DataRecycle: Trustworthy Distillation of Structured Data from Scattered Web Evidence

Can Liu, Tengteng Zhu, Xinyu Jiang, Xiaolin Wen, Yong Wang

***Abstract*—The web contains vast amounts of scattered quantitative evidence that is difficult to organize into comparable and verifiable structures. Yet its unstructured, sparse, and sometimes conflicting nature makes a systematic distillation costly and error-prone. Existing approaches force users to make a trade-off: search engines return document lists that require labor-intensive extraction and reconciliation, while generative AI systems produce fluent summaries that obscure data provenance, introduce reporting bias, and hide inconsistencies. We present *DataRecycle*, an interactive system that helps transform scattered web evidence into structured tables of comparable numerical facts. Given a user's query or statement, *DataRecycle* decomposes it into entities and attributes, retrieves candidate evidence, and organizes scattered, potentially conflicting records in a table-based evidence view. To address key challenges in automated data distillation (i.e., incomplete retrieval, untraceable generation, sparse data distribution, and multi-source inconsistency), *DataRecycle* adopts an ontology-guided, top-down data acquisition strategy, visualizes the recorded lineage from retrieved snippets to extracted values, and aligns heterogeneous data across dimensions such as time, definition, and scale. We evaluate *DataRecycle* through module-level experiments, a usage scenario, and a user study, showing that the system helps users organize numerical web evidence into provenance-linked tables and inspect uncertainty and conflicting sources.**

***Index Terms*—Data aggregation, uncertainty visualization, fact verification**

I. Introduction

While existing curated datasets provide a solid foundation for systematic analysis, they often cover only a limited portion of the knowledge available in the world [1]. Much of the most timely and fine-grained data, such as news, corporate disclosures, and market reports, first appears as unstructured text on web pages [2] before being incorporated into official datasets. Converting such evidence into official structured datasets requires substantial time and effort, leading to an inevitable lag between real-world events and dataset updates. Consequently, to capture emerging trends, underrepresented sectors, or the long tail of entities and metrics outside the scope of major data providers, researchers often need to turn to the web as a data source.

C. Liu is with the School of Information and the Institute for the Community for the Chinese Nation at the Central University of Finance and Economics, and with Nanyang Technological University. This work was conducted primarily while he was a research fellow supervised by Yong Wang at Nanyang Technological University. Email: can.liu@cufe.edu.cn.

X. Wen and Y. Wang are with Nanyang Technological University. Email: xiaolin004@e.ntu.edu.sg, yong-wang@ntu.edu.sg.

Y. Wang is the corresponding author.

T. Zhu is with Beijing Institute of Technology and Nanyang Technological University. Email: bit_zhutengteng@163.com.

X. Jiang is with the University of Sydney. Email: xjia0751@uni.sydney.edu.au

Despite this potential, much of the available web data remains underutilized because it is largely unstructured. Unlike structured datasets, in which entities, attributes, and relationships are explicitly defined, valuable quantitative facts on the web are scattered across different websites. These facts are typically buried within unstructured text, which makes systematic analysis difficult. For instance, assessing the global green energy landscape requires aggregating data points from diverse sector-specific reports. Many such data points are absent from global statistical datasets; even when included, they are often poorly aligned, reported in inconsistent units, or associated with conflicting values because of differences in methodology or reporting cycles. Therefore, distilling comparable structured facts directly from the web is essential for bridging the gap between curated datasets and evidence reported in the wild.

Current tools largely follow two document-centered workflows: search engines return ranked pages, while generative engines [3] summarize retrieved pages into prose. Traditional search engines return ranked lists of web pages, leaving users to navigate multiple pages, locate relevant facts, align them across entities, and manually resolve contradictions. A recent system, Compendia [4], partially reduces this burden by generating visual narratives from retrieved documents, but it remains document-centric. In particular, it does not construct a table of the underlying facts, making fact-level coverage difficult to inspect. Generative engines such as Perplexity[1] and ChatGPT[2] improve usability by synthesizing direct answers, but their answers are still shaped by both retrieval coverage and generation-time selection. Even when relevant pages are retrieved, numerical facts may be selectively summarized, merged into prose, or omitted, making it difficult for users to inspect which entity–attribute cells are supported, missing, or conflicting across sources. Beyond verifying individual data points, users often need context to interpret what a numerical value actually means. This requires the same attributes to be aligned across entities, yet document retrieval often makes such alignment inconsistent because different retrieved pages emphasize different aspects of each entity. For example, if the top retrieved documents for Country A emphasize installed capacity whereas those for Country B focus on subsidies, a generative engine may produce a narrative that fails to compare the same metric across countries. Moreover, generative engines often smooth over such inconsistencies and obscure data sparsity. By collapsing conflicting evidence into a single narrative, they hide the provenance and distribution

[1]https://www.perplexity.ai/

[2]https://chatgpt.com/

of the underlying data, making it difficult for users to assess uncertainty, identify contradictions, or determine where data is fundamentally missing.

To address this problem, we first identify the fundamental challenges that hinder the transformation of scattered web fragments into verifiable and structured data (C1-C4).

- **C1. Incompleteness of retrieval results.** Bottom-up retrieval and summarization methods based on relevance often fail to cover the full scope of a target schema. Search engines and generative engines cannot guarantee retrieval of all data facts required to construct a complete comparison table.
- **C2. Untraceability of generation process.** The data generation process in existing approaches often lacks traceable lineage. Users cannot verify how specific values are extracted or computed from source documents, making it difficult to distinguish grounded facts from model hallucinations.
- **C3. Sparsity of data distribution.** Quantitative data is often sparsely distributed and isolated across unrelated documents. A critical data point may appear only once in a lengthy report, which makes it difficult to aggregate dispersed fragments into a coherent structure. For example, when users compare attributes across multiple countries, relevant evidence may not appear in the same document unless prior analyses have already consolidated it.
- **C4. Inconsistency in multi-source evidence.** Data retrieved from heterogeneous sources often contains conflicts. These discrepancies arise from both factual errors and mismatches in dimensions, such as differences in time frames, scopes, or attribute definitions, which complicate validation.

To address these challenges, we propose *DataRecycle*, an interactive system for trustworthiness-oriented distillation of structured tables of comparable numerical facts from scattered web evidence. Given a user query or statement, *DataRecycle* decomposes it into constituent entities, attributes, and scope, retrieves relevant web sources, and integrates fragmented evidence into a unified view. To guide retrieval toward schema-level coverage (C1), *DataRecycle* adopts an ontology-guided top-down approach. Instead of relying only on relevance-ranked retrieval, which can leave some schema cells under-covered, the system first specifies a target schema and then retrieves and extracts evidence for each entity-attribute cell. To support traceability (C2), *DataRecycle* records and visualizes the system-observed lineage from retrieved snippets to final data values, allowing users to inspect the extraction and computation steps represented in the pipeline. To address data sparsity (C3) and potential conflicts (C4), we introduce a context-aware alignment mechanism that normalizes heterogeneous data points into a unified coordinate system, for example, by standardizing temporal, unit, and attribute dimensions, together with a multidimensional source-quality signal that helps prioritize potentially conflicting evidence for further inspection. To further support the inspection of conflicting multi-source evidence (C4), we design a compact uncertainty encoding based on the complementary cumulative distribution function (CCDF) for dense tabular comparisons, summarizing source-supported value distributions while preserving comparability across entities. With *DataRecycle*, users can transform natural-language queries into structured comparison tables, trace evidence for each cell, and inspect conflicting perspectives interactively.

Our main contributions are as follows:

- We present a trustworthiness-oriented data engine for distilling comparable structured facts from scattered web sources, integrating retrieval, evidence extraction, context-aware alignment, and source-aware aggregation.
- We develop *DataRecycle*, an interactive system that supports provenance-aware table construction and uncertainty inspection, including a compact CCDF-based encoding for dense comparison views.
- We extensively evaluated *DataRecycle*: a usage scenario to showcase its workflow, quantitative benchmarking of the evidence extraction and record-level auditing modules, a controlled comparison of CCDF bars and density plots, and a user study on the provenance interface, which collectively demonstrated its effectiveness and usability.

II. Related Work

This work is related to prior research on web data extraction, data trustworthiness, and uncertainty visualization.

A. Web Data Extraction

Prior work on web data extraction mainly follows two paradigms: automated extraction and structure discovery with constraint modeling.

Automated Extraction. Early work focused on scalable information extraction using predefined patterns or relation-learning models, such as Snowball [5] and OpenIE [6]. Subsequent systems shifted toward probabilistic and declarative frameworks, including DeepDive [7] and Knowledge Vault [8], which integrate heterogeneous signals for large-scale knowledge base construction. These approaches support high-throughput extraction, but they typically assume relatively stable schemas or rely on post hoc normalization. ChartifyText [9] uses an LLM to infer tables from text containing numerical data and generate charts from those tables.

Structure Discovery and Constraint Modeling. Beyond direct extraction, recent studies have investigated how to uncover latent structures from noisy, heterogeneous text. Structure profiling techniques use clustering and representation learning, for example, StructVizor [10] and Tsai et al. [11], or graph-based formulations such as clique-expansion hypergraphs in HINTs [12], to identify implicit schemas. In parallel, constraint-guided methods improve consistency by incorporating domain knowledge, including state-based decoding [13], in-context log encoding [14], purpose-driven inference [15], and hierarchical blueprint representations [16], [17]. These methods improve structural coherence, but they remain largely automated and opaque to end users. SciDaSynth [18] lets users query a screened set of scientific papers, check extracted values against the papers, and refine the table.

Existing systems such as ChartifyText and SciDaSynth structure information from supplied texts or screened collections of papers. *DataRecycle* starts from a comparison query and retrieves numerical evidence from the open web. It uses a

shared ontology to align records by entity, attribute, time, and scope. When reports still disagree, it shows the distribution of reported values and links them to their sources.

B. Data Trustworthiness

Assessing conflicting web sources requires source reliability estimation and consistency validation [19].

Source Reliability Estimation. This line of work assesses the credibility of information providers through both algorithmic modeling and human-centered analysis. Automated approaches infer source reliability from signals, such as interaction patterns and propagation behaviors [20], and structural anomalies using graph-based techniques [21], through fine-grained [22] and bias-aware [23] probabilistic inference. Complementarily, visual analytics research has framed trust as an interpretable and human-in-the-loop process, emphasizing data provenance [24], transparency [25], and interactive sensemaking [26]. Building on these principles, several representative systems include those for the detection of anomalous user-generated content [27], [28], and the analysis of anomalous information spreading [29].

Consistency Validation. The second line of work establishes truth by validating extracted facts against logical constraints or prior knowledge. Existing methods enforce logical rules [30], [31] or align extracted facts with external knowledge bases [8] to detect inconsistencies. In parallel, visual analytics systems, such as QualityTrails [32], enable users to trace data transformations and assess consistency through transparent and interactive workflows. FOX [33] groups facts extracted from a dataset by analytical scope and helps users explore outliers within each group.

Existing approaches often provide limited support for verifying open-domain web content. Our framework addresses this by combining source reliability estimation and cross-source consistency inspection in a transparent process, helping users examine trust-related evidence more explicitly.

C. Uncertainty Visualization

Communicating uncertainty is essential for reliable analytical decision-making, especially in settings where data are incomplete, noisy, or potentially conflicting [34], [35]. Existing visualization approaches for uncertainty communication generally follow two paradigms: extrinsic auxiliary elements and intrinsic channel modifications [36].

Extrinsic Techniques. Extrinsic methods add auxiliary graphical elements to depict distributions or error ranges without changing the original marks. Common examples include error bars [37] and uncertainty glyphs [38]–[40]. For instance, Winglets [38] attach dual-sided strokes to scatterplot points, using their length and orientation to encode cluster association and uncertainty. Such designs make uncertainty explicit and can improve perceptual grouping beyond color alone.

Intrinsic Techniques and Integrations. Intrinsic methods instead modify existing visual channels, such as opacity and sketchiness [35], [41], to communicate confidence [42], [43]. Other encodings, including value-suppressing palettes [44], Bayesian surprise maps [45], and quantile dotplots [46], [47], have been shown to reduce inference errors. Recent systems further integrate uncertainty into analytical workflows, including feature attribution systems such as Explainer [48] and VBridge [49], natural-language interfaces such as DataTone [50] and Eviza [51], temporal ambiguity analysis [52], and storytelling [53].

Existing techniques provide limited support for representing conflicts across multiple sources. To address this limitation, we propose a novel encoding that combines bars with cumulative distribution function curves to visualize uncertainty and data distribution.

III. Overview

As illustrated in Figure 1, *DataRecycle* consists of two major components: (1) a **trustworthiness-oriented data engine** (section IV) for evidence acquisition, organization, and reliability estimation, and (2) an **interactive interface** (section V) for user-facing exploration and analysis. Our system pipeline is designed to address the aforementioned four challenges. To address incomplete retrieval results (**C1**), *DataRecycle* first transforms the users' natural-language input into a structured schema through *ontology-based parsing*. This ontology-derived schema defines the initial layout of the output table and provides top-down guidance for subsequent retrieval. By explicitly specifying target entities and attributes, this schema-guided process directs retrieval toward individual table cells and makes under-supported cells explicit. To address sparse data availability across web sources (**C3**), *DataRecycle* then performs targeted *web crawling* over diverse sources, such as government reports and news articles. The *evidence extraction* module identifies relevant data points from these lengthy and disconnected documents, producing candidate evidence for subsequent integration. To address evidence conflicts from heterogeneous sources (**C4**), *DataRecycle* employs a *data consolidation and conflict resolution* module. This stage aligns values under a unified schema, consolidates comparable evidence, and analyzes discrepancies caused by dimensional mismatches or factual conflicts. To support traceability throughout the process (**C2**), *DataRecycle* quantifies source reliability and presents the consolidated results in an interactive table view augmented with trust scores and provenance evidence. This design allows users to inspect underlying distributions, conflicts, and uncertainties, and to trace specific values back to their supporting documents.

IV. Trustworthiness-Oriented Data Engine

The pipeline begins with **ontology-based parsing** (subsection IV-A), which extracts multidimensional coordinates, such as entity, attribute, time, and unit, from user input to define the fundamental structure of the analytical space. Guided by these coordinates, the system performs **web crawling** and **evidence extraction** (subsection IV-B) over diverse web sources to collect relevant raw facts. To improve data robustness, the retrieved facts then undergo a dual-track refinement process. The **data consolidation and conflict resolution** module (subsection IV-C) mitigates inconsistencies by aligning scales

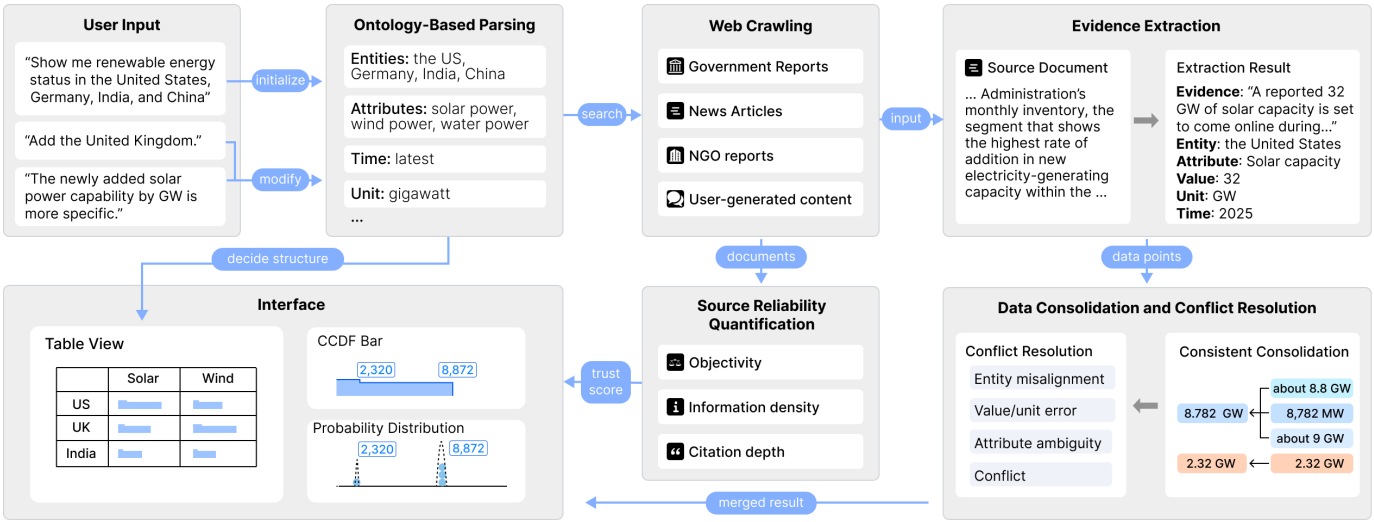


Fig. 1. Overview of the *DataRecycle* workflow. A natural language input is parsed into a schema consisting of entities, attributes, time, and scope, which determines the table structure in the interface. The schema also guides web crawling and evidence extraction from diverse sources. Extracted evidence is then consolidated, assessed for source trustworthiness, and integrated into the final table and visualizations shown to the user.

and resolving contextual mismatches, while the **source reliability quantification** module (subsection IV-D) assesses the credibility of sources.

TABLE I
THE 6-TUPLE ONTOLOGY FOR DATA FACTS.

Component	Description	Example
Entity (**E**)	Subject serving as the comparison anchor.	"Germany"
Attribute (**A**)	Specific dimension or property being measured.	"HSR Length"
Time (**T**)	Temporal timestamp for trend analysis.	"2020", "2023"
Scope (**S**)	Shared contextual constraints that condition the fact.	"Singapore market"
Value (**V**)	Numerical value under comparison.	"1500"
Unit (**U**)	Measurement unit for numerical alignment.	"km", "USD"

A. Ontology-Based Parsing

To address contextual misalignment in unstructured web text, we define an ontology-inspired six-tuple schema that serves as a coordinate system for extracted facts. Each extracted fact is represented in six dimensions: Entity (E), Attribute (A), Time (T), Scope (S), Value (V), and Unit (U), as summarized in Table I. For example, from the sentence *"In 2022, Toyota sold 2 million cars and BMW sold 1.5 million cars in Europe,"* we extract two distinct facts. Both share the same attribute (car sales), time (2022), and geographic scope (Europe), but they capture different entities and values: Toyota at 2 million units, and BMW at 1.5 million units. Crucially, the assignment of these dimensions depends on context. While "Europe" functions as a shared scope constraint in the previous example, in a query such as *"Tesla sales in Germany and India,"* the countries instead become entities because they serve as the primary subjects of comparison. This ontology directly addresses a root cause of data inconsistency. Consider a query for the *"annual car sales of Tesla in Germany,"* which might return two seemingly contradictory values: 200,000 and 250,000. Without contextual awareness, a system might incorrectly identify them as a factual conflict. However, our ontology-based coordinate space distinguishes these values by assigning them to different temporal contexts, for example, one to 2022 and the other to 2023. Formally, because they differ along the time dimension, no true conflict exists. By attaching contextual metadata to every data point, the schema transforms isolated values into context-bearing, inspectable records and enables users to align facts along shared dimensions, such as country, year, and metric, for presentation in a comparable tabular form.



Fig. 2. From natural-language query to table schema. *DataRecycle* parses a comparative query into entities, time, scope, and attribute intent, then expands the attribute into measurable dimensions to generate the target table schema.

Building upon this schema, our operational pipeline defines four transformation types that expand, align, and refine the analytical space. Figure 2 illustrates how ontology-guided parsing converts a comparative natural-language query into an explicit table schema. In this example, the system identifies the target entities and year, interprets the coarse attribute intent ("green energy installed capacities"), and refines it into measurable columns for subsequent retrieval. We operationalize this schema through four transformation types:

- **Entity handling:** Expands the comparison set with related entities to reduce sparsity and support cross-entity analysis.
- **Attribute handling:** Maps underspecified analytical intents to measurable attributes and adjusts attribute granularity when needed.

- **Time handling:** Extracts timestamps to align facts on a shared temporal axis and avoid conflating values from different years.
- **Scope handling:** Enforces scope matching over contextual constraints in S so that only contextually aligned facts are compared.

These operations provide a common vocabulary for the transformations supported by the interface; user interactions are implemented through one or more of these operations.

B. Web Crawling and Evidence Extraction

Based on the proposed ontology, we retrieve and extract evidence for each table cell defined by an entity, attribute, time, and scope.

Web crawling and text collection. For each cell, we construct a query by combining the entity name, attribute keyword, temporal constraint, and, when applicable, the scope, such as a country or region. We send each query to Google via the Serper API[3] and retrieve up to 20 ranked results by default. The limit is configurable. We download the returned pages and clean them by removing HTML tags, navigation menus, and advertisements. We then extract the cleaned text body and remove duplicate paragraphs to avoid repeated content [8]. The resulting corpus spans diverse sources, including government reports, news articles, NGO documents, independent media reports, and personal blogs.

LLM-Assisted Evidence Extraction. We use an LLM-based pipeline to transform unstructured text into ontology-consistent structured evidence. Specifically, we prompt the LLM to extract relevant evidence for each target cell and return the result in a structured format. The prompt includes the field definitions derived from the ontology, together with an illustrative example, to guide the model toward consistent outputs. For each text segment and target cell, defined by scope, entity, attribute, and time, the LLM returns one or more records containing the entity, attribute, time, scope, value, unit, a relevance score, and the supporting sentence(s). For example, suppose the target cell corresponds to *entity* = *Germany*, *attribute* = *solar power total installed capacity*, and *time* = *2025*. If a retrieved web page contains a sentence such as "At the end of 2025, the total installed solar capacity in Germany reached 117 GW," the LLM extracts a record with `value = 117`, `unit = GW`, and the quoted sentence as supporting evidence. These supporting sentences serve as human-readable justifications, while the structured fields are used for downstream aggregation and visualization. Multiple records may correspond to the same cell. We retain all such records for subsequent deduplication, conflict resolution, and confidence estimation.

C. Data Consolidation and Conflict Resolution

The data merging process is conducted in two sequential steps: interval-based data consolidation and LLM-assisted conflict resolution.

[3]https://serper.dev/

Interval-Based Data Consolidation. The first step aims to merge non-conflicting values extracted from different sources whenever possible. Such cases often arise when sources report values at different granularities or in different interval forms. For example, Source A may report a value larger than three, Source B may report 3.5, and Source C may report 3.54. The intervals implied by the three reports overlap, so we treat the reports as compatible. To assess this consistency, we convert the reported values into intervals and examine their overlaps. Numerical precision can implicitly convey uncertainty. When uncertainty is not explicitly reported, the last significant digit implies an interval of approximately ± 0.5 at that level of precision. For example, 3.78 implies an interval of $[3.775, 3.785]$ [54]. By representing values as intervals, we can identify compatible reports and consolidate them accordingly. For a group of mutually consistent values, we use the highest-precision value as a display representative while retaining all supporting records for inspection. This choice preserves the finest reported granularity but does not treat numerical precision itself as evidence of factual accuracy. After this step, multiple distinct results may still remain, and we therefore further resolve these apparent discrepancies through conflict resolution, as their underlying causes may differ.

LLM-Assisted Conflict Resolution. After the consolidation phase, unresolved discrepancies may still exist. To address them, we introduce an LLM-assisted approach to conflict resolution. This module examines the original textual evidence, the extracted values, and the conflicting results to diagnose the underlying cause of each discrepancy. We categorize the causes of conflicts into the following four types and apply tailored resolution strategies:

- **Entity misalignment:** The extraction incorrectly targets the wrong entity or scope (e.g., extracting global values when U.S. values are required). *Resolution:* Data points identified with entity misalignment are filtered out.
- **Value or unit extraction errors:** These are system-level failures during the NLP extraction process, including incorrect numeric parsing or unit conversion errors. *Resolution:* Upon detecting such errors, the system automatically triggers a re-extraction process for the affected data points.
- **Attribute semantic ambiguity:** The extraction is technically correct, but the queried attribute is inherently ambiguous and requires further specification (e.g., "solar capacity" could mean newly installed capacity" or "cumulative capacity"). *Resolution:* The system defers to the user by prompting them in the chat view (subsection V-A). The user can either specify a more precise definition or choose to retain both interpretations as newly defined attributes.
- **Genuine Factual Contradiction:** The extracted records appear consistent with their source texts, but the underlying sources report incompatible values. Because such conflicts cannot be resolved automatically without additional domain evidence, the system retains the competing records and summarizes their relative source-supported weights using the source-quality scores (detailed in subsection IV-D). The distribution and the conflicting sources are then exposed in the user interface, empowering users to make the final

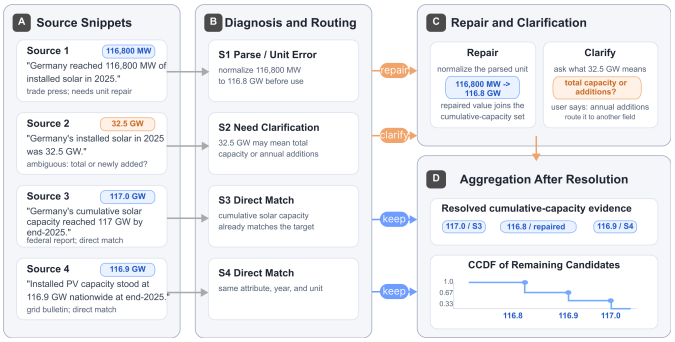


Fig. 3. Conflict resolution via source cases.

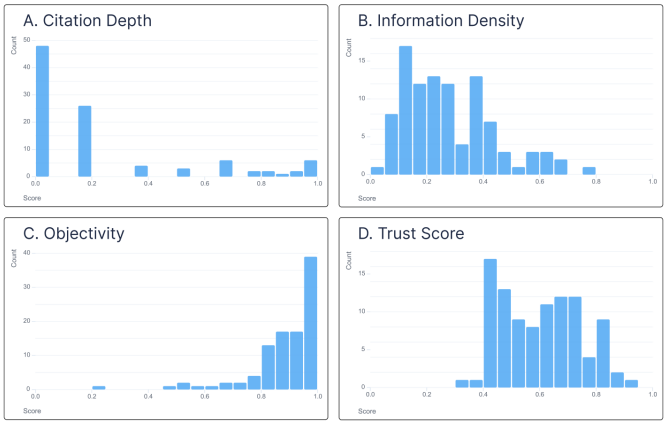


Fig. 4. The distributions of the three components of our source reliability model and the final trust score, computed on 100 web documents.

judgment based on their domain knowledge.

Figure 3 illustrates the conflict-resolution workflow with four source cases. The reports that already match the target attribute proceed directly to aggregation, whereas problematic cases are routed to either repair or clarification. In the example, a value reported as 116,800 MW is normalized to 116.8 GW and then re-enters the cumulative-capacity evidence pool, while an ambiguous 32.5 GW report is clarified as annual additions and therefore rerouted to a different attribute. Only resolved cumulative-capacity candidates are passed to the final CCDF encoding.

D. Source Reliability Quantification

We define the trust score $T(s)$ as a normalized, heuristic source-quality score for a document s. It combines citation evidence, information density, and stylistic neutrality to prioritize sources within the retrieved evidence set; it is not intended as a calibrated probability that an individual claim is correct.

- **Citation Depth (v_{cite}).** Motivated by prior work on knowledge-based trust [55], we use the presence of traceable citations as one inspectable proxy for source quality. We quantify citation depth using a hyperbolic tangent function to normalize the count of references as

$$v_{cite} = \tanh((N_{ext} + \kappa \cdot N_{acad})/\tau),$$

where N_{ext} represents the count of valid external hyperlinks, and N_{acad} denotes the count of references adhering to academic standards. The parameter $\kappa > 1$ weights academic citations higher than general links, and τ is a saturation constant controlling the marginal utility of additional citations. Empirically, we set $\kappa = 2$ to emphasize scholarly rigor and $\tau = 5$ to model the principle of diminishing returns (e.g., 5 weighted citations yield $\tanh(1) \approx 0.76$).
- **Information Density (v_{dens}).** To measure the substantive content ratio (i.e., signal-to-noise ratio) of the text, we calculate the density of named entities and numerical data:

$$v_{dens} = \sum_{t \in T} (\mathbb{I}(t \in \mathcal{E}_{NER}) + \mathbb{I}(t \in \mathcal{N}_{num}))/|T|,$$

where T is the set of valid tokens (excluding stop words) in the text, and $\mathbb{I}(\cdot)$ is the indicator function. $\mathcal{E}_{NER}$ and $\mathcal{N}_{num}$ represent the sets of named entities and numerical tokens.
- **Objectivity (v_{obj}).** Following the definition of data objectivity [56], we penalize sources exhibiting strong emotional polarity, which often correlates with subjective bias. The objectivity score is defined as $v_{obj} = 1 - |\text{Polarity}(T)|^{\gamma}$, where $\text{Polarity}(T) \in [-1, 1]$ is the output of a pre-trained sentiment analysis model (i.e., BERT-Sentiment). We assign $\gamma = 1.5$ to exponentially penalize extreme sentiment polarity while preserving subtle emotional expressions.

The final trust score is computed using a logistic function $T(s) = \sigma(\mathbf{w}^T \mathbf{v}_s + b)$, where $\sigma(\cdot)$ is the sigmoid function mapping the result to $(0, 1)$. We use heuristic defaults of $\mathbf{w} = [1.2, 3.0, 0.6]^T$ and $b = -1.2$, which users can adjust in the configuration panel. When we varied one weight at a time by $\pm 30\%$, all Spearman rank correlations with the default ranking exceeded 0.98.

Distribution of the Trust Scores. Figure 4 shows the distributions of the three components and the final trust scores for 100 web documents sampled from our corpus. Citation depth is highly skewed, information density varies more broadly, and the final scores span a broad range.

V. Interface of *DataRecycle*

The interface of *DataRecycle* is designed to support seamless transitions between high-level exploratory analysis and low-level data verification. It is organized into four coordinated components: the traceability inspector (Figure 6 A) for detailed provenance inspection, the metadata panel (Figure 6 B) for schema management, the main table view (Figure 6 C) for result presentation, and the chat view (Figure 6 D) for intent-driven interaction.

A. Chat View and Metadata Panel

The chat view (Figure 6 D) serves as the natural-language entry point to *DataRecycle*. It supports several key interaction capabilities:

- **Initializing analysis intent.** Users can issue high-level, fuzzy queries (e.g., "Compare green energy capacities"), which the system automatically parses to construct an initial table schema with appropriate entities and attributes.

- **Iterative schema refinement.** Users can refine the table by adding or modifying entities, attributes, or scopes through natural-language commands.
- **Human-in-the-loop semantic clarification.** When attributes are ambiguous or overly coarse, users can negotiate more precise definitions with the system. In response, the system suggests refined schema variants and updates the table accordingly.
- **Proactive ambiguity and conflict surfacing.** When the background data engine detects semantic ambiguities or conflicting definitions (e.g., inconsistent interpretations of a single attribute), the assistant proactively raises an "ambiguity detected" alert and suggests schema refinements, such as splitting or redefining attributes.

The parsed results of user queries are explicitly displayed in the metadata panel (Figure 6 B), which is located above the table. This panel organizes the analysis into three structured facets: *Scope and Time*, *Entities* (e.g., countries), and *Attributes* (e.g., solar power capacity). Users can directly intervene in this structured representation, for example, by clicking the "+ ADD" button to introduce new entities or refine contextual constraints, ensuring that the system's retrieval boundaries remain aligned with their analytical intent.

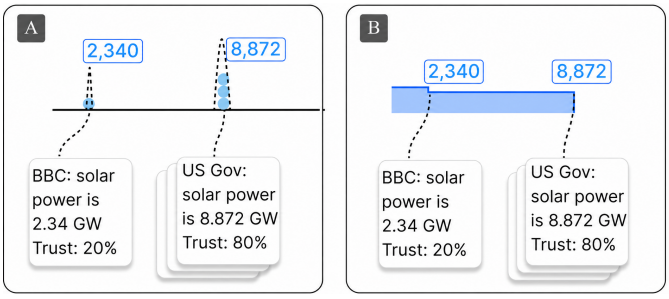


Fig. 5. Two visual encodings for visualizing contradictions in source data. Plot labels use MW, while source cards report GW (1 GW = 1,000 MW). The curves are schematic; mixtures of uniform interval densities yield piecewise-constant densities and piecewise-linear CCDFs.

B. Table View

The table view (Figure 6 C) serves as the primary workspace for data presentation. Each row represents an entity, and each column represents an attribute. To handle the inherent uncertainty and multiplicity of web-extracted data, the table provides two specialized visual encodings, as shown in Figure 5.

- **Probability distribution encoding.** For each cell, we visualize multiple extracted values collected from different sources. We build the density from the original records with finite intervals and use one-sided bounds only to check compatibility. Each eligible extracted value x_i is modeled as a uniform local density $p_i(x)$ over an interval $I_i = [x_i - \delta_i, x_i + \delta_i]$, where the half-width $\delta_i > 0$ is determined by the reported precision of x_i. The density satisfies $p_i(x) \geq 0$ and $\int p_i(x)\,dx = 1$. Specifically, $p_i(x) = 1/(2\delta_i)$ for $x \in I_i$ and $p_i(x) = 0$ otherwise. Given trust scores $\{t_i\}_{i=1}^n$, we aggregate these local densities into a normalized, source-weighted evidence-support distribution, $P(x) = \sum_{i=1}^n \alpha_i p_i(x)$, where $\alpha_i = t_i / \sum_{j=1}^n t_j$ and $\sum_{i=1}^n \alpha_i = 1$. Thus, sources with higher trust scores contribute more strongly to the final distribution. The evidence-support model uses this mixture directly, without kernel smoothing. Its density is piecewise constant, and its CCDF is continuous and piecewise linear. In Figure 5 A, each source is represented by a small dot positioned according to its extracted value along a horizontal value axis shared by all cells in the same attribute column. To reduce overplotting, nearby dots are vertically jittered or stacked. The aggregated distribution $P(x)$ is shown as an overlaid density curve. This encoding reveals variance, consensus peaks, and outliers, allowing users to assess how concentrated or dispersed the evidence is for a given metric. As illustrated in Figure 5, this example contains four sources. Three sources report the value 8.872 GW, each with a trust score of 80% (computed as described in subsection IV-D), whereas one source reports 2.34 GW with a trust score of 20%. Accordingly, the density curve places greater emphasis on the cluster around 8.872 GW while still preserving the lower-confidence evidence at 2.34 GW.
- **CCDF bar encoding.** Although the probability-distribution encoding is intuitive for a single cell, it leaves much of the cell area unused when applied to large tables. To address this limitation, we introduce a more space-efficient visualization based on a CCDF bar (Figure 5 B) to represent the same uncertainty information. When users need to compare magnitudes across entities while retaining uncertainty context, they can switch to the CCDF bar encoding (Figure 6 C2). We introduce a bar-like visual encoding that represents the remaining evidence-support mass along the value axis. Let X follow the aggregated evidence-support distribution, and let $F(x)$ denote its cumulative distribution function (CDF). We define the CCDF as: $R(x) = P(X > x) = 1 - F(x)$. The CCDF bar visualizes $R(x)$ as the profile of a horizontal bar over the value axis. The bar begins at full height (probability 1) and progressively decreases as probability mass accumulates along the axis. For discrete data points clustered from web sources with value probability pairs $D = \{(v_1, p_1), \ldots, (v_n, p_n)\}$, this yields a descending staircase structure: $R(x) = 1 - \sum_{v_i \leq x} p_i$. For continuous distributions with density $f(x)$, it forms a continuous, non-increasing profile: $R(x) = \int_x^\infty f(t)\,dt$. Bar extent marks the upper end of the support. To compare entities, read $R(x)$ at the same threshold x; extent alone gives no general probability ordering. $R(x)$ describes the evidence-support model, not a calibrated probability of truth.

As shown in Figure 7, A presents the probability distribution encoding, and B shows the CCDF encoding. Both encodings compare mobile data prices and average monthly income across several countries. The CCDF encoding can be interpreted as an enhanced bar chart: when the values reported by different sources are highly consistent, it nearly collapses to a conventional bar chart. For the price per GB in Laos, the profile falls over different price ranges. The horizontal gaps between these declines show how far apart the reported

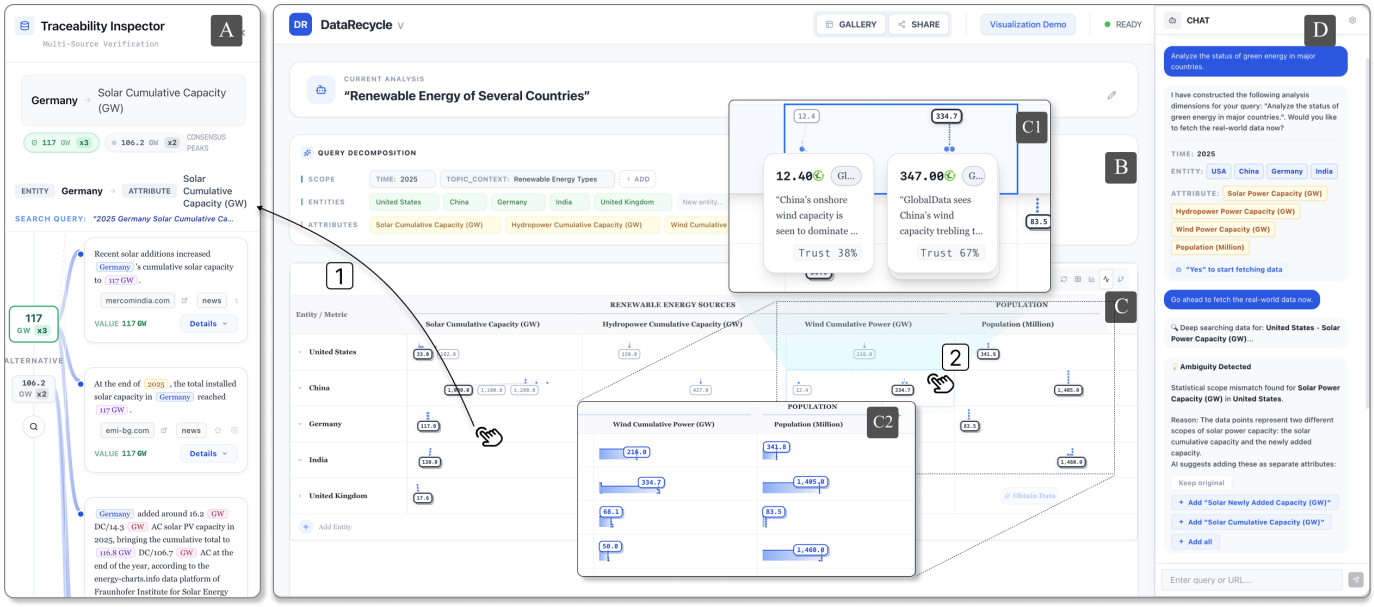


Fig. 6. The user interface of *DataRecycle*. (A) A provenance inspector that reveals the specific data source of a cell's value upon clicking. (B) A metadata panel displaying the entities, attributes, and scope of the current table. (C) The main table view presenting the data results, equipped with a lens tool (C1) that allows users to hover and examine detailed information. (D) A chat interface for users to input natural language queries and interact with the system.



Fig. 7. Cross-entity comparison with (A) probability distributions and (B) CCDF bars. Each CCDF bar ends at the distribution's upper bound. Its height at each threshold shows the share of evidence weight above it.

prices are. Compared with the probability-distribution view, the CCDF view is more compact and supports rapid scanning across rows and columns. Beyond visual encodings, the table view also supports interactive operations that enable users to inspect uncertainty and derive new insights from the data.

- **Semantic Lens.** Due to limited cell space, displaying the full context of conflicting values simultaneously is impractical. To address this issue, we designed a semantic lens tool. When a user hovers over a cell, the lens expands to juxtapose the distinct values. As shown in Figure 6 C1, the lens provides a set of summary cards for each extracted value. Each card displays the numerical value, the supporting extraction context, and a calculated trust score for the source. This design allows users to inspect conflicting data in situ without losing the surrounding table context.
- **Attribute Derivation.** Users can click the attribute derivation button to create a new attribute based on expressions over existing attributes. For example, in Figure 7, after obtaining the data price per GB and the average monthly income, we can derive a new attribute, "affordable data volume", defined as the ratio of income to data price. This derived attribute enables more meaningful comparisons and provides an intuitive understanding of data affordability across countries. This calculation divides representative income by representative data price. The result is a point estimate without joint uncertainty propagation.

C. Traceability Inspector Panel

The Traceability Inspector (Figure 6 A) is a collapsible drawer that exposes provenance details for the selected cell. When a user selects a specific data cell in the table, the corresponding traceability panel is automatically revealed. At the top of the inspector, we present an overview of the consensus peaks, summarizing the aggregated evidence for the selected cell. Beneath this overview, the panel lists a set of evidence cards, each corresponding to an individual source that contributes to the selected cell. Each evidence card exposes rich data lineage, including the source URL and its domain type (e.g., news outlet, government report), the extracted semantic tags (entity, attribute, time, and unit), and the exact text span from the original webpage that serves as evidence. Users can interact with these cards to open the original source in an external browser, assess the quality of the automatic extraction, or explicitly exclude erroneous or low-quality sources from the final aggregation.

VI. Evaluation

Our evaluation comprises four parts: a usage scenario illustrating the workflow, module benchmarks of evidence

extraction and record-level auditing, a controlled perceptual comparison of the two encodings, and a system user study collecting usability and perceived-usefulness feedback.

A. Usage Scenario

Using the green energy sector as an example, relevant data are typically distributed across websites with inconsistent statistical criteria. Traditional search engines often return noisy or localized results.

From Fuzzy Query to Initial Table and Metadata. Our approach allows users to begin exploration with a fuzzy natural language query. For example, a user may ask in the chat view (Figure 6 D), "How do the green energy installed capacities of the United States, China, Germany, and India compare for 2025?" The system semantically parses the query and extracts the corresponding metadata, including Scope (Time), Entities, and Attributes (e.g., Solar, Hydropower, and Wind). The extracted metadata are structured and displayed in the metadata panel (Figure 6 B). Based on these dimensions, the system automatically generates an initial two-dimensional table view (Figure 6 C). At this stage, the user can further refine the table, for example, by adding the UK as a new entity through either the table view or the chat view.

Human-in-the-loop Conflict Resolution. After confirming the table structure, the user prompts the system to fetch data. The backend crawler then performs cell-level retrieval. However, open-web data are often contradictory because of mismatched entity granularity, differences in temporal scope, or attribute ambiguity. To address these issues, the Audit Agent performs preliminary cleaning and supports a human-in-the-loop workflow. As shown in the chat view (Figure 6 D), when the backend detects semantic ambiguity in an attribute, for example, when "solar power capacity" may refer to cumulative capacity or newly added capacity, it issues an "Ambiguity Detected" prompt. The system then suggests either retaining the original attribute or splitting it into more precise columns along additional dimensions, allowing the user to actively guide the data structuring process.

Visual Exploration and Lens-based Inspection. Once the table is populated, the system provides rich visual analytics support. By default, the table presents data as the probability distribution encoding to reflect variance or uncertainty. To facilitate intuitive high-level comparisons across countries, the user can switch to the CCDF bar encoding (Figure 6 C2), which visually encodes data magnitude, for example, when comparing population or wind cumulative power across entities. Furthermore, for cells containing complex or conflicting values, the user can activate the semantic lens (Figure 6 C1). Hovering over or focusing on a specific area opens a detailed lens that juxtaposes distinct extracted values (e.g., 12.40 vs. 347.00) with their corresponding trust scores (computed as described in subsection IV-D) and evidence sentences, helping users quickly assess data reliability.

Multi-Source Traceability and Verification. Ensuring data accuracy is critical in policy research. When a user wants to audit a specific cell's value (e.g., Germany's solar cumulative capacity of 117 GW), they can click on the corresponding data node (as indicated by 1). This action expands the traceability inspector (Figure 6 A) on the left. This panel visualizes the "Consensus Peaks" of the extracted values (e.g., clusters around 117 GW and 106.2 GW). Crucially, it lists the exact provenance for each data point, including source URLs (e.g., *emi-bg.com*), the original textual context, and the extracted time and attribute tags. This transparency allows researchers to verify and select data sources for further analysis.

TABLE II
PERFORMANCE OF DIFFERENT LLMS ON EVIDENCE EXTRACTION AND RECORD-LEVEL EVIDENCE AUDITING.

Family	Model	Evidence Extraction			Evidence Auditing	
		Value Acc. (%)	Unit Acc. (%)	Evid. Acc. (%)	Audit Acc. (%)	Avg. Time (s)
GPT	4o-mini	92.50	92.00	89.50	91.04	3.69
GPT	4o	94.00	89.50	90.00	96.89	1.99
GPT	5.4-mini	92.50	**97.00**	92.50	93.24	**1.66**
GPT	5.4	97.50	91.00	**97.50**	96.34	2.23
Claude	Sonnet-5	**99.00**	95.50	95.00	95.25	5.98
Gemini	3.1-Pro-Preview	95.00	88.00	89.50	**98.17**	9.78
Gemini	3-Flash-Preview	97.50	90.00	94.50	96.53	3.61
Gemini	3.1-Flash-Lite	95.00	91.50	89.00	97.07	3.48
DeepSeek	v4-pro	94.00	93.00	96.00	97.44	8.74
DeepSeek	v4-flash	95.00	93.50	94.00	97.07	28.52
Doubao	seed-1-6-251015	80.50	84.00	85.50	97.99	34.74

B. Quantitative Evaluation

We evaluate two components of the data engine (section IV): evidence extraction and record-level auditing within the conflict-resolution stage. The results are summarized in Table II. The extraction benchmark comprises 200 source–target cases from 186 cleaned source files across 12 attributes. Each provides a target, relevant text, and a reference value, unit, and evidence sentence. Cases were selected with seed 203, using a confidence-based heuristic and query/source diversity. Input lengths are 991–4,776 characters (median 3,785). The inputs are evidence windows rather than full retrieved pages. Given the query target and source text, each LLM was prompted to return a structured `JSON` output containing the extracted value, unit, and evidence sentence. We compared the outputs with the human annotations and report value, unit, and evidence accuracy under *Evidence Extraction*. For conflict resolution, each benchmark case contains an expected query target and a list of candidate extracted records, each with an extracted entity, attribute, value, unit, and textual evidence. Following the conflict types defined in subsection IV-C, each model was asked to assign one audit label to each candidate record. We evaluated this module on 200 audit cases containing 547 records, including 200 with injected errors (Supplement A.4). A prediction was counted as correct when the assigned label matched the human-annotated cause of the discrepancy. Three experts cross-annotated the records, resolved disagreements through discussion, and agreed on acceptable labels before inspecting model outputs. When two diagnoses were acceptable, either counted as correct. We report item-level audit accuracy and average inference time under *Evidence*

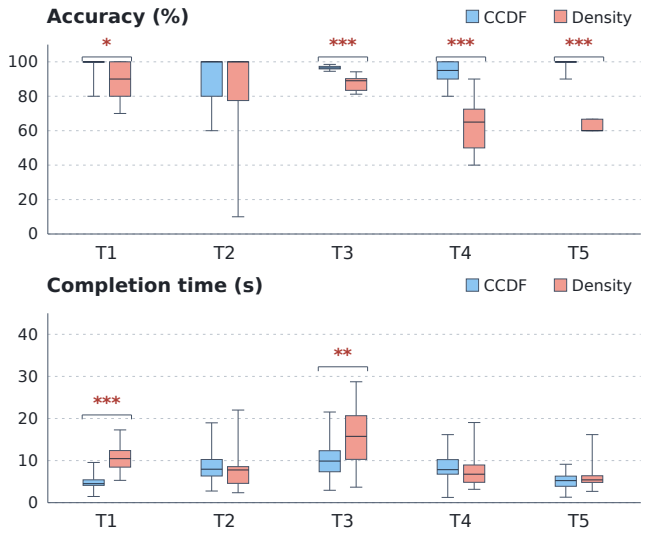


Fig. 8. Controlled user study results comparing the CCDF-based bar encoding with the density-plot encoding on five uncertainty interpretation tasks. Boxes summarize participant-level means: the box spans the 25th–75th percentiles, the central line is the median, and whiskers mark the minimum and maximum. Significance markers use $*$ for $p < .05$, $**$ for $p < .01$, and $***$ for $p < .001$.

Auditing. The error labels correspond to entity/scope misalignment, value/unit extraction errors, and attribute ambiguity in subsection IV-C. The label `extraction_correct` means that a record matches its source evidence and target context, even when sources disagree. The reported accuracy measures record-level auditing.

Results. Model rankings differed between evidence extraction and record-level auditing (Table II). Gemini 3.1-Pro-Preview led auditing (98.17%). Doubao seed-1-6-251015 scored 80.50% on value extraction and 97.99% on auditing.

C. Perceptual Comparison

We conducted a controlled user study comparing our CCDF-based bar encoding with a density-plot encoding (i.e., the probability distribution encoding described in subsection V-B) on uncertainty interpretation tasks.

Task Design. We designed five tasks to cover the core operations involved in analyzing uncertain distributions. T1 (*Extremes*) asks participants to quickly identify the cell with the maximum or minimum value in a dense 5×5 matrix, evaluating how well an encoding supports rapid visual scanning. T2 (*Consistency*) asks participants to identify the most consistent and the least consistent cells in the matrix. T3 (*Probability Estimation*) asks participants to estimate the probability that a value exceeds a given threshold. T4 (*Cluster Weight*) asks participants to select the cluster with the largest probability mass. The component weights determine the reference answer. T5 (*Entity Comparison*) asks participants to identify the entity with the highest probability of exceeding a given threshold. Reference answers are computed from the tail probabilities of the generated distributions. The stimuli are normalized Gaussian mixtures generated with varying component locations, spreads, and weights. Both encodings reuse the same fixed stimulus sequence and answers. T1 and T2 each contain five trials per subgoal. The tutorial explains density area and CCDF height with worked examples and revealed answers. Before each task category, participants familiarized themselves with example cases.

Experimental Setup and Statistical Analysis. A total of 12 participants were recruited for the study. Each participant completed all five tasks using both the CCDF and Density encodings. The five task categories were presented in a fixed order. Within each category, participants completed the trials for one encoding as a block, followed by the trials for the other encoding, before proceeding to the next category. Encoding order was counterbalanced across participants: six used CCDF first and six used Density first. Ten trials were scheduled per task and encoding (100 per participant). Participant means were computed from available trials without imputation. We recorded task completion time and accuracy for each task. For statistical analysis, we first averaged results within each participant for each task, and then conducted two-tailed paired t-tests on participant-level means. Reported p values use the Student t distribution with 11 degrees of freedom and are unadjusted for multiple comparisons. For the probability estimation task (T3), we computed accuracy as $100 -$ absolute percentage-point error.

Results. As shown in Figure 8, CCDF achieved significantly higher accuracy in T1 (*Extremes*, $p < .05$), T3 (*Probability Estimation*, $p < .001$), T4 (*Cluster Weight*, $p < .001$), and T5 (*Entity Comparison*, $p < .001$); no significant difference was observed for T2 (*Consistency*). For completion time, CCDF was significantly faster in T1 ($p < .001$) and T3 ($p = .002$), with no significant time differences in T2, T4, or T5.

D. User Study

To collect usability feedback from users with diverse but relevant backgrounds, we recruited 12 participants (P1–P12, comprising 8 males and 4 females) from a range of academic and professional fields, including data analysis, computer science, international trade, business engineering, and software development. We treated participants with either relevant professional roles or at least three years of experience in visualization, data analysis, software development, or web-based information organization as expert users. By this criterion, 7 participants (P1, P2, P3, P6, P7, P11, P12) were expert users. This cohort included two visualization experts (P3, P11), three data analysts (P1, P2, P7), two software developers (P6, P12), four graduate students (P4, P8, P9, P10), and one undergraduate student (P5). All participants reported substantial experience in retrieving unstructured data from the Internet and performing data cleaning and analysis in their daily academic research or professional work. We also collected participants' experience with data visualization tools; on average, they reported 4.08 years of experience. The study was conducted in a local experimental setting. Participants used a 27-inch monitor and interacted with the system through a Chrome browser.

Experimental Procedure. Each session lasted approximately 45 minutes. Researchers first introduced the study

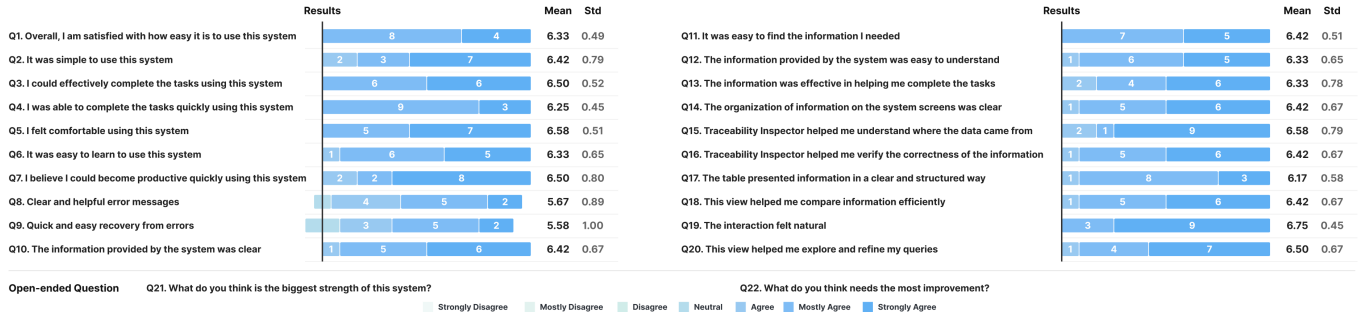


Fig. 9. Quantitative results of the post-study questionnaire. The figure illustrates the score distributions, means, and standard deviations for the 20 closed-ended questions (Q1-Q20) evaluated on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree).

background and the system workflow: users submit natural-language queries, and the system retrieves web evidence, structures numerical records, and aligns them for inspection. Participants then completed a 10-minute tutorial covering the core interactions, including query input, backend crawling, cell-level inspection, uncertainty visualization, and provenance tracking. After 10 minutes of free exploration, participants formulated multi-dimensional comparison queries based on their own interests or professional backgrounds. Their queries covered diverse real-world scenarios, including cross-country housing prices, national land areas, macroeconomic price levels, new energy vehicle sales, and monthly active users (MAU) of popular games.

After submitting a query in the chat interface, the system first parsed the intent and prompted users to confirm the inferred retrieval scope. For queries with approximately ten entities and ten attributes (about 100 cells), observed wall-clock time was 1–2 minutes, from scope confirmation to completion of crawling, extraction, and aggregation for the entire table. Cells run concurrently, and results are returned and displayed as soon as each cell finishes processing. Users can inspect completed cells while the remaining cells are still being processed. Additional cells increase processing and token consumption. The token cost of *DataRecycle* mainly comes from input evidence snippets rather than full web pages or long-form generation. For a query with m attributes and n entities, if each cell retains at most s sources and each evidence snippet contains k tokens on average, an illustrative snippet-input estimate is $T_{in} \approx mnsk$. The output is usually short structured JSON; if each cell produces q output tokens on average, then $T_{out} \approx mnq$. Using illustrative rates of USD 2.50 per million input tokens and USD 15.00 per million output tokens, the corresponding token-cost component is $C \approx 2.5T_{in}/10^6 + 15T_{out}/10^6$. In practice, the same report may support multiple entities or attributes, and DataRecycle deduplicates retained sources across cells. If u denotes the number of unique retained evidence snippets after deduplication, the input scale is closer to uk, where $u \leq mns$. This snippet-token estimate excludes search-service fees, crawling, and additional parsing, auditing, retry, and re-extraction calls.

After the data was retrieved, participants completed the following three steps: (1) *Overview inspection:* participants examined the table view to understand the overall distribution, including both the CCDF bar encoding and the probability distribution visualization, and assessed whether the results could answer their query; (2) *Focused exploration:* participants interacted with the table using the semantic lens to inspect specific cells in the table view; and (3) *Verification and provenance tracking:* participants inspected individual cells using the traceability inspector and followed source links to examine data provenance. After completing the tasks, participants completed a post-study questionnaire (see Figure 9). The questionnaire included items adapted from PSSUQ [57]. It contained 20 closed-ended Likert-scale questions (Q1-Q20) evaluating system effectiveness, efficiency, and usability, as well as two open-ended questions (Q21, Q22) regarding *useful features* (Q21) and *suggestions for improvement* (Q22), respectively.

Results. Most questionnaire items had mean scores above 6.0 on the 7-point scale (Figure 9). The item on natural interaction in the chat view had the highest mean (Q19, $M = 6.75$). P12 noted that the unified table made it faster to compare sources and check values. Ten participants identified provenance as useful in Q21. Participants rated the inspector as helpful for understanding data sources (Q15, $M = 6.58$) and checking accuracy (Q16, $M = 6.42$). P8 contrasted checking evidence directly in the inspector with manually extracting and verifying facts from long AI-generated answers. We submitted 12 comparative factual queries to ChatGPT with web search enabled and checked the answers for factual accuracy, support from sources, and source coverage. Six answers contained factual errors or claims that their sources did not support. Three others relied heavily on a single source or reproduced values from an existing table.

VII. Discussion and Future Work

Limitations. *DataRecycle* organizes numerical web evidence into comparable tables with links to the sources. Several limitations remain. **1) Evaluation scope.** The module benchmark measures extraction and auditing, the perceptual study tests how participants interpret the encodings, and the user study captures perceived usefulness. Even when retrieved sources agree, search rankings and limited retrieval depth can leave out relevant or contradictory reports. We did not compare evidence coverage, contradiction detection, or cost across retrieval depths. For source weighting, we tested how rankings change as the weights vary. We have yet to compare the scores with independent assessments of source reliability or compare

verification outcomes under uniform or alternative weights. In the user study, we did not systematically assess whether participants correctly detected or corrected errors. We also did not compare final table accuracy, evidence coverage, or verification time in a controlled study. **2) Schema omissions.** Omitted attributes may leave no empty cells. Users can add them by refining the schema. The module benchmark uses manually specified targets and supplied texts, so it cannot assess retrieval coverage, initial schema completeness, or whether users find and recover omissions. **3) Source dependence.** Paragraph deduplication cannot identify dependent sources. Weighting republications and paraphrases separately can overstate support for the same evidence. Grouping dependent sources before weighting could reduce this effect.

Context-Aware Data via Local RAG. *DataRecycle* currently focuses on open-web data, while many organizations and individuals maintain large volumes of private "dark data" [58], such as business logs, analytical reports, emails, and historical dashboards. A promising direction is to incorporate local RAG mechanisms for structuring such private, heterogeneous repositories without exposing sensitive content to external LLM services. This extension requires clear processing boundaries: sensitive data could be retrieved and modeled locally, whereas abstracted or sanitized representations could be passed to cloud-based LLMs for further reasoning. Such local–cloud collaboration may help balance organizational context, analytical capability, and data security.

Beyond Numerical Representations. *DataRecycle* currently focuses on numerical data, where facts can be aligned in a shared ontology space and presented in tabular form. However, many real-world scenarios are better expressed through links and relationships among entities, forming relational structures that require conceptual disambiguation. A future direction is to extend the ontology space beyond attribute-value pairs to represent entity relationships and graph-structured data. This extension would support richer relational semantics, but also introduces challenges in modeling higher-order relationships and maintaining semantic alignment across heterogeneous sources.

Inspecting Potential Source Manipulation. Recent research highlights the risk of manipulating content retrieved by generative engines [3], [59]. In such attacks, malicious actors flood the web with manipulated content to intentionally contaminate the retrieval pool. Retrieved content can then lead to misleading answers, for example, a scam phone number [60]. The provenance and duplication views let users trace apparent agreement back to its sources and inspect repeated or conflicting reports. Future work could test whether displaying source relationships and publication timing helps analysts spot coordinated manipulation, such as sudden spikes in publication volume or clusters of highly repetitive reports.

VIII. Conclusion

This work addresses the challenge of transforming scattered web evidence into verifiable structured data for analysis. *DataRecycle* adopts a schema-guided, top-down retrieval strategy that targets explicit table cells and makes coverage gaps visible, together with a provenance-centered design that makes the data lineage from raw snippets to final cells inspectable. The evaluations indicate that *DataRecycle* supports evidence extraction and helps users inspect and interpret conflicts across sources. By combining algorithmic trust estimation with interactive provenance inspection and schema refinement, *DataRecycle* helps users stay in control of the analytical process. Looking ahead, several directions may extend the system. Incorporating local retrieval-augmented generation could support context-aware structuring over private organizational data while respecting privacy constraints. Extending the system beyond numerical tables could support richer relational representations in a broader ontology space.

Acknowledgment

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (NTU Tier 1: RG104/25) and Academic Research Fund Tier 2 (Proposal ID: T2EP20222-0049). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

[1] G. Weikum, X. L. Dong, S. Razniewski, and F. Suchanek, "Machine knowledge: Creation and curation of comprehensive knowledge bases," *Found. Trends Databases*, vol. 10, no. 2–4, pp. 108–490, Jul. 2021. [Online]. Available: https://doi.org/10.1561/1900000064

[2] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques," *Know.-Based Syst.*, vol. 70, no. C, pp. 301–323, Nov. 2014. [Online]. Available: https://doi.org/10.1016/j.knosys.2014.07.007

[3] P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, and A. Deshpande, "GEO: Generative engine optimization," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2024, pp. 5–16.

[4] M. Karunathilaka, L. Lei, Y. Gao, Y. Wang, and J. Li, "Compendia: Automated visual storytelling generation from online article collection," *IEEE Transactions on Visualization and Computer Graphics*, 2026.

[5] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *Proceedings of the fifth ACM conference on Digital libraries*, 2000, pp. 85–94.

[6] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web," *Communications of the ACM*, vol. 51, no. 12, pp. 68–74, Dec. 2008. [Online]. Available: https://doi.org/10.1145/1409360.1409378

[7] C. Zhang, C. Ré, M. Cafarella, C. De Sa, A. Ratner, J. Shin, F. Wang, and S. Wu, "Deepdive: declarative knowledge base construction," *Commun. ACM*, vol. 60, no. 5, pp. 93–102, Apr. 2017. [Online]. Available: https://doi.org/10.1145/3060586

[8] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: a web-scale approach to probabilistic knowledge fusion," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 601–610.

[9] S. Zhang, L. Wang, T. J.-J. Li, Q. Shen, Y. Cao, and Y. Wang, "ChartifyText: Automated chart generation from data-involved texts via LLM," *arXiv preprint arXiv:2410.14331*, 2024.

[10] Y. Huang, Y. Miao, D. Weng, A. Perer, and Y. Wu, "Structvizor: Interactive profiling of semi-structured textual data," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 2025, pp. 1–18.

[11] E. Tsai, N. Mangaokar, B. Zheng, H. Zheng, and A. Prakash, "Harmful terms and where to find them: Measuring and modeling unfavorable financial terms and conditions in shopping websites at scale," in *Proceedings of the ACM Web Conference*. ACM, 2025, pp. 990–1003.

[12] S. Y.-T. Lee and K.-L. Ma, "Hints: Sensemaking on large collections of documents with hypergraph visualization and intelligent agents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 9, pp. 5532–5546, 2025.

[13] H. Li, Y. Ren, Y. Cao, Y. Li, F. Fang, Z. Lin, and S. Wang, "Bridging the gap: Aligning language model generation with structured information extraction via controllable state transition," in *Proceedings of the ACM Web Conference*. ACM, 2025, pp. 1811–1821.

[14] Z. Ni, P. Fan, S. Dai, B. Zhang, H. Wan, and X. Zhao, "Fg-cibgc: A unified framework for fine-grained and class-incremental behavior graph classification," in *Proceedings of the ACM Web Conference*. ACM, 2025, pp. 1166–1181.

[15] B. Chen, J. Lyu, T. Wu, M. B. Chhetri, and G. Bai, "Semantics-aware cookie purpose compliance," in *Proceedings of the ACM Web Conference*. ACM, 2025, pp. 1602–1613.

[16] Q. Chen, Y. Wang, Y. Yu, X. Zhu, X. Yu, and R. Wang, "Conch: Competitive debate analysis via visualizing clash points and hierarchical strategies," *IEEE Transactions on Visualization and Computer Graphics*, 2025, early Access.

[17] L. Ferreira, G. Moreira, and F. Miranda, "Va-blueprint: Uncovering building blocks for visual analytics system design," *IEEE Transactions on Visualization and Computer Graphics*, 2025, early Access.

[18] X. Wang, S. L. Huey, R. Sheng, S. Mehta, and F. Wang, "SciDaSynth: Interactive structured data extraction from scientific literature with large language model," *Campbell Systematic Reviews*, vol. 21, no. 4, p. e70073, 2025.

[19] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *SIGKDD Explor. Newsl.*, vol. 17, no. 2, pp. 1–16, Feb. 2016. [Online]. Available: https://doi.org/10.1145/2897350.2897352

[20] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[21] P. Kaghazgaran, J. Caverlee, and A. Squicciarini, "Combating crowdsourced review manipulators: A neighborhood-based approach," in *Proceedings of the ACM International Conference on Web Search and Data Mining*, ser. WSDM '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 306–314. [Online]. Available: https://doi.org/10.1145/3159652.3159726

[22] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han, "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 745–754.

[23] Y. Li, H. Sun, and W. H. Wang, "Towards fair truth discovery from biased crowdsourced answers," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020, pp. 399–409.

[24] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen, "Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 31–40, 2016.

[25] A. Dasgupta, J.-Y. Lee, R. Wilson, R. A. Lafrance, N. Cramer, K. Cook, and S. Payne, "Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 271–280, 2017.

[26] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren, "The state of the art in enhancing trust in machine learning models with the use of visualizations," *Computer Graphics Forum*, vol. 39, no. 3, pp. 713–756, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14034

[27] N. Cao, C. Shi, S. Lin, J. Lu, Y.-R. Lin, and C.-Y. Lin, "Targetvue: Visual analysis of anomalous user behaviors in online communication systems," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 280–289, 2015.

[28] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2649–2658, 2012.

[29] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins, "#FluxFlow: Visual analysis of anomalous information spreading on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1773–1782, 2014.

[30] S. Ortona, V. V. Meduri, and P. Papotti, "Robust discovery of positive and negative rules in knowledge bases," in *Proceedings of the IEEE International Conference on Data Engineering*, 2018, pp. 1168–1179.

[31] C. Ye, H. Wang, K. Zheng, Y. Kong, R. Zhu, J. Gao, and J. Li, "Constrained truth discovery," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4859–4872, 2022.

[32] C. Bors, T. Gschwandtner, S. Miksch, and J. Gärtner, "Qualitytrails: Data quality provenance as a basis for sensemaking," in *Proceedings of the International Workshop on Analytic Provenance for Sensemaking*, 2014.

[33] Y. Li and Y. Wang, "FOX: Visual exploration of data fact outliers," *arXiv preprint arXiv:2608.08671*, 2026.

[34] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha, "Approaches to uncertainty visualization," *The Visual Computer*, vol. 13, no. 8, pp. 370–390, 1997.

[35] H. Song and D. A. Szafir, "Where's my data? evaluating visualizations with missing data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 914–924, 2019.

[36] N. D. Gershon, "Visualization of an imperfect world," *IEEE Computer Graphics and Applications*, vol. 18, no. 4, pp. 43–45, 1998.

[37] M. Correll and M. Gleicher, "Error bars considered harmful: Exploring alternate encodings for mean and error," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2142–2151, 2014.

[38] M. Lu, S. Wang, J. Lanir, N. Fish, Y. Yue, D. Cohen-Or, and H. Huang, "Winglets: Visualizing association with uncertainty in multi-class scatterplots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 770–779, 2020.

[39] J. Sanyal, S. Zhang, G. Bhattacharya, P. Amburn, and R. J. Moorhead, "A user study to compare four uncertainty visualization methods for 1d and 2d datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1209–1218, 2009.

[40] T. M. Athawale, Z. Wang, Q. Gong, S. Klasky, K. Moreland, D. Pugmire, C. R. Johnson, and P. Rosen, "Uncertainty visualization of critical points of 2d scalar fields for parametric and nonparametric probabilistic models," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[41] N. Boukhelifa, A. Bezerianos, T. Isenberg, and J.-D. Fekete, "Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2769–2778, 2012.

[42] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler, "Visualizing geospatial information uncertainty: What we know and what we need to know," *Cartography and Geographic Information Science*, vol. 32, no. 3, pp. 139–160, 2005. [Online]. Available: https://doi.org/10.1559/1523040054738936

[43] K. Potter, P. Rosen, and C. R. Johnson, "From Quantification to Visualization: A Taxonomy of Uncertainty Visualization Approaches," in *Uncertainty Quantification in Scientific Computing*, A. M. Dienstfrey and R. F. Boisvert, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 226–249.

[44] M. Correll, D. Moritz, and J. Heer, "Value-suppressing uncertainty palettes," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–11. [Online]. Available: https://doi.org/10.1145/3173574.3174216

[45] M. Correll and J. Heer, "Surprise! bayesian weighting for de-biasing thematic maps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 651–660, 2017.

[46] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson, "When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, New York, NY, USA, 2016, pp. 5092–5103. [Online]. Available: https://doi.org/10.1145/2858036.2858558

[47] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay, "Uncertainty displays using quantile dotplots or cdfs improve transit decision-making," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–12. [Online]. Available: https://doi.org/10.1145/3173574.3173718

[48] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady, "explAIner: A visual analytics framework for interactive and explainable machine learning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1064–1074, 2020.

[49] F. Cheng, D. Liu, F. Du, Y. Lin, A. Zytek, H. Li, H. Qu, and K. Veeramachaneni, "VBridge: Connecting the dots between features and data to explain healthcare models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 378–388, 2022.

[50] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios, "DataTone: Managing ambiguity in natural language interfaces for data visualization," in *Proceedings of the ACM Symposium on User*

Interface Software and Technology. New York, NY, USA: Association for Computing Machinery, 2015, pp. 489–500. [Online]. Available: https://doi.org/10.1145/2807442.2807478

[51] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang, "Eviza: A natural language interface for visual analysis," in *Proceedings of the ACM Symposium on User Interface Software and Technology*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 365–377. [Online]. Available: https://doi.org/10.1145/2984511.2984588

[52] T. Gschwandtnei, M. Bögl, P. Federico, and S. Miksch, "Visual encodings of temporal uncertainty: A comparative user study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 539–548, 2016.

[53] D. Ren, M. Brehmer, B. Lee, T. Höllerer, and E. K. Choe, "ChartAccent: Annotation for data-driven storytelling," in *Proceedings of the IEEE Pacific Visualization Symposium*, 2017, pp. 230–239.

[54] I. ISO. and B. OIML, *Guide to the Expression of Uncertainty in Measurement*. Aenor Madrid, Spain, 1993.

[55] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang, "Knowledge-based trust: Estimating the trustworthiness of web sources," *Proc. VLDB Endow.*, vol. 8, no. 9, pp. 938–949, 2015. [Online]. Available: http://www.vldb.org/pvldb/vol8/p938-dong.pdf

[56] R. Y. Wang and D. M. Strong, "Beyond accuracy: what data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, Mar. 1996.

[57] J. R. Lewis, "Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ," in *Proceedings of the human factors society annual meeting*, vol. 36, no. 16, 1992, pp. 1259–1260.

[58] A. Corallo, A. M. Crespino, V. D. Vecchio, M. Lazoi, and M. Marra, "Understanding and defining dark data for the manufacturing industry," *IEEE Transactions on Engineering Management*, vol. 70, no. 2, pp. 700–712, 2023.

[59] R. Mochizuki, S. Komatsu, S. Noguchi, and K. Ataka, "Exposing citation vulnerabilities in generative engines," *arXiv preprint arXiv:2510.06823*, 2025.

[60] Q. Deng and A. A. Labs, "When AI recommends scammers: New attack abuses LLM indexing to deliver fake support numbers," https://aurascape.ai/llm-search-poisoning-fake-support-numbers/, Dec. 2025, accessed: 2026-03-29.



Can Liu is an assistant professor at the Central University of Finance and Economics. He was previously a research fellow in the College of Computing and Data Science at Nanyang Technological University. He received his Ph.D. in Intelligent Science (2023), B.Sc. in Computer Science (2018), and B.Ec in Economics (2018) degrees from Peking University. His research interests lie in the field of deep learning-driven visualization, especially intelligent interaction for visualization. Refer to https://vvoliucano.github.io for more details.



Tengteng Zhu received the B.S. degree from Northwest Agriculture and Forestry University in 2017, and the M.S. degree from Beijing Information Science and Technology University in 2021. He is currently pursuing a Ph.D. degree at the Beijing Institute of Technology.



Xinyu Jiang is a Master of Engineering (Software Engineering) at the University of Sydney, Australia. He received his B.Sc. degree in Software Engineering from Beijing Institute of Technology. His research interests include intelligent interactive systems, multimodal AI, and visual analytics.



Xiaolin Wen is currently a Ph.D. student in the College of Computing and Data Science, Nanyang Technological University (NTU). His research interests include on visualization for FinTech and LLM-assisted design study. He received his master's degree in Computer Science and Technology from Sichuan University in 2023 and his dual bachelor's degree in Computer Science and Financial Engineering from Sichuan University in 2016. For more information, kindly visit https://wenxiaolin.com/.



Yong Wang is currently an assistant professor in the College of Computing and Data Science, Nanyang Technological University. Before that, he worked as an assistant professor at Singapore Management University from 2020 to 2024. His research interests include data visualization, HCI, and human-AI collaboration, with an emphasis on their application to FinTech, quantum computing and online learning. He obtained his Ph.D. in Computer Science from Hong Kong University of Science and Technology. He received his B.E. and M.E. from Harbin Institute of Technology and Huazhong University of Science and Technology, respectively. For more details, please refer to http://yong-wang.org.