data workers source and evaluate candidate datasets and how this evaluation changes when discovery is mediated by an LLM-assisted interface. Study 1 examines dataset discovery in current practice through a survey and follow-up interviews. Study 2 reanalyzes qualitative data from a separate user study of an LLM-assisted discovery interface built on top of Pneuma-Seeker [9]. Codebooks and additional study materials are available in the online supplemental materials.

3.1 Study 1: Current Dataset Discovery Practice

Study 1 examined current dataset discovery practice through a platform review, a survey, and follow-up interviews. The platform review grounded our prompts in current interface features. The survey captured broader patterns across participants' discovery experiences, while the interviews elicited more open-ended accounts of workers' reasoning processes and interface needs.

3.1.1 Preliminary Review of Data Discovery Platforms. We reviewed seven data platforms spanning three provider types: commercial data marketplaces, open community platforms, and government open-data portals. The platforms included Snowflake [7], AWS Data Exchange [1], Datatrade [3], Kaggle [5], Hugging Face [4], Data.gov [2], and NYC Open Data [6]. On each platform, we recorded features that could support dataset evaluation, such as documentation, samples, profiling tools, usage signals, reviews, ratings, provider contact mechanisms, and example use cases. These features grounded our survey questions and interview prompts about how data workers assess candidate datasets using current tools.

3.1.2 Participants. We recruited 41 survey participants, each with at least two years of data work experience. Participants came from a range of roles, including data analysis, data science, consulting, software engineering, research, journalism, visualization design, and media work. Table 1 summarizes participant backgrounds, with follow-up interview participants marked in bold.

3.1.3 Survey. The survey asked participants to describe concrete dataset search experiences. Each participant reported one successful case and one unsuccessful case.

For each experience, we asked about the participant's objective, where they obtained candidate datasets, how they assessed them, and what information supported that assessment. We also asked whether the available information was sufficient to judge task fit, whether the data seemed adequate in quality, and whether the dataset would have been worth paying for if it was not free.

3.1.4 Interviews and Co-Design Activity. The follow-up interviews extended the survey by eliciting more detailed accounts of participants' discovery processes, evaluation strategies, and expectations for tool support. Each interview included a retrospective account, a co-design activity, and a debriefing interview. Participants first described a recent dataset search, including the process they followed, the tools they used, and the information they relied on. In the co-design activity, they imagined assessing several candidate datasets for a data-driven task. We suggested common interface elements from current discovery platforms, including documentation, samples, profiling tools, reviews, usage information, access information, and contact channels. Participants selected elements they would use, explained how they would use them, suggested missing elements, and walked through an imagined evaluation workflow. The debriefing focused on how participants assess dataset usefulness, define data quality, reason about dataset value, and use visualizations to evaluate data.

3.1.5 Study 1 Analysis. We analyzed the survey responses and interview transcripts separately through iterative qualitative coding. For each source, the first author open-coded an initial subset and discussed emerging codes and themes with the research team to develop and refine a codebook. The first and second authors then divided the remaining responses and transcripts for coding. After coding was complete, the first author reviewed all coded materials, discussed ambiguous or inconsistent code applications with the second author, and revised the codebooks and coding as needed. The first author also periodically discussed emerging findings and supporting evidence with the research team.

The survey codebook covered task types, evaluation criteria, information sources, and search practices. The interview codebook focused on search strategies, discovery difficulties, information access, interface elements, and participants' rationales for deciding whether a dataset was worth pursuing. We treated the two analyses as complementary and synthesized their findings in Section 4.1.

3.2 Study 2: Dataset Discovery with an LLM-Assisted Interface

Study 2 examines how dataset evaluation changes when discovery is mediated by an LLM-assisted interface. We reanalyzed qualitative data from a separate user study of an interface built on Pneuma-Seeker [9]. The study included 36 participants with prior experience using data, whom we refer to as U1–U36. Our analysis focuses on how participants evaluated system-surfaced datasets, intermediate artifacts, and draft answers.

3.2.1 Interface and Task Setting. Pneuma-Seeker is an LLM-assisted interface for data discovery and question answering over relational

data. Participants used Pneuma-Seeker in two conditions: in the baseline condition, they interacted only through its conversational interface. In the comparison condition, we added an interface layer that displayed intermediate steps, code, tables, and source data in a sidebar and linked the draft answer to these supporting artifacts.

The tasks came from KramaBench and concerned legal and environmental domains [34]. For example, one task asked participants to assess consistency across sources in FTC consumer fraud data, where reported counts were aggregated from multiple reporting pipelines. Participants were asked to report both the system-generated answer and their own answer, and they were free to browse online or consult external resources. This open setting allowed us to observe how they checked system outputs and candidate data in ways that more closely resembled real-world practice.

3.2.2 Study 2 Analysis. We conducted a qualitative reanalysis of the think-aloud data collected as participants completed the Study 2 tasks. Our goal was to examine how participants evaluated system-surfaced datasets, intermediate artifacts, and draft answers, and how these practices differed from unassisted discovery in Study 1.

The first author reviewed an initial subset of the think-aloud data to develop preliminary codes, which were refined through discussion with the research team. The first author then coded all think-aloud data and periodically discussed emerging findings and supporting evidence with the team. The analysis focused on how participants interacted with and checked the interface's outputs, as well as comments about LLM trust, including concerns about fabrication and whether outputs could be checked against the data. We then compared these patterns with the account developed in Study 1 and report the resulting findings in Section 4.2.

4 FINDINGS

We organize the findings in two parts. First, we characterize how data workers evaluate candidate datasets in current discovery practice. Second, we examine how these evaluation practices appear in an LLM-assisted setting where candidate datasets, intermediate artifacts, and draft answers are surfaced by the system. Reading the two studies together, we find a stable pattern across settings: data seekers ration limited effort under uncertainty. Interface mediation changes this pattern by making some checks easier than others, shaping where evaluation effort concentrates.

4.1 Study 1: Current Dataset Discovery Practice

Within Study 1, five patterns recurred. Discovery was iterative because workers often had to clarify both what they needed and what candidate datasets could support. Task context shaped which evaluation criteria became active. Cheap signals gated costlier verification. Social communication supplied situated evidence that search and documentation could not provide. Finally, selection centered on finding a workable dataset whose remaining costs and limitations were acceptable, rather than the best dataset in general.

Discovery Was Iterative Because Fit and the Need Remained Provisional. Dataset discovery tools often assume that workers can state an information need and match it to candidate datasets. Our participants' accounts showed that this assumption often did not hold: both the requirement and the dataset's fit became clearer only through engagement with candidate data. This pattern is consistent with prior empirical work characterizing dataset discovery as iterative and exploratory [10, 26, 31, 33]. Our findings explain this iteration as a response to uncertainties that are costly to resolve upfront. Data seekers were not only refining search terms, they were jointly resolving two uncertainties: what a candidate dataset contained, and what use that data could support.

Participants moved among recurring activities: sourcing candidate datasets, screening them through visible signals, inspecting promising candidates more closely, and deciding whether remaining limitations could be repaired, disclosed, worked around, or tolerated. These activities formed a loop rather than a pipeline [10, 33]. A failed screening check could send participants back to sourcing, while problems found during deeper inspection could change what they considered relevant, usable, or worth pursuing.

One driver of this iteration pattern was that information needed to assess fit often was not exposed at search time. Search results and dataset pages might show titles, descriptions, providers, or samples, but participants often needed more specific evidence. P_{31}, for example, needed column-level information to judge whether a dataset contained the variables their task required. That information was available only inside the CSV file rather than in searchable metadata, and the column names sometimes were too abstract to interpret without a separate dictionary [20, 29]. As P_{31} put it, *"It took me a lot of efforts to understand the column. But if I don't understand the column, I don't understand this distribution."* The evidence existed, but accessing and interpreting it required additional work.

A second driver was that participants often could not specify the requirement before seeing candidate data. Their sense of what would count as useful data formed through engagement with available datasets [10]. P_{34} described skimming Kaggle projects to see what others had done before deciding what approach to take: *"I would just search Kaggle [for] existing projects along that domain just to see what other professionals are doing."* P_{36}, who was trying to improve a fraud-detection model, had to examine candidate datasets alongside the model's existing data before they could tell what additional data might help. Storytelling and visualization tasks showed a similar pattern, where fit depended on whether the data held a trend worth building a story around. P_{28} said *"I would make a chart before I decide to use it, because I don't know if the choice will be interesting to visualize it,"* while P_{32} described downloading multiple datasets and testing visually which one fit the case.

Taken together, these findings show that dataset discovery was iterative because participants were learning both what to look for and what the available dataset space contained.

Task Context Determined Which Search Criteria Gained Active Attention. Search and ranking systems often evaluate candidates against general notions of relevance and quality. However, our findings suggest participants did not weigh these dimensions uniformly. To characterize how participants evaluated datasets, we coded the criteria they described against Koesten et al.'s taxonomy of relevance, usability, and quality sub-criteria [33]. Working bottom-up from participants' own descriptions, we adapted several sub-criteria to better match the concerns they raised. Table 2 lists the resulting criteria alongside the original terms.

Table 2: Evaluation criteria adapted from Koesten et al.'s information needs for selecting datasets [33]. We grounded terms in participants' own descriptions. † marks a criterion new to our coding, with no counterpart in the original taxonomy; "replaces" marks a renamed criterion; the remainder are retained.

Criterion	Description
Relevance	
context	the broader setting or domain the data describes
coverage	whether the data spans the scope the task needs
required attributes†	whether the dataset contains the specific variables the task requires
original purpose	why the data was originally collected
documented use†	examples of prior use (e.g., notebooks, prior analyses)
community adoption†	whether the dataset is widely used, peer-reviewed, or endorsed
granularity	the level of detail or aggregation
preview	an inspectable view of the data (replaces *summary*)
time frame	the period the data covers
Usability	
labeling	clarity of field names and labels
documentation	availability and clarity of documentation
license	terms of use
cost†	monetary cost of access
access	how the data can be obtained
tool compatibility	whether the data works with the participant's tools (replaces *machine readability*)
language used	the natural language the data is written in
format	the data's file or encoding format
schema	the structure of the data
ability to share	whether the data can be redistributed
provider contactability†	whether the data provider can be reached
Quality	
collection methods	how the data was gathered
provenance	the data's source and history
validation evidence	evidence the data is accurate or has been checked (replaces *consistency of formatting/labeling*)
completeness	the extent of missing data
what has been excluded	what the data omits

Table 3: Task groups. We divide [33]'s goal-oriented category into *analytical* and *communicative* tasks, distinguished by what consumes the data.

Group	What consumes the data	*n*
Process-oriented	A computational workflow that cannot easily adapt to the data (e.g., training a model, building a pipeline).	18
Analytical	A person whose conclusions rest on the data (e.g., exploratory analysis, answering a bounded question).	37
Communicative	An audience that consumes the output, not the dataset (e.g., data journalism, storytelling, visualization).	17

We also build on Koesten et al.'s distinction between process- and goal-oriented tasks. We retained the process-oriented category, but found that goal-oriented tasks separated into two forms with different evaluation patterns: *analytical* tasks, where data supported an analysis and its conclusions, and *communicative* tasks, where data was visualized or reported to an audience. We therefore grouped tasks into three categories: process, analytical, and communicative. Table 3 summarizes these categories.

Across these task groups, participants did not apply a single fixed checklist to every candidate dataset. Task context determined which criteria gained active attention and which could remain secondary.

For process-oriented uses, such as model training or data pipelines, participants foregrounded criteria that affected whether the data could enter and support an automated workflow. Schema, collection methods, and distribution mattered because problems in these properties could propagate into downstream outputs. P_{34}, describing their modeling practice, treated statistical adequacy as a precondition for continuing: *"if the data is not prepped [and] statistically valid to some degree, then I can't even go further with that."* In these tasks, a dataset could be topically relevant yet still be unsuitable if it could not be processed reliably.

For analytical tasks, participants foregrounded criteria that affected whether the data could support a conclusion. They attended to whether the dataset contained the required columns, how complete the data was, how missingness was handled, how the data was collected, and whether the source was credible. For these tasks, the central question was whether the data provided sufficiently complete and credible evidence to support the intended analysis.

For communicative tasks, such as reporting, storytelling, or public-facing visualization, participants foregrounded criteria that affected whether the data could support a clear and defensible communication. They considered whether it revealed a pattern worth presenting, whether variables were clearly labeled, and whether limitations could be explained to an audience. P_{28} cared less about pristine data than about being able to account for it: *"as long as you clearly denote what you're measuring and what you're not measuring, [it's] worth writing about in the news."* In these tasks, limitations in what the data measured or represented did not necessarily disqualify it if those limitations could be clearly communicated.

These findings showed that dataset fit was task-conditioned rather than intrinsic. Process tasks emphasized workflow compatibility and computational constraints, analytical tasks emphasized evidentiary requirements, and communicative tasks emphasized presentability and explainability. Dataset discovery systems should accordingly prioritize evaluation criteria based on inferred or elicited task context rather than present them as a flat checklist.

Costly Verification Became Worthwhile Only After Cheap Signals Indicated Potential Fit. While interfaces tend to expose similar evidence for each candidate dataset, participants did not evaluate every candidate dataset with the same depth. Consistent with prior work that treats relevance as a prerequisite for further dataset evaluation [33], they first used cheap, visible signals to decide whether a candidate was worth deeper inspection. Titles, descriptions, sources, column names, and previews acted as early gates. P_{31}, for instance, used the sample as a quick screen *"I like to see a sample of the data set just to know this is really what I want, and also get a judgment about the quality of the data set"*. Such signals filtered candidates quickly rather than confirming that a dataset would actually fit the task.

Costlier checks became worthwhile only after a dataset appeared promising enough to pursue or uncertainty became consequential.

This pattern was especially visible in participants' use of documentation. Prior work emphasizes complete and standardized documentation as a means of supporting dataset reuse [20, 29], but participants used it in two distinct ways. During initial screening, the presence of documentation served as a low-cost signal of credibility. P_{31} found it reassuring even when its contents were not directly useful: *"although they are not super useful, but it gave me an impression that this data set is very well documented."* Later, close reading became a targeted verification step. P_{38} consulted documentation when something did not add up: *"when I go to documentation or files, I should have some misunderstanding about the dataset."* Documentation thus functioned both as an early credibility signal and as a resource for resolving specific uncertainties during deeper evaluation. Effort also tracked stakes and time pressure: P_{28}, a journalist, preferred to verify by talking to people, but under deadline relied on available documentation to understand the data quickly.

Selective documentation use was therefore not simply a consequence of missing or poorly designed documentation. It was part of how participants managed the cost of evaluation. Better documentation can lower the cost of deeper verification, but it does not follow that workers will read full documentation for every candidate.

Social Communication Helped Resolve Situated Questions That Neither Search or Documentation Could Answer. Keyword search and documentation present information prepared in advance, making them informative signals for a wide range of tasks. Yet, situated answers are still indispensable in many cases. Prior work characterizes social communication in dataset work as a complement to incomplete documentation [22, 29, 31] and as a mechanism through which trust develops [58]. Our findings show a related role: participants turned to other people to resolve questions that were specific to their task, organization, or domain.

Participants turned to colleagues, prior users, and data providers when platform-visible information was insufficient. P_{15} contacted a known provider to ask whether they had collected a particular kind of data that was unlikely to appear elsewhere. P_{38} found keyword search over an internal database difficult, but a colleague could point them to the specific dataset directly. P_{27} and P_{28} talked through their goals with colleagues who had written on similar topics to decide where to look and what evidence to trust. As P_{28} put it, *"I think chat and contact is better if I can easily access the person who publishes this data. I always learn more from that conversation than I do through things like a sheet."*

In these cases, social communication functioned as an information source on its own. It helped participants obtain situated, interactional evidence about whether a dataset fit a use. At the same time, this support was uneven: it depended on knowing and having access to who to ask, and having time to seek their input.

Selection Sought a Workable Dataset, Not the Best One. Discovery systems often aim to surface the most relevant or highest-quality dataset [14]. Our findings complicate this framing. At selection time, participants did not choose the highest-quality dataset in the abstract. They chose a candidate whose remaining costs and limitations were manageable for the task.

A dataset that appeared stronger on one dimension could be rejected if it introduced too much downstream work, while one with known limitations could still be selected if those limitations could be repaired, worked around, disclosed, or tolerated. P_{35}, for example, settled on a freely available dataset whose *"precision was good enough"* rather than pursuing a more precise option that cost money. The free dataset was not necessarily better, but the additional precision did not justify the access cost for that task.

Participants weighed what a dataset offered against the work required to use it, including obtaining access, understanding documentation, cleaning values, reconciling definitions, combining sources, and defending its appropriateness. P_{38} compared a richer but poorly structured JSON source with an easier-to-use dataset that offered narrower coverage. As they explained, *"[I] could deliver more information based on the messed-up JSON one, but it [would] be very time consuming."* P_{38} therefore chose the narrower dataset because the additional information did not justify the effort required to make the richer source usable.

Known limitations also did not automatically rule out a dataset, particularly when no better alternative existed. P_{27}, for instance, explained: *"If the dataset has problems, but it's the only one that exists, then I would consider using it with caveats. I could write a story but warn readers: "You should also know this.""*

This suggests a different target for discovery systems: beyond ranking candidates, they should help workers estimate the effort required to make a dataset usable, compare that effort across alternatives, and determine whether its remaining limitations can be repaired or tolerated for the task.

4.2 Study 2: Dataset Discovery with an LLM-Assisted Interface

Study 2 showed that an LLM-assisted interface reorganized dataset evaluation in three ways: participants worked from system-surfaced artifacts, grounded trust in traceability to data and computation, and focused more heavily on whether the system had used the data correctly.

Surfacing Candidate Data Shifted Evaluation to System Outputs. LLM-assisted systems are often framed as automating the work of finding and preparing data and producing answers [9, 16, 39, 55]. In our study, these capabilities lowered the effort required to reach a candidate answer, but redirected evaluation toward the datasets, operations, and results returned by the system.

Participants used system outputs as starting points for targeted evaluation rather than as end-to-end substitutes for discovery. They traced intermediate steps, inspected generated code and source tables, searched for corroborating information, and recomputed results when something appeared wrong. U_{05}, for example, found that values in an intermediate table disagreed and explained: *"I saw that in [the] intermediate table the numbers were different. I then went back to step one and did the task myself."* U_{07} set the interface's answer aside and used the surfaced dataset name to retrieve figures from official reports. U_{27} manually recomputed a comparison and found that the system had matched records using `beach_name`, a non-unique join key. Together, these cases show that the interface did not complete discovery on participants' behalf, but provided concrete artifacts that focused and guided their subsequent evaluation.

In Study 1, reaching candidate data was itself costly: keyword search often fell short, and data seekers turned to colleagues, providers,

papers, and prior use to locate promising datasets. In Study 2, the interface reduced much of that sourcing cost, but it did not replace the iterative loop of search, inspection, and corroboration. Instead, it seeded that loop: the surfaced answer and artifacts gave participants concrete starting points from which to search, recompute, and corroborate. The effort saved by LLM-assisted retrieval therefore reappeared as effort spent evaluating the candidates, operations, and answers surfaced by the system.

Trust in LLM Assistance Required Data-Grounded Traceability. Prior work shows that explanations can influence trust in intelligent systems [15] and support appropriate reliance when they enable users to verify system outputs with reasonable effort [35, 53]. Our study, however, suggests that explanations alone were often insufficient for data-intensive tasks: trust depended on whether participants could trace an answer through code, intermediate values, tables, and source data.

Participants used traceability as evidence that an answer was grounded in the data. U_{02} contrasted the interface with a text-only chatbot, explaining that *"the scripts analyze the actual database, [which] gives me confidence that the AI is not faking any data. [With] just text-based GPT, I wouldn't be able to trust it that much."* When traceability broke down, trust weakened. For example, when U_{08} could not access a result table the interface had promised, the answer remained unchecked: *"I couldn't figure out how to access the result table that I was promised."*

Trust depended less on explanation alone than on whether participants could follow an answer through code, intermediate values, tables, and source data. This marks a shift from Study 1, where trust drew on data collection methods, community adoption, documentation, and provider contactability. In the LLM-assisted setting, trust hinged more narrowly on whether the answer could be traced to the data the system used.

LLM Assistance Shifted Evaluation Toward System Fidelity. Making the system's intermediate work inspectable also introduced a new object of evaluation. Beyond considering whether a dataset fit the task, participants examined whether the system had used the surfaced data and operations correctly to produce its answer.

Participants' concerns reflected this shift. Some worried that the model might fabricate information, reason incorrectly, or use the wrong data. U_{02}'s concern that a plain chatbot might be *"faking"* data made the interface's connection to the underlying database reassuring. Other participants checked generated SQL, traced intermediate values, compared outputs against source tables, or recomputed results by hand.

These checks supported evaluation of system fidelity but did not address the full question of dataset fit. Participants focused more on whether the answer followed from the data used by the system than on broader properties such as documented use, data collection methods, completeness, and other task-specific limitations. A result could be correctly computed from a table even when the underlying data was poorly documented, collected for a different purpose, largely incomplete, or otherwise ill-suited to the task.

The effect of inspectability was also mixed. Traceable artifacts such as code, intermediate tables, and source rows enabled targeted checking, but operations and outputs that appeared reasonable could also discourage further evaluation. U_{34} accepted the system's summary without checking it against the CSV, while U_{17} suggested that their general confidence in the system led them not to recompute the answer. These cases show the other side of system mediation: when surfaced artifacts or the system as a whole appeared credible, participants sometimes reduced further evaluation.

Together, these findings show that LLM-assisted discovery changed the cost and target of evaluation rather than removing it. By surfacing candidate datasets, intermediate steps, code, and tables, the interface made targeted checks cheaper and gave trust a traceable basis. At the same time, it bounded evaluation around what was visible: participants more consistently checked whether the system had used its evidence correctly than whether that evidence fit the task, leaving broader dataset-level properties less thoroughly examined.

5 DISCUSSION

The findings show that evaluation work persists across discovery settings, but its focus shifts with the checks an interface makes easiest to perform. As LLM-assisted systems lower the cost of dataset retrieval, preparation, and answer generation, the design bottleneck shifts toward evaluation. A generated answer can show how it was produced, but it may not capture a data seeker's task context. Future discovery systems should therefore be designed around this evaluation work.

Drawing on the practices observed in our studies, we derive design implications and use existing systems to illustrate concrete interface patterns.

5.1 Support Evolving Information Needs

Both studies show that information needs develop through interaction rather than being fixed in advance. Discovery systems should therefore allow users to begin with provisional information needs and refine them iteratively as candidate datasets are surfaced. LLM-assisted systems are well positioned to support this process because they can interpret natural-language descriptions and return candidates or draft answers for users to react to. The design opportunity is to make the system's evolving interpretation visible and revisable, so that subsequent search and evaluation can build on what the data seeker has already learned.

Pneuma-Seeker takes a step in this direction by reifying the system's current interpretation as an explicit relational specification and exposing intermediate artifacts [8, 9]. Our findings motivate further exploration of how such representations can help data seekers understand and revise how their information needs are represented and operationalized throughout discovery.

5.2 Design for Task-Conditioned Evaluation

Study 1 showed that data seekers weighted evaluation criteria differently depending on what the data was expected to support. A single ranking over dataset properties may not capture how data seekers evaluate candidates, because criteria are not independent or always active at the same time.

Existing systems illustrate design patterns for making dataset evaluation responsive to task context. DataScout supports attribute- and granularity-based filtering and, when users inspect a dataset,

generates task-specific relevance indicators that summarize its utilities and limitations [37]. Extending this direction, our findings suggest broadening these indicators to cover a wider range of dataset evidence and organizing them within a more explicit evaluation structure. Intended use and evaluation stage could jointly determine which criteria are foregrounded and whether they serve as early screening requirements or considerations for deeper inspection.

5.3 Surface Evidence Progressively

Although the evidence that was cheapest to inspect differed across the two studies, both showed that data seekers began with low-cost signals before investing in more demanding checks. Discovery systems should therefore align the cost of inspection with the user's current evaluation stage, while also directing attention toward evidence that may be less immediately accessible but necessary for judging a candidate's suitability. Progressive disclosure provides one interaction pattern for supporting this movement from overview to deeper inspection [50].

Existing dataset-search interfaces already stage information progressively. Google Dataset Search moves from result lists to dataset metadata [11], while Auctus moves from compact summaries to attributes and statistics [12]. Our findings suggest breaking this progression into more fine-grained steps: first presenting compact, task-critical signals for initial screening; then exposing the data and operations needed to assess a generated claim; and finally surfacing provenance, coverage, and other limitations needed for deeper evaluation and comparison.

5.4 Preserve Traceability While Supporting Broader Fit Judgments

Study 2 showed that traceability was central to trust in LLM-assisted discovery. Participants trusted system outputs when they could follow an answer back to the data and operations that produced it. Existing interfaces make data transformations easy to trace and inspect by decomposing them into intermediate steps and visualizing the effects on the data [45, 49, 52].

However, for LLM-assisted systems, traceability should not stop at answer–data alignment. Documentation frameworks provide structured representations of broader evidence about datasets [20, 24]. Systems could further indicate which of these evaluation dimensions users have examined and which remain unresolved. Lumos provides a interaction pattern by recording users' attention to data and comparing their observed behavior with a target distribution [44]. Discovery interfaces could adapt this pattern to a broader set of dataset-evaluation dimensions.

5.5 Support Situated Comparison Across Candidates

Study 1 showed that selection did not center on finding the best dataset in general. Data seekers chose among candidates whose value depended on the intended use, circumstances, and constraints. Discovery systems should therefore support comparison across candidates rather than only rank or surface a single best result. Ver takes a step in this direction by preserving multiple candidate views and eliciting users' preferences through questions about the distinctions among them [21]. Our findings suggest extending this situated comparison beyond schema and viewing preferences as a richer set of task-relevant dimensions that help users assess which tradeoffs are preferable for their intended use.

Social and contextual evidence can also support situated comparison, but it should be surfaced carefully. When search and documentation could not answer situated questions, data seekers turned to colleagues, prior users, data providers, papers, citation trails, and organizational knowledge. Early platforms such as Many Eyes illustrate how shared datasets and visualizations can accumulate community discussion and contextual knowledge [54], and more recent work demonstrates that visualizations are sometimes understood through a sociological lens, whereby audiences draw inferences about the motives and credibility behind charts [42]. Discovery systems could incorporate this pattern by attaching prior use cases, known caveats, and user notes to candidate datasets, while preserving the task context in which that evidence was produced.

5.6 Limitations

Our studies identify analytic patterns in dataset evaluation rather than estimate their prevalence. Study 1 relies on survey responses and retrospective interviews, which provide insight into participants' reasoning but may omit details from direct observation. Participants were recruited primarily through professional networks, so the samples are not intended to represent all data workers.

Study 2 is based on a separate user study of an LLM-assisted interface, with a different sample, task setting, and system design. We therefore do not treat it as a controlled comparison with Study 1. Future work should test these design implications through controlled comparisons, longitudinal deployments, and studies of LLM-assisted discovery in more open-ended search contexts.

Finally, we derive design implications from observed evaluation practices and illustrate them through existing systems rather than implementing them in a unified interface. Evaluating these directions in a unified system remains future work.

6 CONCLUSION

Dataset discovery requires data workers to decide not only which datasets are relevant, but which are good enough for an intended use. Across two qualitative studies, we found that dataset evaluation is staged, task-conditioned, and cost-sensitive. Data seekers screen candidates with accessible signals, escalate to deeper checks when uncertainty becomes consequential, and select datasets whose remaining limitations can be managed. In an LLM-assisted interface, this evaluation work did not disappear. It shifted toward the artifacts the system surfaced.

These findings suggest that discovery systems should support evaluation as much as retrieval. Human-in-the-loop and LLM-assisted tools should help data seekers refine evolving information needs, inspect evidence at the right level of detail, trace outputs to source data without narrowing fit judgments, compare candidates in context, and keep unresolved uncertainty visible.

REFERENCES

[1] [n.d.]. AWS Data Exchange. https://aws.amazon.com/data-exchange. Accessed: September 2024.
[2] [n.d.]. Data.gov. https://data.gov. Accessed: May 2025.
[3] [n.d.]. Datatrade. https://datatrade.ai. Accessed: September 2024.

[4] [n.d.]. Hugging Face. https://huggingface.co. Accessed: September 2024.
[5] [n.d.]. Kaggle. https://www.kaggle.com. Accessed: September 2024.
[6] [n.d.]. NYC Open Data. https://opendata.cityofnewyork.us. Accessed: May 2025.
[7] [n.d.]. Snowflake. https://app.snowflake.com/marketplace. Accessed: September 2024.
[8] Muhammad Imam Luthfi Balaka and Raul Castro Fernandez. 2026. Demonstration of Pneuma-Seeker: An Agentic System for Reifying and Fulfilling Information Needs on Tabular Data. In *Proceedings of the ACM Conference on AI and Agentic Systems*. 1199–1203.
[9] Muhammad Imam Luthfi Balaka and Raul Castro Fernandez. 2026. The Pneuma Project: Reifying Information Needs as Relational Schemas to Automate Discovery, Guide Preparation, and Align Data with Intent. *arXiv preprint arXiv:2601.03618* (2026).
[10] Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–424.
[11] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *The world wide web conference*. 1365–1375.
[12] Sonia Castelo, Rémi Rampin, Aécio Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. 2021. Auctus: A dataset search engine for data discovery and augmentation. *Proceedings of the VLDB Endowment* 14, 12 (2021), 2791–2794.
[13] Raul Castro Fernandez. 2025. What is the Value of Data? A Theory and Systematization. *ACM / IMS J. Data Sci.* 2, 1, Article 3 (June 2025), 25 pages. https://doi.org/10.1145/3728476
[14] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L-D. Ibáñez, E. Kacprzak, and P. Groth. 2020. Dataset search: a survey. *The VLDB Journal* 29 (2020), 251–272. https://doi.org/10.1007/s00778-019-00564-x
[15] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.
[16] Meihao Fan, Ju Fan, Nan Tang, Lei Cao, Guoliang Li, and Xiaoyong Du. 2025. AutoPrep: Natural Language Question-Aware Data Preparation with a Multi-Agent Framework. *Proc. VLDB Endow.* 18, 10 (June 2025), 3504–3517. https://doi.org/10.14778/3748191.3748211
[17] Raul Castro Fernandez. 2025. Data discovery is a socio-technical problem: the path from document identification and retrieval to data ecology. *IEEE Computer Society Data Engineering Bulletin. Preprint available at Semantic Scholar (CorpusID: 281957015). https://api. semanticscholar. org/CorpusID* 2819 (2025), 57015.
[18] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. Aurum: A data discovery system. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 1001–1012.
[19] Sainyam Galhotra, Yue Gong, and Raul Castro Fernandez. 2023. Metam: Goal-oriented data discovery. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 2780–2793.
[20] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
[21] Yue Gong, Zhiru Zhu, Sainyam Galhotra, and Raul Castro Fernandez. 2023. Ver: View discovery in the wild. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 503–516.
[22] Kathleen Gregory, Paul Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. 2019. Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. *Journal of the Association for Information Science and Technology* 70, 5 (2019), 419–432. https://doi.org/10.1002/asi.24165 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24165
[23] Thomas L Griffiths, Falk Lieder, and Noah D Goodman. 2015. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science* 7, 2 (2015), 217–229.
[24] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data protection and privacy* 12, 12 (2020), 1.
[25] Kevin Hu, Snehalkumar'Neil'S Gaikwad, Madelon Hulsebos, Michiel A Bakker, Emanuel Zgraggen, César Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayan, and Çağatay Demiralp. 2019. Viznet: Towards a large-scale visualization learning and benchmarking repository. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
[26] 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* (Santiago, AA, Chile) *(HILDA 24)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3665939.3665959
[27] Emilia Kacprzak, Laura Koesten, Jeni Tennison, and Elena Simperl. 2018. Characterising Dataset Search Queries. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1485–1488. https://doi.org/10.1145/3184558.3191597
[28] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.
[29] Javen Kennedy, Pranav Subramaniam, Sainyam Galhotra, and Raul Castro Fernandez. 2022. Revisiting online data markets in 2022: A seller and buyer perspective. *ACM SIGMOD Record* 51, 3 (2022), 30–37.
[30] Aamod Khatiwada, Grace Fan, Roee Shraga, Zixuan Chen, Wolfgang Gatterbauer, Renée J Miller, and Mirek Riedewald. 2023. Santos: Relationship-based semantic table union search. *Proceedings of the ACM on Management of Data* 1, 1 (2023), 1–25.
[31] Laura Koesten, Kathleen Gregory, Paul Groth, and Elena Simperl. 2021. Talking datasets – Understanding data sensemaking behaviours. *International Journal of Human-Computer Studies* 146 (2021), 102562. https://doi.org/10.1016/j.ijhcs.2020.102562
[32] Laura Koesten, Pavlos Vougiouklis, Elena Simperl, and Paul Groth. 2020. Dataset Reuse: Toward Translating Principles to Practice. *Patterns* 1, 8 (2020), 100136. http://dblp.uni-trier.de/db/journals/patterns/patterns1.html#KoestenVSG20
[33] Laura M. Koesten, Emilia Kacprzak, Jenifer F. A. Tennison, and Elena Simperl. 2017. The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1277–1289. https://doi.org/10.1145/3025453.3025838
[34] Eugenie Lai, Gerardo Vitagliano, Ziyu Zhang, Om Chabra, Sivaprasad Sudhir, Anna Zeng, Anton A Zabreyko, Chenning Li, Ferdi Kossmann, Jialin Ding, et al. 2025. Kramabench: A benchmark for ai systems on data-to-insight pipelines over data lakes. *arXiv preprint arXiv:2506.06541* (2025).
[35] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[36] Oliver Lehmberg and Christian Bizer. 2017. Stitching web tables for improving matching quality. *Proceedings of the VLDB Endowment* 10, 11 (2017), 1502–1513.
[37] Rachel Lin, Bhavya Chopra, Wenjing Lin, Shreya Shankar, Madelon Hulsebos, and Aditya G Parameswaran. 2025. Rethinking dataset discovery with DataScout. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
[38] Chang Liu, Arif Usta, Jian Zhao, and Semih Salihoglu. 2023. Governor: Turning open government data portals into interactive databases. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
[39] Lei Liu, So Hasegawa, Shailaja Keyur Sampat, Maria Xenochristou, Wei-Peng Chen, Takashi Kato, Taisei Kakibuchi, and Tatsuya Asai. 2024. AutoDW: Automatic Data Wrangling Leveraging Large Language Models. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* (Sacramento, CA, USA) *(ASE '24)*. Association for Computing Machinery, New York, NY, USA, 2041–2052. https://doi.org/10.1145/3691620.3695267
[40] Jock Mackinlay. 1986. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)* 5, 2 (1986), 110–141.
[41] Anthony J Million, Jeremy York, Sara Lafia, and Libby Hemphill. 2025. Data, not documents: Moving beyond theories of information-seeking behavior to advance data discovery. *Journal of the Association for Information Science and Technology* 76, 4 (2025), 649–664.
[42] Michelle Morgenstern, Amy Rae Fox, Graham M Jones, and Arvind Satyanarayan. 2025. Visualization Vibes: The Socio-Indexical Function of Visualization Design. *IEEE Transactions on Visualization and Computer Graphics* (2025).
[43] Dominik Moritz, Chenglong Wang, Greg L Nelson, Halden Lin, Adam M Smith, Bill Howe, and Jeffrey Heer. 2018. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 438–448.
[44] Arpit Narechania, Adam Coscia, Emily Wall, and Alex Endert. 2021. Lumos: Increasing awareness of analytic behavior during visual data analysis. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 1009–1018.
[45] Arpit Narechania, Adam Fourney, Bongshin Lee, and Gonzalo Ramos. 2021. DIY: Assessing the Correctness of Natural Language to SQL Systems. In *26th International Conference on Intelligent User Interfaces (IUI 2021)*. Association for Computing Machinery, College Station, TX, USA, 597–607. https://doi.org/10.1145/3397481.3450667
[46] Fatemeh Nargesian, Erkang Zhu, Ken Q Pu, and Renée J Miller. 2018. Table union search on open data. *Proceedings of the VLDB Endowment* 11, 7 (2018), 813–825.
[47] Zheng Ning, Zheng Zhang, Tianyi Sun, Yuan Tian, Tianyi Zhang, and Toby Jia-Jun Li. 2023. An empirical study of model errors and user error discovery and repair strategies in natural language database queries. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 633–649.
[48] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.
[49] Xiaoying Pu, Sean Kross, Jake M Hofman, and Daniel G Goldstein. 2021. Datamations: Animated explanations of data analysis pipelines. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[50] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.

[51] Chris Stolte, Diane Tang, and Pat Hanrahan. 2002. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (2002), 52–65.

[52] Yuan Tian, Jonathan K. Kummerfeld, Toby Jia-Jun Li, and Tianyi Zhang. 2024. SQLucid: Grounding Natural Language Database Queries with Interactive Explanations. arXiv:2409.06178 [cs.HC] https://arxiv.org/abs/2409.06178

[53] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[54] Fernanda B Viegas, Martin Wattenberg, Frank Van Ham, Jesse Kriss, and Matt McKeon. 2007. Manyeyes: a site for visualization at internet scale. *IEEE transactions on visualization and computer graphics* 13, 6 (2007), 1121–1128.

[55] Qiming Wang and Raul Castro Fernandez. 2023. Solo: Data Discovery Using Natural Language Questions Via A Self-Supervised Approach. *Proc. ACM Manag. Data* 1, 4, Article 262 (Dec. 2023), 27 pages. https://doi.org/10.1145/3626756

[56] Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024. Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=4L0xnS4GQM

[57] Liwenhan Xie, Chengbo Zheng, Haijun Xia, Huamin Qu, and Chen Zhu-Tian. 2024. Waitgpt: Monitoring and steering conversational llm agent in data analysis with on-the-fly code visualization. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[58] Ayoung Yoon. 2017. Data reusers' trust development. *Journal of the Association for Information Science and Technology* 68, 4 (2017), 946–956.

[59] Yunjia Zhang, Jordan Henkel, Avrilia Floratou, Joyce Cahoon, Shaleen Deep, and Jignesh M. Patel. 2024. ReAcTable: Enhancing ReAct for Table Question Answering. *Proc. VLDB Endow.* 17, 8 (April 2024), 1981–1994. https://doi.org/10.14778/3659437.3659452

[60] Yi Zhang and Zachary G Ives. 2020. Finding related tables in data lakes for interactive data science. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1951–1966.