# *Compendia*: Automated Visual Storytelling Generation from Online Article Collection

Manusha Karunathilaka, Litian Lei, Yiming Gao, Yong Wang, and Jiannan Li

***Abstract*—In the digital age, readers value quantitative journalism that is clear, concise, analytical, and human-centred. To understand complex topics, they often piece together scattered facts from multiple articles. Visual storytelling can transform fragmented information into clear, engaging narratives, yet its use with unstructured online articles remains largely unexplored. To fill this gap, we present *Compendia*, an automated system that analyzes online articles in response to a user's query and generates a coherent data story tailored to the user's informational needs. *Compendia* addresses key challenges of storytelling from unstructured text through two modules covering: *Online Article Retrieval*, which gathers relevant articles; *Data Fact Extraction*, which identifies, validates, and refines quantitative facts; *Fact Organization*, which clusters and merges related facts into coherent thematic groups; and *Visual Storytelling*, which transforms the organized facts into narratives with visualizations in an interactive scrollytelling interface. We evaluated *Compendia* through a quantitative analysis, confirming the accuracy in fact extraction and organization, and through two user studies with 16 participants, demonstrating its usability, effectiveness, and ability to produce engaging visual stories for open-ended queries.**

***Index Terms*—Data Storytelling, Scrollytelling, Text/Document Data**

## I. Introduction

In today's information-rich era, people rely on diverse online sources, such as news articles, blogs, reports, and expert analyses, to stay updated on events and facts [1], [2]. Audience studies indicate that readers particularly value quantitative journalism that is constructive, concise, analytical, and human-centred [3], and visualizations have been widely used to convey such quantitative information clearly and engagingly [4]. Following this trend and growing audience demand from news consumers and analysts alike, major information outlets, such as The New York Times, The Guardian, and Pew Research Center[1], have made statistics and explanatory visual journalism a core part of their reporting. Some of them maintain dedicated data-driven sections, including The Guardian's Datablog[2] and New York Times' Upshot[3].

However, when approaching complex phenomena with multiple facets, readers, such as news consumers and analysts, often have to consult multiple sources to gain a more complete picture [2]. For example, a reader who is trying to understand the climate impact of artificial intelligence (AI) might need to integrate technology landscapes, energy estimates, and carbon footprint assessments, which are scattered across news articles, company white papers, and government reports. Similarly, an analyst studying global poverty might draw on income distributions and social welfare policies from multiple government websites and think tank reports. In practice, readers must iteratively refine web search queries, carefully scan extensive contents to extract relevant quantitative information, and then relate findings across sources to construct a coherent understanding [5]. This manual effort creates substantial cognitive burden [3], often obscuring overarching themes and making it difficult to reconcile related facts into a clear mental model.

To ease the burden, existing aggregator systems, such as Google News, Newsblaster [6], and NewsBird [7], consolidate content from multiple articles into a single interface. Yet, they tend to focus on analyzing surface-level information, such as headlines and short summaries, leaving readers to perform the cognitively demanding task of synthesizing quantitative information scattered across sources. Similarly, AI-powered search engines like Perplexity[4] produce cited summaries of web content, but they remain largely text-centric and lack the quantitative synthesis and visual scaffolding that readers strongly prefer when navigating complex, data-rich topics [3].

Visual storytelling transforms complex topics into coherent narratives that combine quantitative evidence with visual presentations, enhancing the comprehension, engagement, and recall of the underlying data and insights [8]–[10]. Automated visual story generation saves significant manual efforts, but prior work on automated visual storytelling required structured data (e.g., tables) [11]–[13] or a single article as source [14], [15] and lacked the ability to integrate information across multiple articles [4]. We address this gap by combining automated text analysis with visual data story generation over online article collections, aiming to help readers synthesize quantitative facts and form a coherent mental model of a topic in support of their analytical tasks.

Generating visual data stories from a collection of online articles remains largely unexplored and intrinsically challenging. The challenges of generating a coherent data story from online article collections originate from the fact that these data sources are often vastly **diverse** and **unstructured**. First, relevant data facts are often scattered across multiple articles and sections, mixed with irrelevant content, and expressed in varied formats and units (e.g., 3.7K/3700/3,700). *Extracting* data facts from these sources requires a systematic

M. Karunathilaka and J. Li are with Singapore Management University. Email: gmik.vidana.2023@phdcs.smu.edu.sg, jiannanli@smu.edu.sg.

L. Lei, Y. Gao, and Y. Wang are with Nanyang Technological University. Email: {litian001, gaoy0053}@e.ntu.edu.sg, yong-wang@ntu.edu.sg.

J. Li and Y. Wang are the corresponding authors.

[1]https://www.pewresearch.org/

[2]https://www.theguardian.com/news/datablog

[3]https://www.nytimes.com/international/section/upshot

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

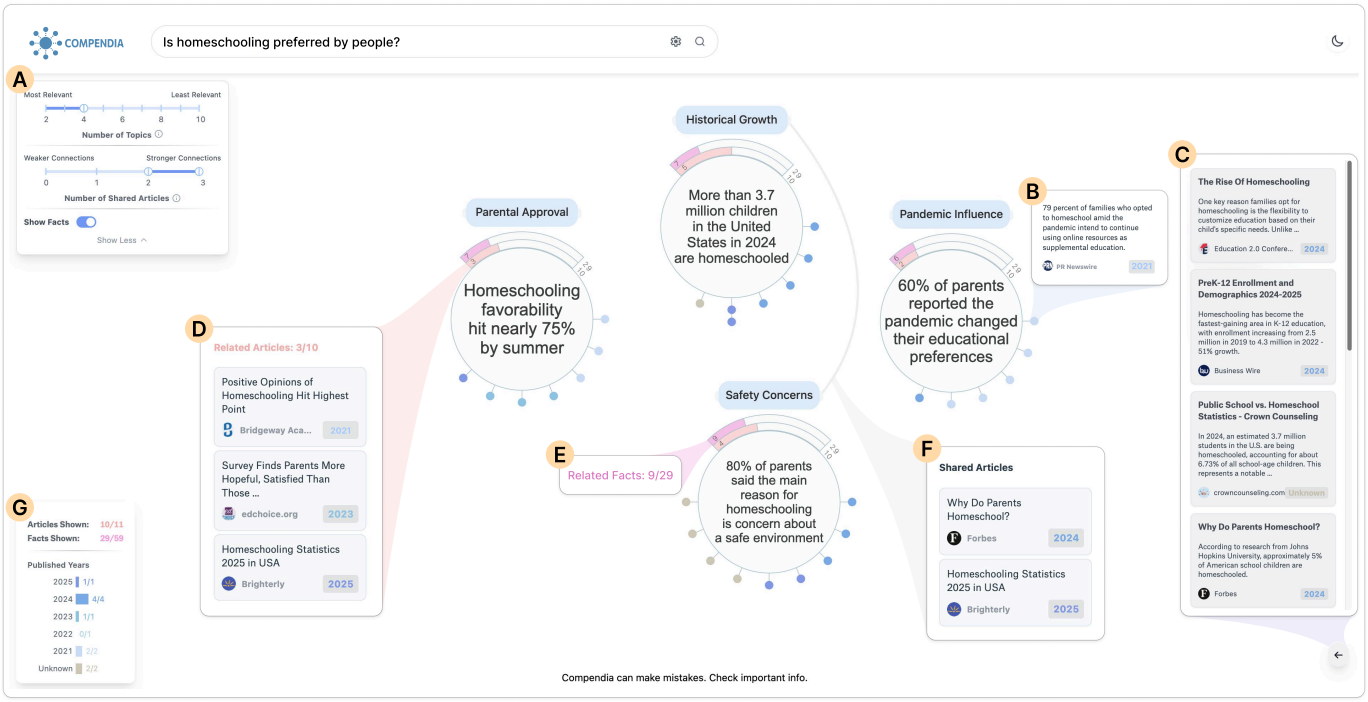



Figure 1. *Compendia* transforms the query *"Is homeschooling preferred by people?"* into a structured data story by extracting, clustering, and visualizing key facts using unstructured data from a collection of online articles. **Thematic Overview** uses **Thematic Circles** to visualize clustered facts across different themes. (A) Filter widget provides control over the overview, (B) Detailed fact panel provides fact content and source details, (C) Articles panel presents all retrieved articles relevant to the story, (D) Related Articles panel displays articles relevant to the topic, (E) Related facts panel provides the number of facts belonging to the topic, (F) Shared articles panel lists sources covering multiple aspects of the topic, and (G) Summary panel displays article and fact statistics.

approach to identify key insights while effectively filtering out irrelevant or noisy information. Second, online articles often differ significantly in their themes, angles, and writing styles. Such diversity enhances the richness of data stories but also complicates the task of *organizing* fragmented facts into clear, coherent narratives. Third, the large quantity and structural complexity of the extracted facts require a narrative strategy and a visual *presentation* design that can preserve the richness of the information without overwhelming readers.

To tackle these challenges, we present ***Compendia***, an automated system that retrieves and analyzes collections of online articles in response to a user query and generates a coherent, meaningful visual data story. We propose a framework that includes two main modules: the **Data Fact Extraction and Organization** module and the **Visual Storytelling** module, leveraging Large Language Models (LLMs) to overcome the complexities of unstructured online text and convert it into a coherent data story. The *Data Fact Extraction and Organization* module consists of: **(1)** ***Online Article Retrieval*** gathers relevant articles in bulk using web scraping; **(2)** ***Data Fact Extraction*** filters out irrelevant text and extracts query-relevant quantitative facts, capturing not only their values but also associated units and contextual information to preserve meaning across diverse formats; and **(3)** ***Fact Organization*** clusters multifaceted data facts into thematic groups, distinguishes core from supporting facts, and enhances clarity by merging closely related facts, resolving ambiguities, and ensuring each group forms a coherent narrative unit. The *Visual Storytelling* module weaves narrative units into a coherent story, presenting a **Thematic Overview** allowing users to form a high-level mental model of the topic, followed by interactive **Story Views** in an *"overview first, details on demand"* manner [16]. Users then drill down into the story through intuitive page scrolling, leveraging the scrollytelling technique [17], [18].

We evaluated *Compendia* through (1) quantitative evaluation, confirming its ability to accurately identify, extract, and organize information from diverse article collections, and (2) two qualitative user studies with 16 participants: an initial perception study establishing *Compendia*'s effectiveness and usability, and an open-ended query exploration study demonstrating its ecological validity and utility in real-world use.

The major contributions of this paper can be summarized as follows:

- **An end-to-end framework** that automatically generates relevant, accurate, and well-organized data stories by extracting and structuring highly varied unstructured textual data from online articles.
- **An interactive visualization system**, *Compendia*, demonstrating the feasibility of generating and presenting data stories from online articles based on user queries.
- **A comprehensive evaluation**, including a quantitative evaluation and two user studies, confirming *Compendia*'s information accuracy and its effectiveness in producing coherent, informative, and engaging data stories.

## II. Related Work

The related work falls into two categories: data storytelling, and information retrieval and extraction techniques.

### A. Data Storytelling

A visual data story is a series of story elements arranged in a meaningful sequence, forming a narrative visualization [10],

[19]. Research in this area has examined diverse genre from annotated charts and comics to timelines and animated narratives, and has shown that storytelling enhances memory, engagement, comprehension, and communication [8]–[10], [20]. Furthermore, data stories have been shown to enhance both the efficiency and effectiveness of comprehension tasks, outperforming conventional visualizations [20]. Recent studies further refine storytelling by exploring narrative roles, AI support levels, and human-AI collaboration [21]–[24]. Yet, creating those data stories poses significant challenges.

Early tools let users manually build these visual stories using templates, such as annotated charts [25], infographics [26], timelines [27], and data comics [28]. Later, systems like Idyll [29] and VisFlow [30] supported dynamic formats such as scrollytelling and slideshows, but they often required design and coding skills. More recent semi-automated tools add smart features that help link text and visuals [31], infographic recommendation [32], slide generation [33], and data story editing [34], but they are primarily designed for expert users like data analysts and designers.

Fully automated systems like DataShot [11] and Calliope [12] can build stories from structured data (e.g., tables and spreadsheets), but they often miss deeper meanings. Similarly, recent research has extended this automation towards generating scrollytelling narratives from structured data [35]. New approaches that use LLMs for narrative generation [13], [36], [37] still rely mostly on structured data, leaving the rich information embedded in unstructured text (e.g., text articles) data underexplored. Furthermore, most of these tools do not explicitly support the direct communication of data stories to audiences and often lack mechanisms to incorporate user intent, limiting their effectiveness in tailoring the narrative to the user's needs [23].

Our work addresses these challenges by automatically generating data stories from unstructured text sources such as online articles, tailored to users' informational needs.

### B. Information Retrieval and Extraction

Information Retrieval (IR) focuses on retrieving relevant information from large-scale text corpora in response to user queries. Traditional IR systems rely on statistical and probabilistic models, such as Term Frequency–Inverse Document Frequency (TF–IDF) [38] and the Okapi BM25 ranking function [39]. These methods compute lexical similarity between queries and documents, providing a foundation for early search engines and document retrieval systems. However, these approaches struggle with semantic understanding, as they primarily depend on keyword matching rather than contextual comprehension. More recent neural retrieval methods, particularly those based on transformer architectures like BERT [40], have improved search effectiveness by leveraging contextual embeddings. However, these methods still require extensive domain-specific training and fine-tuning, making them less adaptable to diverse and evolving information needs.

Beyond retrieval, Information Extraction (IE) is essential in structuring retrieved content by identifying key entities, relationships, and events. Traditional IE methods typically rely on separate models for different tasks, such as Named Entity Recognition (NER), Relation Extraction (RE), and Event Extraction (EE) [41]. However, the need for multiple independent models for different tasks made these approaches resource-intensive. Recent advancements in LLMs, such as GPT-4 [42], have introduced generative IE methods that offer a more scalable alternative by generating structured knowledge directly from text, reducing the need for task-specific models [43]. This shift allows more flexible and efficient extraction from complex, evolving data.

In the context of LLMs, prompt engineering is fundamental for guiding the behavior of LLMs to optimize their performance [44]. Techniques such as Chain-of-thought (CoT) prompting [45], which provides a structured reasoning process to direct model responses, and Few-shot prompting [46], which utilizes in-context learning by incorporating example demonstrations, are commonly applied to improve the performance. *Compendia* leverages LLMs to enhance information retrieval and extraction. Carefully designed prompts ensure that retrieved content is both relevant and structured to support automated data story generation.

## III. Compendia Framework Overview

*Compendia* is the first work to achieve *automated* visual storytelling generation from online articles, to the best of our knowledge. Yet, there has been much research on visual storytelling [10], [17], [18], [47], exploratory research [5], [48]–[50], and information visualization principles [16]. To guide our framework and visualization designs, we have established the following visualization design requirements by conducting an extensive literature review:

- **(R1) Provide an overview of topic clusters:** During the exploratory research or query, users need a quick overview of the topics covered in the online articles to orient themselves and guide further inquiry [5], [48]. This is well-aligned with the visualization mantra *"overview first, details-on-demand"* [16]. Prior work shows that thematic overviews reduce cognitive load and improve sensemaking [49], [50], especially when information is grouped by topic rather than listed flatly [51], [52].
- **(R2) Reveal the connections among topics:** Revealing the correlation between the topics covered help users gain a more structured understanding when exploring multiple related documents [47]. Prior work has also shown that linking articles by shared topics helps trace storylines across documents [53], [54].
- **(R3) Enable smooth navigation between topics and data facts:** There are often multiple topics and data facts covered in a large collection of online articles. Existing research on visual storytelling [10], [17], [18] has emphasized the importance of maintaining data fact continuity and smooth navigation to reduce cognitive load. Also, many online data storytelling examples [55], [56] have often used a map metaphor to enable the transition and navigation among different topics.
- **(R4) Support visualizations with narratives:** Data visualizations coupled with narrative text help readers interpret quantitative facts more effectively than raw facts or

isolated charts [10], [57]. Furthermore, weaving facts into story-driven visuals enhances comprehension, retention, and engagement [20].

**(R5) Enable direct access to facts and their sources:** To promote transparency, users must be able to trace each fact to its original source article. Prior work shows that source attribution improves credibility and trust [58]–[60], with interfaces like TileBars enabling direct access to source passages [61]. We aim to ensure that every fact is displayed with its original details, citation, and publication year.

To satisfy these design requirements, the *Compendia* framework consists of two core modules: the *Data Fact Extraction and Organization* module and the *Visual Storytelling* module, as shown in Fig. 2. Unlike the prior systems such as DataShot [11], which operate on structured table data, *Compendia* works directly on natural language text where quantitative numbers and related descriptions are often scattered across multiple documents. This introduces challenges such as noisy paragraphs, vague expressions (e.g., *"this year," "currently"*), fragmented statements, and inconsistencies in units or time frames. To address these issues, the two modules progressively extract, validate, and organize the data facts, finally presenting them as a data story.

The first module (Fig. 2A) retrieves and processes unstructured text from online articles through three stages. The **Online Article Retrieval** expands the user query with variations and automatically searches and scrapes relevant articles. The **Data Fact Extraction** then filters paragraphs, identifies and extracts quantitative facts with their values, units, and context. Finally, **Fact Organization** clusters semantically related facts and merges coherent facts. These stages parallel the workflow of DataShot [11], but extend it from structured tables to unstructured articles: replacing direct table parsing with LLM-guided fact extraction, redefining fact composition through clustering and merging.

The second module (Fig. 2B) transforms structured fact clusters into interactive narratives, weaving them into a coherent story flow that combines visualization with scrollytelling for effective presentation. This module consists of two views, beginning with the **Thematic Overview** (in Fig.2B), which provides a high-level summary of all extracted data facts as clusters; and followed by the **Story View** (in Fig.2B), accessible by scrolling or selecting a cluster, which enables detailed exploration of data facts in that cluster. We discuss the two modules in the following sections.

## IV. Data fact extraction and organization

Our data fact extraction and organization module (Fig. 2A) processes unstructured text from online articles and transforms them into structured quantitative facts that form the foundation for coherent storytelling. In this section, we describe each stage of the module and the implementation details.

### A. Online Article Retrieval

The online article retrieval stage collects the information sources for data stories. It begins with a user-provided natural language query, which reflects their informational intent. As search results are increasingly affected by spam and ranking manipulation [62], we expand the original query with LLM-generated variants to improve coverage and reduce unrelated content [62]. Based on these queries, *Compendia* uses an online search engine to retrieve relevant articles, emulating how a user would search for information. The source article is tracked at every stage for transparency (***R5***).

### B. Data Fact Extraction

*Compendia*'s focus is to find and extract quantitative evidence that supports the user's query. This stage aims to extract relevant data facts from the retrieved article text, through multiple phases to progressively filter, identify, validate, and refine the data, ultimately generating a well-structured dataset, which is easier to visualize (***R4***).

*1) Paragraph Filtering:* This phase filters the paragraphs related to the user's query from scraped article text, where paragraphs preserve more context than individual sentences and enable better relevance assessment and disambiguation [63]. These texts often contain extraneous information such as headers, footers, references, advertisements, and unrelated content. The system first prompts an LLM to filter the paragraphs using the article title to remove extraneous information, and then re-filters the paragraphs based on the relevance to the query to align with the user's search intent, producing a set of **filtered paragraphs** as foundation for subsequent phases.

*2) Data Fact Identification:* This phase focuses on extracting relevant quantitative information, termed **fact content**, from the *filtered paragraphs* with the aid of an LLM. This process introduces several key challenges. First, distinguishing data facts from factual statements is challenging, as not all factual sentences contain analytically useful quantitative information. For example, the sentence *"Joshua and Samuel are two popular names for boys in 2004"* represents a categorization-type data fact [11], but lacks a numerical value for quantitative analysis or visualization. Therefore, we primarily focus on identifying and isolating quantifiable facts, such as numerical data, measurements, and statistics, while excluding subjective opinions and interpretations. Second, multiple facts may appear within a single paragraph, often interwoven with contextual or explanatory content. Specifically, we focus on identifying sentence or phrase level data facts, where each fact contains at least one data value. For example, *"More than 3.7 million children in U.S. in 2024 are homeschooled"* illustrates a sentence-level data fact with a clearly stated numerical value. Identifying such sentence or phrase level fact content is also crucial for downstream analysis, particularly for grouping similar facts, as outlined in Section IV-C2. To systematically identify data facts, we embed Datashot's [11] quantifiable fact type definitions and examples into the prompt.

*3) Data Extraction:* This phase extracts numerical data by parsing each *fact content*, termed **data points** using an LLM. Several challenges arise during this process. First, numerical values appear in varied writing styles, such as *"3,700,000"*, *"3.7 million"*, or *"3.7M"*. Second, a single fact may contain multiple numerical values, each requiring separate extraction and contextual alignment. For example, the sentence *"From*

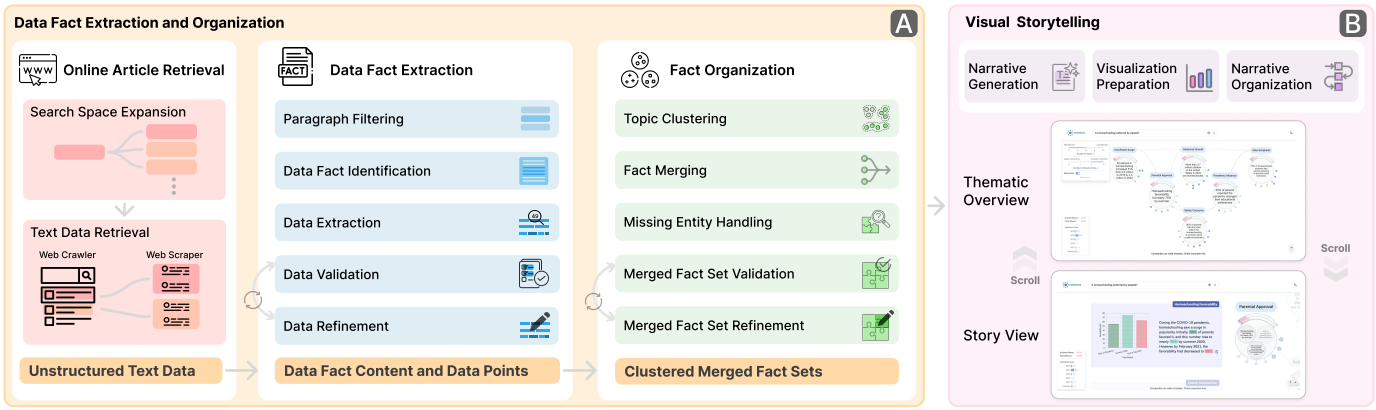



Figure 2. The framework of our system, *Compendia*. It consists of two main modules: A) the Data Fact Extraction and Organization module, which includes Online Article Retrieval, Data Fact Extraction, and Fact Organization; and B) the Visual Storytelling module, which prepares the data for display and presents it as an interactive scrollytelling story. Each stage involves steps performed in the given order, with iterative validation and refinement phases. The boxes highlight the outputs at each stage.

*1999 to 2020, the number of homeschooled students in the U.S. increased from 850,000 students to 2.5 million students"* includes two distinct values associated with different timestamps, which must be captured individually while preserving their temporal semantics. In response, we treat each numerical value individually as a **data point**, extracting them into a unified structured representation consisting of three key fields: **label**, **value**, and **unit**. This returns a set of data points per fact content. For example, given a *fact content "The number of homeschoolers in the U.S. is 3.7 million"*, the system extracts the label *"Homeschooled Children"*, value *"3.7"* with the unit *"million."* Finally, contextual ambiguity in the text can make accurate data extraction challenging. For example, *"this year"* could be misinterpreted as the current year (2025) when the article was actually published in 2024. To resolve such ambiguities, we incorporate article metadata alongside the content in the prompt, which helps clarify ambiguous data.

*4) Data Validation:* This phase assesses the correctness, consistency, and completeness of extracted *fact contents* (Section IV-B2) and *data points* (Section IV-B3) using an LLM and suggests corrections for discrepancies and errors. The process begins by prompting an LLM to verify *fact contents* against the original text and identify discrepancies. Next, it validates the *data points*, prompting an LLM to check their correctness and completeness against the original text data and ensure the unit consistency within data points of a fact. For example, if a paragraph states, *"The number of homeschoolers in the U.S. is 3.7 million,"* but the extracted data point's value is 3, the system detects the error and suggests correcting the value to 3.7. Similarly, if units such as *"billion"* and *"million"* appear inconsistently within a fact's data points, the system flags the inconsistency and recommends consistently normalizing them.

*5) Data Refinement:* This phase corrects errors in *fact contents* and *data points* by prompting an LLM based on validation feedback. The system reviews the feedback by comparing each fact with the source text and then applies minimal edits to preserve original context. Common fixes include adjusting numerical values, normalizing units, and resolving misinterpretations. The validation and refinement phases are performed iteratively to ensure the final structured fact data is correct, consistent, and complete.

### C. Fact Organization

The facts from diverse online articles are often fragmented and inconsistent in style, and vary in detail and focus. This stage clusters facts into themes (***R1***) and merges related facts into coherent narratives (***R4***), reducing redundancy while preserving diversity, yielding a consistent and comprehensive view of the themes drawn from multiple sources.

*1) Thematic Clustering:* This phase facilitates clustering the facts into themes, returning **clustered facts** (***R1***). Those extracted facts originate from a wide range of online articles, each written with different styles, emphases, and levels of detail. As a result, the collected facts are inconsistent in granularity, with some capturing high-level summaries, while others offering fine-grained statistics, yet often sharing common themes. Identifying such thematic overlap across facts is important to organize the fragmented facts.

To organize the extracted facts into meaningful groups, we perform thematic clustering. The aggregate length of the text facts often exceeds the context window limits of LLMs, and chunking the input to fit these limits tends to degrade clustering quality [64]. To better support large-scale clustering in an unsupervised setting, we adopt an embedding-based approach using the Gaussian Mixture Model (GMM) [65]. Specifically, we first obtain vector representations of the *fact contents* using an LLM-based embedding service, and then apply GMM to cluster these embeddings based on their semantic similarity, producing clustered facts. We then generate a concise **cluster topic** using an LLM for each cluster to help users quickly interpret and compare them. However, generating topics independently for clusters can result in ambiguity or overlap due to their conciseness. To improve clarity and differentiation across topics, we first prompt an LLM to generate a brief summary for each cluster. We then provide these summaries, together with the initial topics, back to the LLM to refine them into a final set of distinct, thematically coherent topics.

Furthermore, we calculate each fact's relevance score as the cosine similarity between its embedding and the query embedding. This relevance score helps to identify the most representative fact (highest relevance score) and the top three facts in a cluster. Moreover, we calculate each cluster's rele-

vance score as the average of the fact relevance scores within the cluster, which is then used for filtering (Section V-A2).

*2) Fact Merging:* This phase consists of two steps: first, resolve redundancies by constructing sets of factually similar facts as **fact sets**; and second, merging semantically similar fact sets into **merged fact sets**, within a cluster.

**Resolve Redundancy.** Multiple articles often express the same fact in different wording or styles, yet convey equivalent underlying factual content. For example, *"3.7 million children were homeschooled in 2024"* and *"In 2024, 3,700,000 students homeschooled."* To handle the redundancy, we use an LLM to construct **fact sets** within each cluster based on their factual similarity in *fact content* and *data points*. We observe that semantically similar statements may include conflicting values. Resolving these contradictions requires more sophisticated fact checking, with potential approaches noted in Section VIII.

**Merge Fact Sets.** The extracted facts are often fine-grained, captured at the sentence or phrase level. While this granularity is beneficial for identifying similar facts as discussed above, it also introduces significant challenges such as fragmented information that disrupts narrative coherence and leads to cognitive overload, when facts are presented in isolation. In response, we propose a merging process that combines semantically similar *fact sets* within each cluster. These **merged fact sets** form the basis of the *Narrative Units* discussed in the Section V-B1. This improves the readability and reduces the number of narratives that users have to scroll through.

This merging is challenging due to variations in perspectives, time frames, and units, while also needing to ensure the compatibility with the visualizations that support the narrative (***R4***). To achieve coherence and compatibility, our merging process follows a structured approach within a prompt, incorporating **Unit Consistency**, where all facts use the same measurement units; **Thematic Relevance**, where facts contribute to a cohesive story; **Temporal Consistency**, where facts share logically connected time frames; and **Visualization Readiness**, where data share a consistent x- and y-axis.

*Unit Consistency.* The fact sets are grouped by unit type to maintain consistency, while normalizing the units when necessary (e.g., *"3,000,000"* $\rightarrow$ *"3 million"*). The fact sets with different measurements or unit types are not merged. For example, *"Three million children are homeschooled"* cannot be merged with *"20% of students were homeschooled,"* as one uses a count and the other a proportion.

*Thematic Relevance.* The *fact sets* are merged when they share a common theme. For example, *"23.1% said that the reason for homeschooling was the child's special needs"* and *"15.6% of parents said that the child had a physical or mental problem"* can be merged as: *"23.1% of parents cited their child's special needs as the reason for homeschooling, while 15.6% pointed to physical or mental health challenges."*

*Temporal Consistency.* The *fact sets* are merged only when they share compatible time ranges; those from different periods remain separate. For example, *"Homeschooling increased by 25% between 2020 and 2021"* cannot be merged with *"Homeschooling rates have doubled over the past decade,"* as the first reflects a short-term trend and the second a long-term pattern.

*Visualization Readiness.* Finally, we assess whether the *merged fact set* is compatible with a visual representation, ensuring that the data can be aligned along common axes (e.g., consistent time or category labels on the x-axis and comparable values on the y-axis).

*3) Missing Entity Handling:* After obtaining the *merged fact sets*, we often encounter missing entities, such as geographical locations, time references, or subject identifiers, that are essential for clarity. These gaps can make the narrative ambiguous or incomplete and typically arise from the fragmented nature of fact extraction, where individual sentences may omit context present in the original source. Identifying all necessary contextual information during the initial extraction stage is often infeasible due to the complexity and variability of how information is presented across different articles. For example, consider the following two facts within a merged group: *"23.1% of parents in the U.S. cited special needs as a reason for homeschooling"* and *"15.6% of parents said that the child had a physical or mental problem."* While the first fact explicitly mentions the U.S., the second fact lacks a geographical reference, making it ambiguous.

To tackle this, we employ a multi-step approach to identify and fill missing entities. First, we use Named Entity Recognition (NER) to detect and classify entity types in each fact. Next, we compare the entity types across facts within each *merged fact set* to detect missing entities. If an entity is present in one fact but absent in others, our algorithm flags it as a potential missing entity. To fill those missing entities, we leverage an LLM to analyze the original article content and infer the missing entities. Finally, these *merged fact sets* undergo an iterative validation and refinement phase similar to the previous stage. The validation checks their correctness against the original article and provides feedback, while the refinement applies corresponding edits.

### D. Implementation Details

We employ Google Search for article retrieval, using SearchAPI[5] to collect organic Search Engine Results Page (SERP) and Serper API[6] to extract page text, in compliance with source site terms and copyright laws. For the LLM mentioned in both Sections IV and V, we use OpenAI's *gpt-4o-2024-08-06* model, which supports structured outputs[7] for consistent formatting. We define the output schema for each LLM task and apply techniques such as Chain-of-Thought (CoT), prompt chaining, and few-shot prompting. Detailed prompts are provided in the supplemental materials. For the clustering, we utilize OpenAI's *text-embedding-3-large* model for vector embeddings and the Gaussian Mixture Model (GMM) method, which supports soft clustering in an unsupervised manner [65]. To accommodate screen size constraints and mitigate information overload when visualizing clusters as discussed in Section V-A, we limit the maximum number of clusters to 10. For the NER task (Section IV-C3), we use SpaCy's transformer-based model, *en_core_web_trf*[8],

[5]https://www.searchapi.io
[6]https://serper.dev/
[7]https://openai.com/index/introducing-structured-outputs-in-the-api/
[8]https://spacy.io/models/en#en_core_web_trf

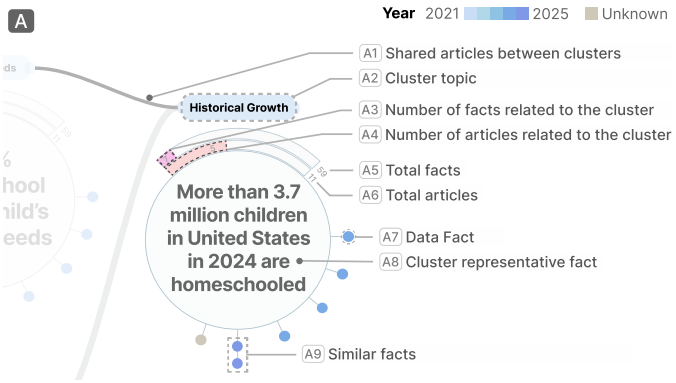



Figure 3. Thematic Circle represents one cluster of facts, with the cluster representative fact (A8) displayed in the inner circle, and individual data facts (A7) positioned in the bottom half of the outer area, where each circle is color-coded with publication year of its associated article.

due to its robust performance and comprehensive entity coverage. We have released *Compendia*'s source code here: https://github.com/compendia-project for public use.

## V. Visual Storytelling

Our visual storytelling module builds upon the above module and consists of a **Thematic Overview** and **Story View**, as shown in Fig. 2B. The *Thematic Overview* provides users with a brief visual summary of major topics related to their query, and the *Story View* enables users to explore detailed data facts in an interactive and engaging manner.

### A. Thematic Overview

The **Thematic Overview** (Fig. 1) groups extracted facts into semantically-coherent topics, allowing users to gain a quick summary of all the relevant topics for a given query (***R1***). To avoid overwhelming users, the *Thematic Overview* integrates a **Filter Widget** that allows the users to simplify the view and focus on what matters the most.

*1) Thematic Circle:* The **Thematic Circle** (Fig. 3) is designed to visualize details of a cluster, including a concise topic and related fact details, allowing users to gain a quick summary of a cluster. The concise *cluster topic*, displayed at the top (Fig. 3A2), indicates the main theme of the cluster. The circle (Fig. 3A8), fixed in size, encloses the text of the cluster's most representative fact. The dots (Fig. 3A7) positioned in the bottom half of the circle display individual data facts in the cluster. Each dot is color-coded by the publication year of its source article (***R5***), allowing users to quickly assess the timeliness of the information. As shown in Fig. 3A9, connected dots indicate similar facts identified through the merging process described in Section IV-C2. The two radial bars indicate the number of facts (Fig. 3A3) and the number of contributing articles (Fig. 3A4) related to the cluster, enabling easier comparison between clusters (***R1***).

The design balances overview and detail to satisfy ***R1***. While displaying the details of all the facts could be overwhelming [66], showing only a brief topic overview, such as a word cloud of keywords, is often ineffective for interpretation [67]. Our design adopts a focus+context strategy [68] that balances overview and detail. Also, inspired by the *Circle Map* used in structured brainstorming [69]: the inner circle conveys the main theme, while the outer ring presents supporting details.

A link (Fig. 3A1) between two clusters indicates the number of shared articles between them (***R2***), where those articles contain information relevant to multiple topics. The thickness of each link reflects the number of shared articles.

*2) Filter Widget:* As shown in Fig.1A, **Filter Widget** helps to manage visual complexity and support diverse exploration needs. The filters include: (1) **Relevance-based topic filtering**, a single-thumb slider, which limits the number of displayed *Thematic Circles* (six by default) based on their relevance to the query (Section IV-C1); (2) **Shared article connection filtering**, a range slider, which adjusts the visibility of inter-topic links based on the number of shared articles to reduce clutter and support targeted exploration; and (3) **Fact visibility toggling**, which lets users show or hide fact dots to maintain a cleaner view. Based on the applied filters, the summary panel (Fig. 1G) updates to summarize both the statistics of the current story and the overall distribution of articles and facts, providing a concise snapshot of the information landscape.

*3) Interactions: Compendia* enables rich interactions to allow users to smoothly explore the overview of the story. Users can hover over a data fact dot (Fig. 3A7) to reveal the detailed fact panel (Fig. 1B) displaying the detailed fact, its source, and publication year (***R5***). Furthermore, each favicon (Fig. 4A3), which indicates the source article, is clickable and redirects users to the original article (***R5***). Hovering over the related articles bar (Fig. 3A4) reveals the list of the articles contributing to the cluster as shown in Fig. 1D. Similarly, hovering over the radial bar for related facts (Fig. 3A3) displays a panel showing the count of related facts (Fig. 1E). Hovering over a link (Fig. 3A1) reveals the shared articles panel (Fig. 1F), which lists the information about the shared articles (***R2***), including their titles, publication years, and sources. Users can also toggle the articles panel (Fig. 1C) by clicking on the bottom-right arrow button as shown in Fig. 1. It shows all the articles contributing to the story, including their titles, snippets, sources, and published years (***R5***).

### B. Story View

The **Story View** (Fig. 4B) allows users to move from the overview to detailed exploration of the data facts of a topic, helping them understand the quantitative evidence. This consists of a **Narrative Unit** (Fig. 4A), which provides a detailed story with a chart, and a zoomed *Thematic Circle* (Fig. 4B), which presents the top three most relevant data facts as text within circles (Fig. 4B4), rather than showing only the most representative fact.

*1) Narrative Unit:* The **Narrative Unit**, shown in Fig.4A, summarizes a *merged fact set* derived in Section IV-C2, combining a text description (Fig.4A2) with a chart (Fig.4A4) in detail (***R4***). The **narrative title**, shown in the top right of the *Narrative Unit* (Fig. 4A1), conveys the core message of the *merged fact set*, generated by prompting an LLM to give a concise title of the *merged fact set*.

The **narrative caption**, a text with highlights, shown in Fig. 4A2, expresses the facts, in a *merged fact set*, in natural language to form a meaningful story. The caption highlights

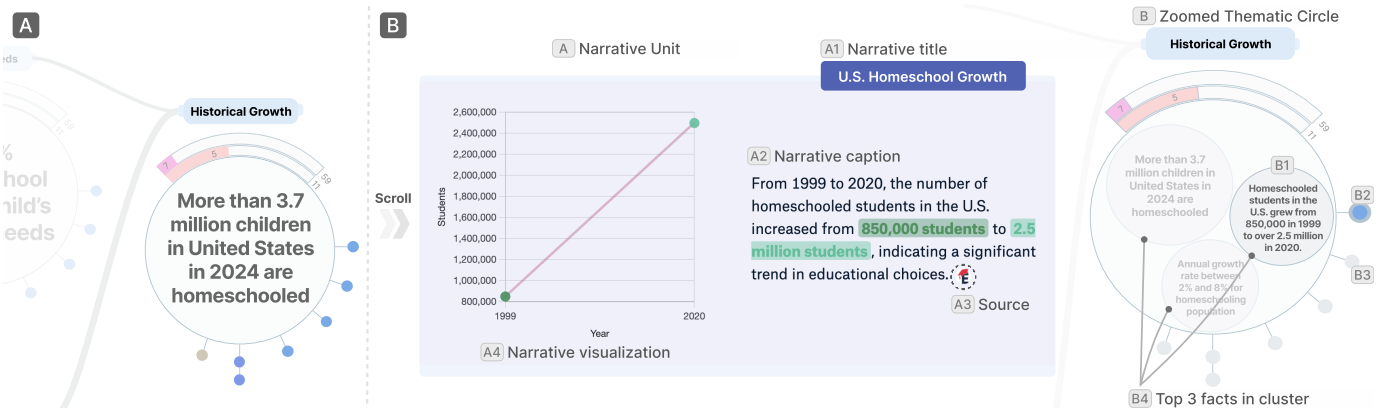



Figure 4. Scrolly flow of *Compendia*. The system transitions from **Thematic Overview** (A) to **Story View** (B) when scrolling down. (B) illustrates the narrative related to U.S. homeschool growth, where the corresponding fact is highlighted in the outer area of the zoomed thematic circle. Since this narrative is derived from one of the top 3 facts in the cluster, the related fact is also highlighted within the zoomed thematic circle.

key numerical values, applying consistent colors to match the data values across both the caption and the visualization (***R4***). The favicons at the end of the text (Fig. 4 A3) indicate the source articles for the narrative (***R5***). This *narrative caption* is generated during the **Narrative Generation** phase of our framework (Fig. 2 B), as HTML code, using an LLM to highlight and summarize all the facts within a *merged fact set* and produce a concise, human-readable story.

The **narrative visualization**, shown in Fig. 4 A4, illustrates the underlying data of a *merged fact set* (***R4***). It can take the form of bar, pie, line, isotype, range/dumbbell, or text, which are the commonly-used charts in news and data-driven articles, as also noted in prior research [11], [12]. The **Visualization Preparation** phase of our framework (Fig. 2 B) recommends the chart type using an LLM, informed by prior work [70] that uses LLMs to recommend appropriate visualizations.

*2) Interactions: Compendia* provides a scroll-driven narrative interface that implements scrollytelling techniques [71] to support both linear and non-linear exploration (***R3***).

To enable a smooth story flow among different *Narrative Units*, we arrange them in a coherent sequence at two levels (**Narrative Organization** in Fig. 2 B): across clusters and within each cluster, as each cluster contains one or more *Narrative Units*. First, the inter-cluster ordering follows the inverted pyramid paradigm for engaging news reporting [72] by placing the most relevant clusters first based on the clusters' relevance scores (Section IV-C1). Second, the intra-cluster ordering uses an LLM to arrange *Narrative Units* within a cluster, producing a coherent sequence with smooth transitions between *Narrative Units*. This layered approach transforms all *Narrative Units* into a coherent data story.

**Linear exploration.** As users scroll down the page, *Compendia* progressively unfolds the story following the above-mentioned order. When users begin scrolling on the *Thematic Overview*, *Compendia* smoothly transitions to the first *Story View* (Fig. 4 B) using an animated pan and zoom to focus on the first *Thematic Circle*, while keeping other circles subtly visible in the background, following the *"Pan-And-Zoom"* scrollytelling technique [71] (***R3***). The zoomed *Thematic Circle* highlights facts associated with the current *Narrative Unit* (Fig. 4 B1, B2), while dimming others (Fig. 4 B3) (***R5***). As users scroll to the next *Narrative Unit* within the same cluster, the zoomed *Thematic Circle* animates to highlight the facts related to the new *Narrative Unit*, smoothly transitioning between visual states (***R3***). For example, when scrolling, we animate the *Thematic Circle* from Fig. 5 B to Fig. 5 C, updating the highlights such as most representative facts (Fig. 5 B1) and fact dots (Fig. 5 B2). When users scroll past the last narrative of the current cluster to a new cluster, the system pans and zooms to its *Thematic Circle* and unfolds the *Narrative Units*.

**Non-linear exploration.** In addition to scroll-driven progression, users can jump to any cluster by clicking its *Thematic Circle* (Fig. 3). The system navigates to that cluster's first *Narrative Unit*, from which users continue scrolling to reveal the remaining *Narrative Units* as in linear exploration. Users can return to the *Thematic Overview* by clicking the outside (Fig. 5 C1) of the *Thematic Circle*.

## VI. Use Case

In this section, we present a use case scenario where Bob, a university student and avid news reader, is analyzing TikTok's global trends for a media studies assignment. Rather than looking for subjective takes, Bob aims to uncover objective, quantitative facts about the TikTok's worldwide impact.

Bob discovers our system, *Compendia*, and enters the query *"TikTok trends worldwide"*. Instead of returning a flat list of articles, the system expands the query with variations, retrieves relevant articles, and extracts quantitative facts. Within moments, *Compendia* transforms his query into a rich, interactive visual overview (Fig. 5 A). From 12 retrieved articles, the system identifies 106 distinct facts and organizes them into thematic clusters such as Global Reach, Gen Z Focus, App Dominance and others. Initially, it presents 64 facts drawn from the six most relevant clusters, sourced from 9 of the 12 articles as displayed in summary panel (Fig.5 A5), giving Bob a first glance of the most relevant themes. Yet, he wants to dig deeper into the evidence.

Eager to explore the story behind the clusters, he scrolls down, and the system smoothly pans and zooms into the *Gen Z Focus* cluster from the overview. The story begins with *TikTok User Demographics* (Fig.5 B), where Bob learns that


Figure 5. *Compendia* generated story for a query *"TikTok trends worldwide"* showcasing: (A) the Thematic Overview; (B–C) the continuous scroll-down flow from (A) to (C) to explore the story; and (D) jumping to the story in the **Youth Appeal** theme by clicking the thematic circle (A6) in (A).

70% of TikTok's users are from Generation Z. The highlighted circles (Fig.5 B1) show that this story draws on three key facts, while the favicons (Fig. 5 B2) and highlighted dots (Fig. 5 B3) show that the story draws on two articles and three data facts. Wanting to check the credibility of the data, Bob hovers over the fact dot (Fig. 5 B3) to see the concise fact details as in the Fig. 1 B panel, and clicks on the favicon to trace it back to the original sources, confirming their credibility.

As he scrolls further, new narratives emerge, for example, on *Platform Preference by Gender* (Fig. 5 C). In the background, the *Thematic Circle* dynamically updates, highlighting the facts relevant to each unfolding story. This scrollytelling flow helps Bob stay connected to the overall story while drawing him deeper into the evidence like user growth, download trends, revenue growth and many more fulfilling his goal of finding quantitative facts. After following the initial clusters, Bob grows curious about what else he might uncover. With a simple click outside the thematic circle (Fig.5 C1), he moves back into the overview (Fig.5 A). Using the filter controls (Fig.5 A1), he expands the number of visible topics and pays his attention to the themes with more facts and broader article coverage, easily identified by the radial bars (Fig. 5 A3).

His curiosity leads him to the *Youth Appeal* theme. Clicking on the *Thematic Circle* (Fig. 5 A6), he jumps into a related narrative, where he discovers that 30% of TikTok users express interest in women's sports (Fig. 5 D). With each new discovery, Bob becomes increasingly impressed by *Compendia*'s ability to distill scattered news coverage into clear, engaging, and objective visual narratives.

## VII. Evaluation

We extensively evaluated *Compendia* through both quantitative evaluation and two user studies. The quantitative evaluation measured the accuracy of results at different stages of our framework. The two user studies provided a comprehensive evaluation of *Compendia*, with the first study assessing usability and effectiveness, and the second study evaluating ecological validity and real-world utility. More details related to the studies are provided in the supplementary materials.

### A. Quantitative Evaluation

***Procedure:*** We generated four data stories for topics in distinct domains (e.g., *"Tiktok trends worldwide"*) using *Compendia*. In total, 399 extracted data facts from 36 articles, producing 196 *merged fact sets* along with their associated *narrative captions* and *visualization recommendations*. Two co-authors independently reviewed and coded each story without prior discussions to maintain objectivity. They examined five key aspects: the accuracy of *fact content* and *data points* from the *Data Fact Extraction* stage; the quality of *merged fact sets* from the *Fact Organization* stage; and the quality of *narrative captions* and *visualization recommendations* from the *Visual Storytelling* module. To evaluate the accuracy of *fact content* and *data points*, the coders manually compared each extracted fact with its source article. A *fact content* was marked as correct if its meaning and context were faithfully retained, and a *data point* was marked correct if it exactly matched the value in the source. The quality of *narrative captions* was marked as correct if they faithfully reflected the underlying data facts. A *merged fact set* was marked as correct if the underlying data facts logically fit together. A *visualization recommendation* was marked as correct if it effectively represented the extracted data; otherwise, coders provided their own recommendation.

***Results:*** We adopted accuracy (i.e., the proportion of correct items in all extracted items) as the primary evaluation metric. *Compendia* achieved 97.2% accuracy in both *fact content* and *data points*, 95.9% accuracy in *merged fact sets*, 92.3% accuracy in *narrative caption*, and 77% accuracy in *visualization recommendations*. These results demonstrate

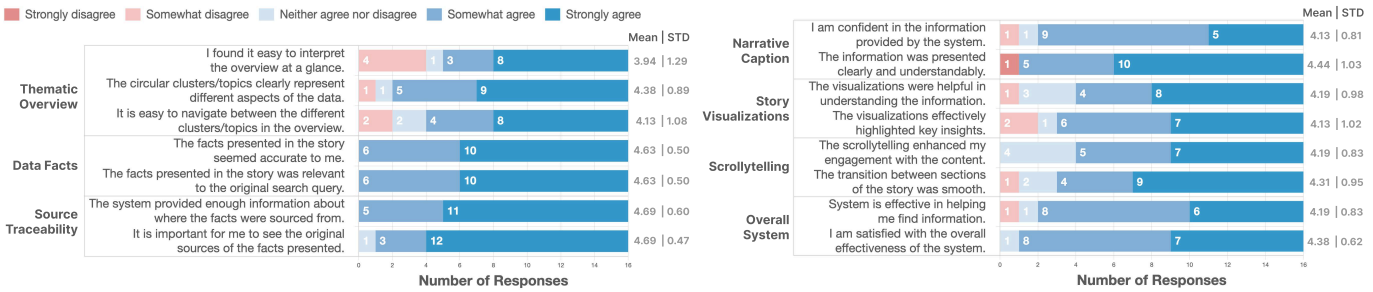



Figure 6. The effectiveness of *Compendia*. Responses cover aspects of the Thematic Overview, including interpretability, theme clarity, and navigation; Data Facts with their perceived accuracy and relevance; Source Traceability for factual grounding; Story View in terms of Narrative Caption clarity and confidence, usefulness of Story Visualizations, Scrollytelling engagement and transitions, and Overall System effectiveness and satisfaction.

*Compendia*'s effectiveness in accurately extracting structured data from unstructured text and the ability to connect fragmented facts into coherent stories leveraging the LLMs.

The failure cases in *Data Fact Extraction* stage arose from the difficulties in interpreting comparative expressions (e.g., *"A is 30% more than B"*) to derive the value of one item from the other. The *Fact Organization* failures arose primarily in cases requiring complex reasoning, such as when some facts shared the same context (e.g., year, region) while some of them included an additional qualifier (e.g., *"among men"*), leading to incorrect merges. The visualization recommendation failures were mostly with *Narrative Units* that contained a single data point. For example, a text visualization was recommended for *"About 57% of TikTok users discover brands through social media"*, possibly because of the approximate value (*"about 57%"*) was treated as plain text, whereas coders expected an isotype chart to better highlight the proportion. We anticipate the presence of such failure cases to decrease as LLMs become better at reasoning about data context.

### B. User Studies

We conducted two user studies with the same participants under controlled and non-controlled conditions. The **Initial Perception Study** adopted a controlled design with a pre-generated data story using *Compendia* for all participants, ensuring consistent evaluation of usability and effectiveness, and enabling a fair comparison with the AI-powered search engine Perplexity. Informed by the findings from the initial study, we updated *Compendia* by adding interactive filters (Fig. 1 A) prior to conducting the second study. The **Open-Ended Query Exploration Study** adopted a non-controlled design, allowing participants to pursue self-directed queries to assess ecological validity and real-world utility. All the studies were approved by our university's Institutional Review Board.

*1) **Initial Perception Study**:* This study evaluates the usability and effectiveness of the initial system and compares it with the baseline AI-powered search engine Perplexity.

***Participants**:* We recruited 16 participants (P1–P16) from two universities (8 females, 8 males, $\mu_{\text{age}}$= 25.75 years, $\sigma_{\text{age}}$= 3.75). All participants reported daily use of web search and AI-based search tools, primarily ChatGPT and Perplexity. Additionally, 13 participants were familiar with scrollytelling.

***Procedure**:* The studies were conducted through video calls where participants accessed an online version of *Compendia*, while sharing their screens. We first collected participants' demographic information and then introduced them to the system through a data story (*"Does AI lead to new jobs or job displacements?"*), which they explored to become familiar with its features. Then, they were tasked to examine a data story (*"Is homeschooling preferred by people?"*) using a think-aloud protocol, freely navigating while ensuring full coverage. Next, they used Perplexity to search the same query and reviewed its results. Finally, participants completed a post-study questionnaire (5-point Likert) on *Compendia*'s effectiveness (Fig. 6), rated usability with the System Usability Score (SUS) [73], compared *Compendia* with Perplexity (Fig. 7), and joined semi-structured interviews. The whole study lasted about 50 minutes.

***Results**:* Fig. 6 summarizes the questionnaire responses. As it shows, participants rated *Compendia* favorably in terms of its visualization design (Thematic Overview and Story Visualization), the quality of the data facts (Data Facts and Source Traceability), and the clarity and coherence of the stories (Narrative Caption and Scrollytelling). They also found the overall system effective and helpful. *Compendia* received a SUS score of 73.3, which falls within the good range [74].

*Thematic Overview.* Participants found the thematic overview was easy to interpret ($\mu = 3.94$), effective for representing different aspects of the information ($\mu = 4.38$), and smooth for navigating between clusters ($\mu = 4.13$). These ratings support our overview first approach. P3 valued *"seeing the big picture"* beyond a simple Google search, and P10 appreciated the links and timeline cues. We further observed that most participants used hovering features to see more details of the facts and the articles. However, while participants appreciated the information richness in the *Thematic Overview*, others, especially P2 and P11, found the overview somewhat complex and suggested adding more controls, such as toggling connections on or off (P2). This led us to integrate filter widget into the updated system.

*Facts and Source Traceability.* Participants valued the concise, accurate facts ($\mu = 4.63$) and strong source traceability ($\mu = 4.69$), with links enabling verification ($\mu = 4.69$). P11 appreciated that *"the facts are present in short and crisp format."* We observed that the participants used source links not only for verification but also to explore broader context, enhancing the trust and engagement.

*Story View.* Participants found the story-like format effective for organizing information and navigating between narratives, with smooth animated transitions ($\mu = 4.31$) and engaging scrollytelling ($\mu = 4.19$). Most did not feel lost, though some

suggested adding a minimap-style navigator, similar to those in IDEs, to show their current position in the story. Narrative captions were rated clear ($\mu = 4.44$) and coherent ($\mu = 4.13$), although a few (e.g., P15) wanted more background details. Visualizations were valued for aiding comprehension ($\mu = 4.19$) and highlighting key insights ($\mu = 4.13$). Meanwhile, some participants noted failure cases, including duplicated narratives caused by hallucination, incoherent narratives (e.g., where different facts described the same issue but reported conflicting numerical values were merged into a single narrative), and instances where the data visualizations did not fit the content. During the interviews, P3 suggested incorporating images and using charts from the original articles, while P5 recommended retrieving more data to make the visualizations richer.

*Overall System.* Overall, *Compendia* was seen as effective for gaining a comprehensive view ($\mu = 4.19$) with high satisfaction ($\mu = 4.38$). Most participants valued its ability to condense scattered quantitative information into clear narratives and indicated they would use it for research (P8), academic work (P12), and sharing stories (P13).

**Comparison.** Participants compared *Compendia* with Perplexity in terms of comprehensiveness, informativeness, clarity, and engagement as shown in Fig. 7. The interviews revealed two clear strengths. First, participants valued the integrated overview of the topics in *Compendia*, which let them see the full topical landscape before drilling into specifics. As P3 explained, *"Compendia shows me the whole picture, so I know where to start."* Second, they valued the system's storytelling approach, where we combine data visualizations with the text, as the visuals helped them see trends and make comparisons that pure text could not. P10 remarked, *"Perplexity is more like an essay answer but this combines visuals with text."* However, not all the participants preferred this style. For example, P15, who mentioned a preference for reading text, leaned towards text-only summaries. Furthermore, P2 suggested adding a brief summary paragraph, similar to the opening paragraph in Perplexity's results.

*2) **Open-Ended Query Exploration Study**:* This study examines the real-world use of *Compendia* by letting the participants try their own open-ended queries.

***Participants**:* We recruited the same 16 participants (P1–P16) involved in the initial perception study to join this study.

***Procedure**:* The studies were conducted through video calls where participants accessed an online version of *Compendia*, while sharing their screens. Participants were asked to formulate a search query of their interest and searched in *Compendia*. Once the story was generated, they explored it without any guidance or restrictions. We observed participants' interactions throughout the study. Finally, participants completed a post-study questionnaire and semi-structured interviews. The sessions lasted about 30 minutes. They received $25 for their time participating in both user studies.

***Results**:* We summarize the study results in terms of the generalizability of *Compendia*, information foraging and sensemaking strategies, and improvement suggestions.

*Generalizability across domains and task types.* All participants successfully performed a self-directed query using *Compendia*. The queries spanned multiple domains, including real estate (e.g., *"What are the property prices for Sydney 2025?"* – P10), demographics (e.g., *"Marriage statistics in Singapore"* – P4), healthcare (e.g., *"The distribution of AIDS patients in Sweden"* – P12), finance (e.g., *"What are the market shares of global pharmaceutical companies?"* – P3), and architecture (e.g., *"Typical Gothic architectural style in Notre Dame de Paris?"* – P11). These queries also reflected multiple information-seeking task types, such as comparative analysis (e.g., *"BYD car sales compared to Tesla car sales in the recent years"* – P1), trend or temporal analysis (e.g., *"AI usage trends"* – P9), forecasting (e.g., *"Gold prices predictions by different firms"* – P6), and descriptive statistical overview (e.g., *"World population by country"* – P15). Together, these findings suggest broader applicability of *Compendia* across different domains and task types in real-world scenarios.

*Information foraging and sensemaking strategies.* All the participants agreed that the resulting information generally matched their expectations. For example, P7 mentioned *"One of my friends shared this [resulted story] Harvard funding freeze with me a few weeks ago."* We observed three recurring information foraging strategies that supported effective sensemaking:

***Overview-first exploration.*** Most of the participants began by looking at the *Thematic Overview* to assess topic coverage before proceeding to detailed stories; as P9 explained, *"The circular layout [thematic overview] helped me pick up on some high-level AI trends I didn't thought first."* Then, they followed the story in the given sequence by scrolling.

***Back-and-forth navigation and iterative sensemaking.*** Rather than examining the entire overview at once, some participants (P4 and P14) moved iteratively between the *Thematic Overview* and the detailed *Narrative Units* to refine their understanding as new information emerged. For instance, P4 initially navigated to detailed narrative related to the Divorce Trends, which appeared top left of the overview, and then returned to explore other connected topics.

***Filtering for sensemaking.*** All participants agreed that the filters were easy to understand and useful for information foraging. Beyond surface-level navigation, several participants used filtering as an active sensemaking strategy to identify salient patterns and relationships. For example, P3, who explored market shares of global pharmaceutical companies, first filtered by the most relevant topics, which helped them quickly surface dominant players, noting that *"Pfizer is dominating this."* Similarly, P4 used the shared-articles filter to examine cross-article connections, which prompted reflection on broader patterns, observing that *"These links suggest educated people tend to have fewer children."*

*Improvement suggestions.* During the study, some participants encountered minor issues that highlight opportunities for further refinement of the system in future work.

***Handling outdated and time-sensitive information.*** Some participants observed outdated results for future-oriented predictions. For example, P6 searched for gold price predictions and expected forecasts for the coming years. While the results included such forecasts, they also surfaced predictions from older articles that were no longer valid at the time of exploration (e.g., a 2023 article forecasting prices for 2024,

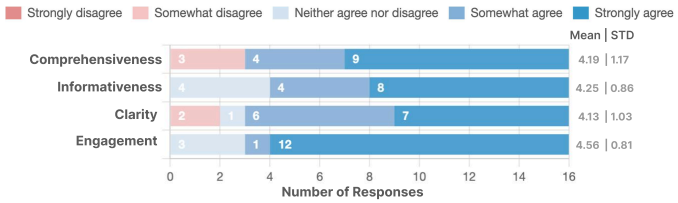



Figure 7. The comparison between *Compendia* and Perplexity.

which was outdated in 2025). This limitation primarily occurs in cases involving time-sensitive predictive information, which the current system does not explicitly distinguish. As future work, this issue could be addressed by incorporating explicit temporal reasoning into the framework (in *Data Validation*).

***Redundancy and narrative coherence.*** Some participants noticed somewhat similar content appearing across different clusters (P12), as well as occasional disjointed story flow (P8). These issues primarily stemmed from phrasing variations across sources that express equivalent facts differently, causing them to be clustered into separate groups, as well as the high diversity of extracted facts, which can disrupt smooth narrative transitions. Although relatively infrequent, these cases highlight opportunities for future work to improve clustering and narrative organization and are likely to diminish as LLM context windows continue to expand.

## VIII. Discussion

This section discusses key lessons learned from generating data stories using unstructured text and outlines limitations and future work.

### A. Lessons Learned

**LLMs for handling unstructured text:** Systematically transforming unstructured information into structured narratives requires effective information retrieval and extraction. Transformer-based retrieval models (e.g., BERT) often require domain-specific fine-tuning, limiting cross-domain applicability, while rule-based and statistical extraction methods (e.g., NER) struggle with linguistic ambiguity and stylistic or structural variation. In contrast, LLMs offer a more flexible alternative, as they can leverage contextual reasoning even in the absence of explicit task-specific training. However, LLMs are inherently unreliable, and their effectiveness in data storytelling relies on the integration with a broader, more robust processing framework. Several strategies in our processing framework were helpful for reliability. Text chunking and parallel processing enable handling articles that exceed an LLM's context window, by splitting text into manageable segments that are processed independently and later aggregated. Structured output and schema enforcement constrain model outputs to predefined formats, improving consistency and downstream usability. Validation and retry logic ensure that extracted results satisfy schema and quality requirements. Finally, prompt engineering plays a central role, with clear instructions and few-shot examples guiding model behavior without requiring model fine-tuning.

**Challenges in LLM-based systems:** While LLMs offer flexibility in handling unstructured text, they also introduce practical challenges when it comes to larger context, specific tasks, and time efficiency. Scalability remains a primary concern, particularly in the later stages of the framework (e.g., *Fact Organization*), where all extracted facts must be considered at once and chunking is not feasible. For example, our experiments with LLM-based clustering yielded poor and unstable results (e.g., missing or hallucinated facts), due to both the larger context and the inherently unsupervised nature of the task. This prompted us to adopt a more robust machine learning approach for clustering. In practice, the system can effectively manage around 300 individual facts per story, constrained by output context length limits (especially in merging and validation) that may result in incomplete output structures. We believe these constraints will eventually be resolved with the rapid context length expansion of modern LLMs [75]. Another drawback of using LLMs is the latency. This limitation is not unique to our system; most advanced research agents that autonomously search, analyze, and synthesize information from the Internet, such as DeepResearch[9], also require longer processing time than regular LLM-based chat. As LLM architectures continue to evolve, we expect the future models to achieve lower latency with high accuracy.

**Support personalized data stories:** Throughout the user studies, we observed that participants employed distinct sensemaking strategies and expressed different information needs when interpreting the presented content. For example, P3 began by filtering for only the most relevant information, whereas P15 preferred richer contextual details to better understand extracted facts. These variations highlight the inherently personal nature of sensemaking and suggest future research opportunities in adaptively personalizing data stories based on factors such as users' interaction histories (e.g., previously applied filters), information granularity (e.g., preference for detailed contextual information), task framing (e.g., comparative analysis), and source preferences.

**Storytelling for search interfaces:** We believe this work opens new directions for search interfaces, beyond presenting ranked lists such as map-based visualizations [76], [77], where narrative storytelling can enhance how users engage with and interpret search results. In our study, participants generally preferred *Compendia* over traditional text-based interfaces like Perplexity, citing its structured presentation and visual storytelling. This aligns with emerging trends in search technologies, such as Bing's Copilot Search[10], which are beginning to integrate AI-generated summaries, images, and infographics to give visually immersive search experiences. A promising avenue for future work is to expand beyond factual content toward nuanced perspectives such as opinions, interpretations, and contested views while preserving clarity and trust.

### B. Limitations and Future Work

**Fact-checking and bias awareness:** *Compendia* provides source links alongside the extracted facts and generated narratives to support transparency and verification. However, the system does not independently verify extracted facts or assess

[9]https://openai.com/index/introducing-deep-research/
[10]https://blogs.bing.com/search/2021_03

framing and bias, which are beyond the scope of this work. As a result, inaccurate, outdated, or selectively reported statistics in the source articles may still propagate into the generated story, potentially undermining the result quality. Automated fact-checking and bias assessment remain active research areas [78], [79]. Future work could integrate retrieval-based verification, knowledge-graph validation, or claim–evidence reasoning, and employ visual cues to explicitly communicate uncertainty or confidence levels.

**Objectivity–context trade-off:** To preserve objective grounding, we exclude subjective opinions and interpretations, keeping only minimal context for each fact (e.g., its source sentence). While this reduces the risk of introducing speculative explanations and makes facts easier to grasp, it might omit qualitative context that helps readers interpret *"why"* quantitative changes occur (e.g., policy shifts). As a result, facts might appear isolated and connections across narratives could feel less coherent in some cases, where broader contexts help weave a more coherent overall narrative. Future work could add richer context and carefully scoped interpretive statements to better connect facts, but this may also lead to longer stories and introduce subjectivity.

**Story customization:** *Compendia* currently supports only coarse–grained user control through the provided filters, such as the number of clusters. Beyond these controls, the system does not yet support deeper narrative shaping, such as aggregating additional facts or expanding contextual information, to match different sensemaking intents. Future work could move toward more dynamic and user-steerable storytelling approaches, for example, systems like Socrates [80] that adapt data stories based on user feedback, enabling stories to be iteratively refined, such as by adjusting the amount of contextual scaffolding or narrative detail.

**Sparse, missing, and heterogeneous data:** When only a few quantitative data points are available, narrative visualizations might become less expressive and inadvertently mislead interpretation (e.g., missing intermediate values masking volatility). Moreover, the extracted quantities may be heterogeneous across sources (e.g., following different population definitions), which could reduce comparability even if each value is accurate. Future work could improve merging and normalization rules (e.g., aligning definitions where possible), explicitly surface missing data through visual cues, or explore approaches that represent absent values and uncertainty (e.g., as in ChartifyText [81]). However, richer alignment and uncertainty handling may increase computational overhead and generation latency.

## IX. Conclusion

We present *Compendia*, an automated system that retrieves and analyzes a collection of online articles in response to a user query and delivers a coherent and meaningful visual data story tailored to the user's informational needs through scrollytelling. Our framework comprises two main modules: the Data Fact Extraction and Organization module and the Visual Storytelling module, each designed to address the unique challenges of working with unstructured textual data. We evaluate *Compendia* through a quantitative analysis and two user studies, demonstrating that *Compendia* effectively extracts accurate facts, organizes them into meaningful narratives, and presents them in a visually compelling and user-friendly manner. In the future, we plan to enhance system usability by reducing latency through more efficient, high-performing LLMs and by integrating fact-checking techniques to increase the trustworthiness. We also believe this work opens new directions for search interfaces, where narrative storytelling can transform user engagement with search results by incorporating not only objective facts but also subjective opinions.

## Acknowledgments

This project is supported by NTU Start Up Grant awarded to Yong Wang.

## References

[1] D. O. Case and L. M. Given, "Looking for information: A survey of research on information seeking, needs, and behavior," 2016.

[2] K. Purcell, L. Rainie, A. Mitchell, T. Rosenstiel, and K. Olmstead, "Understanding the participatory news consumer," *Pew Internet and American Life Project*, vol. 1, pp. 19–21, 2010.

[3] F. Stalph, N. Thurman, and S. Thäsler-Kordonouri, "Exploring audience perceptions of, and preferences for, data-driven 'quantitative'journalism," *Journalism*, vol. 25, no. 7, pp. 1460–1480, 2024.

[4] Y. Fu and J. Stasko, "More than data stories: Broadening the role of visualization in contemporary journalism," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 8, pp. 5240–5259, 2024.

[5] G. Marchionini, "Exploratory search: from finding to understanding," *Commun. ACM*, vol. 49, no. 4, p. 41–46, 2006.

[6] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman, "Tracking and summarizing news on a daily basis with columbia's newsblaster," in *Proceedings of the human language technology conference*. Morgan Kaufmann San Francisco, 2002, pp. 280–285.

[7] F. Hamborg, N. Meuschke, and B. Gipp, "Matrix-based news aggregation: Exploring different news perspectives," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2017, pp. 1–10.

[8] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond memorability: Visualization recognition and recall," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 519–528, 2015.

[9] G. Dove and S. Jones, "Narrative visualization: Sharing insights into complex data," 2012.

[10] E. Segel and J. Heer, "Narrative visualization: Telling stories with data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1139–1148, 2010.

[11] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma, and D. Zhang, "Datashot: Automatic generation of fact sheets from tabular data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 895–905, 2019.

[12] D. Shi, X. Xu, F. Sun, Y. Shi, and N. Cao, "Calliope: Automatic visual data story generation from a spreadsheet," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 453–463, 2020.

[13] M. S. Islam, M. T. R. Laskar, M. R. Parvez, E. Hoque, and S. Joty, "DataNarrative: Automated data-driven storytelling with visualizations and texts," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2024, pp. 19 253–19 286.

[14] J. Hullman, N. Diakopoulos, and E. Adar, "Contextifier: automatic generation of annotated stock visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. Association for Computing Machinery, 2013, p. 2707–2716.

[15] M. Leake, H. V. Shin, J. O. Kim, and M. Agrawala, "Generating audio-visual slideshows from text articles using word concreteness," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–11.

[16] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of The Craft of Information Visualization*. Elsevier, 2003, pp. 364–371.

[17] A. Tjärnhage, U. Söderström, O. Norberg, M. Andersson, and T. Mejtoft, "The impact of scrollytelling on the reading experience of long-form journalism," in *Proceedings of the European Conference on Cognitive Ergonomics 2023*, ser. ECCE '23. Association for Computing Machinery, 2023.

[18] D. Seyser and M. Zeiller, "Scrollytelling–an analysis of visual storytelling in online journalism," in *Proceedings of 2018 22nd International Conference Information Visualisation (IV)*. IEEE, 2018, pp. 401–406.

[19] B. Lee, N. H. Riche, P. Isenberg, and S. Carpendale, "More than telling a story: Transforming data into visually shared stories," *IEEE Computer Graphics and Applications*, vol. 35, no. 5, pp. 84–90, 2015.

[20] H. Shao, R. Martinez-Maldonado, V. Echeverria, L. Yan, and D. Gasevic, "Data storytelling in data visualisation: Does it enhance the efficiency and effectiveness of information retrieval and insights comprehension?" in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–21.

[21] C. Tong, R. Roberts, R. Borgo, S. Walton, R. S. Laramee, K. Wegba, A. Lu, Y. Wang, H. Qu, Q. Luo *et al.*, "Storytelling and visualization: An extended survey," *Information*, vol. 9, no. 3, p. 65, 2018.

[22] Q. Chen, S. Cao, J. Wang, and N. Cao, "How does automation shape the process of narrative visualization: A survey of tools," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[23] H. Li, Y. Wang, and H. Qu, "Where are we so far? understanding data storytelling tools from the perspective of human-ai collaboration," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–19.

[24] Y. He, K. Xu, S. Cao, Y. Shi, Q. Chen, and N. Cao, "Leveraging foundation models for crafting narrative visualization: A survey," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–20, 2025.

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

[26] W. Cui, J. Wang, H. Huang, Y. Wang, C.-Y. Lin, H. Zhang, and D. Zhang, "A mixed-initiative approach to reusing infographic charts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 173–183, 2021.

[27] A. Offenwanger, M. Brehmer, F. Chevalier, and T. Tsandilas, "Timesplines: Sketch-based authoring of flexible and idiosyncratic timelines," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[28] D. Kang, T. Ho, N. Marquardt, B. Mutlu, and A. Bianchi, "Toonnote: Improving communication in computational notebooks using interactive data comics," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[29] M. Conlen and J. Heer, "Idyll: A markup language for authoring and publishing interactive articles on the web," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018, pp. 977–989.

[30] N. Sultanum, F. Chevalier, Z. Bylinskii, and Z. Liu, "Leveraging text-chart links to support authoring of data-driven articles with vizflow," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–17.

[31] S. Latif, Z. Zhou, Y. Kim, F. Beck, and N. W. Kim, "Kori: Interactive synthesis of text and charts in data documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 184–194, 2021.

[32] A. Tyagi, J. Zhao, P. Patel, S. Khurana, and K. Mueller, "User-centric semi-automated infographics authoring and recommendation," *arXiv preprint arXiv:2108.11914*, 2021.

[33] H. Li, L. Ying, H. Zhang, Y. Wu, H. Qu, and Y. Wang, "Notable: On-the-fly assistant for data storytelling in computational notebooks," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–16.

[34] M. Sun, L. Cai, W. Cui, Y. Wu, Y. Shi, and N. Cao, "Erato: Cooperative data story editing via fact interpolation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 983–993, 2022.

[35] J. Lu, W. Chen, H. Ye, J. Wang, H. Mei, Y. Gu, Y. Wu, X. L. Zhang, and K.-L. Ma, "Automatic generation of unit visualization-based scrollytelling for impromptu data facts delivery," in *Proceedings of 2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. IEEE, 2021, pp. 21–30.

[36] N. Sultanum and A. Srinivasan, "Datatales: Investigating the use of large language models for authoring data-driven articles," in *Proceedings of 2023 IEEE Visualization and Visual Analytics (VIS)*. IEEE, 2023, pp. 231–235.

[37] L. Cheng, D. Deng, X. Xie, R. Qiu, M. Xu, and Y. Wu, "Snil: Generating sports news from insights with large language models," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[38] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[39] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[41] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[42] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[43] X. Li, J. Jin, Y. Zhou, Y. Zhang, P. Zhang, Y. Zhu, and Z. Dou, "From matching to generation: A survey on generative information retrieval," *ACM Trans. Inf. Syst.*, vol. 43, no. 3, May 2025.

[44] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[45] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[46] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[47] M. Dörk, N. H. Riche, G. Ramos, and S. Dumais, "Pivotpaths: Strolling through faceted information spaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2709–2718, 2012.

[48] R. W. White and R. A. Roth, *Exploratory search: Beyond the query-response paradigm*. Morgan & Claypool Publishers, 2009, no. 3.

[49] K. Hornbæk and M. Hertzum, "The notion of overview in information visualization," *International Journal of Human-Computer Studies*, vol. 69, no. 7-8, pp. 509–525, 2011.

[50] B. Kules and B. Shneiderman, "Users can change their web search tactics: Design guidelines for categorized overviews," *Information Processing & Management*, vol. 44, no. 2, pp. 463–484, 2008.

[51] M. A. Hearst and J. O. Pedersen, "Reexamining the cluster hypothesis: Scatter/gather on retrieval results," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 1996, pp. 76–84.

[52] O. Zamir and O. Etzioni, "Grouper: a dynamic clustering interface to web search results," *Computer networks*, vol. 31, no. 11-16, pp. 1361–1374, 1999.

[53] F. Das-Neves, E. A. Fox, and X. Yu, "Connecting topics in document collections with stepping stones and pathways," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005, pp. 91–98.

[54] B. Sun, P. Mitra, C. L. Giles, J. Yen, and H. Zha, "Topic segmentation with shared topic detection and alignment of multiple documents," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 199–206.

[55] "The Great Flood of 2019: A Complete Picture of a Slow-Motion Disaster (Published 2019) — nytimes.com," https://www.nytimes.com/interactive/2019/09/11/us/midwest-flooding.html, [Accessed 31-08-2025].

[56] "What city is the microbrew capital of the US? — pudding.cool," https://pudding.cool/2017/04/beer/, [Accessed 31-08-2025].

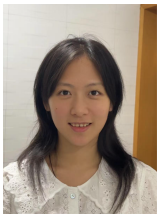
[57] J. Hullman and N. Diakopoulos, "Visualization rhetoric: Framing effects in narrative visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2231–2240, 2011.

[58] K. Solovev and N. Pröllochs, "References to unbiased sources increase the helpfulness of community fact-checks," *Scientific Reports*, vol. 15, no. 1, p. 25749, 2025.

[59] S. Diel and C. Roberts, "News story aggregation and perceived credibility," *Newspaper Research Journal*, vol. 42, no. 2, pp. 162–181, 2021.

[60] P. Borah, "The hyperlinked world: A look at how the interactions of news frames and hyperlinks influence news credibility and willingness to seek information," *Journal of Computer-Mediated Communication*, vol. 19, no. 3, pp. 576–590, 2014.
[61] M. A. Hearst, "Tilebars: Visualization of term distribution information in full text information access," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1995, pp. 59–66.
[62] J. Bevendorff, M. Wiegmann, M. Potthast, and B. Stein, "Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines," in *Proceedings of Advances in Information Retrieval: 46th European Conference on Information Retrieval*. Springer-Verlag, 2024, p. 56–71.
[63] R. Barzilay and L. Lee, "Catching the drift: Probabilistic content models, with applications to generation and summarization," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2004, pp. 113–120.
[64] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.
[65] U. Katz, M. Levy, and Y. Goldberg, "Knowledge navigator: LLM-guided browsing framework for exploratory search in scientific literature," in *Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024, pp. 8838–8855.
[66] M. J. Eppler and J. Mengis, "The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines," *The information society*, vol. 20, no. 5, pp. 325–344, 2004.
[67] J. Harris, "Word clouds considered harmful," *Nieman Journalism Lab*, vol. 13, p. 124, 2011.
[68] A. Cockburn, A. Karlson, and B. B. Bederson, "A review of overview+detail, zooming, and focus+context interfaces," *ACM Comput. Surv.*, vol. 41, no. 1, Jan. 2009.
[69] D. Hyerle, "Thinking maps: Visual tools for activating habits of mind," *Learning and leading with habits of mind*, vol. 16, pp. 149–174, 2008.
[70] L. Wang, S. Zhang, Y. Wang, E.-P. Lim, and Y. Wang, "LLM4Vis: Explainable visualization recommendation using ChatGPT," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, 2023, pp. 675–692.
[71] J. Oesch, A. Renner, and M. Roth, "Scrolling into the newsroom: A vocabulary for scrollytelling techniques in visual online articles," *Information Design Journal*, vol. 27, no. 1, pp. 102–114, 2022.
[72] H. Po¨ttker, "News and its communicative quality: the inverted pyramid—when and why did it appear?" *Journalism Studies*, vol. 4, no. 4, pp. 501–511, 2003.
[73] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability Evaluation in Industry*, vol. 189, no. 194, pp. 4–7, 1996.
[74] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," *Journal of Usability Studies*, vol. 4, no. 3, pp. 114–123, 2009.
[75] Y. Ding, L. L. Zhang, C. Zhang, Y. Xu, N. Shang, J. Xu, F. Yang, and M. Yang, "Longrope: Extending llm context window beyond 2 million tokens," *arXiv preprint arXiv:2402.13753*, 2024.
[76] J. Peltonen, K. Belorustceva, and T. Ruotsalo, "Topic-relevance map: Visualization for improving search result comprehension," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, ser. IUI '17. Association for Computing Machinery, 2017, p. 611–622.
[77] E. Clarkson, K. Desai, and J. Foley, "Resultmaps: Visualization for search interfaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1057–1064, 2009.
[78] T. Agunlejika, "Ai-driven fact-checking in journalism: Enhancing information veracity and combating misinformation: A systematic review," *SSRN*, 2025.
[79] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Transactions of the association for computational linguistics*, vol. 10, pp. 178–206, 2022.
[80] G. Wu, S. Guo, J. Hoffswell, G. Y.-Y. Chan, R. A. Rossi, and E. Koh, "Socrates: Data story generation via adaptive machine-guided elicitation of user feedback," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 1, p. 131–141, Jan. 2024.
[81] 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.

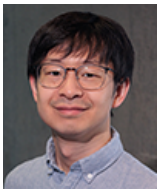
**Manusha Karunathilaka** is currently a Ph.D. candidate in School of Computing and Information Systems at Singapore Management University (SMU). His research interests include data visualization, storytelling, and human–computer interaction, with a particular focus on automated visual story generation. He received his bachelor's degree in Computer Science and Engineering from the University of Moratuwa, Sri Lanka. For more details, kindly visit https://manusha-karunathilaka.com.



**Litian Lei** is currently a master student in Artificial Intelligence at the College of Computing and Data Science, Nanyang Technological University. Her research interests include data visualization, human-computer interaction and computer vision. She received her bachelor's degree in Information and Communications Technology from Royal Institute of Technology in Sweden.

**Yiming Gao** is currently an undergraduate student at the College of Computing and Data Science, Nanyang Technological University. His research interests lie in the areas of Large Language Model (LLM) evaluation pipelines and reinforcement learning (RL).

**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.



**LI Jiannan** is currently an assistant professor in School of Computing and Information Systems at Singapore Management University. His research interests include Human-Computer Interaction and Human-Robot Interaction, with an emphasis on improving interactions between humans and intelligent agents. He obtained his Ph.D. in Computer Science from University of Toronto. He received his MSc in Computer Science from University of Calgary and BEng in Automation and Control from Southeast University, China. For more details, please refer to https://jchrisli.github.io/.