DataDive: Supporting Readers' Contextualization of Statistical Statements with Data Exploration

Hyunwoo Kim
khw0726@kaist.ac.kr
School of Computing, KAIST
Daejeon, Korea

Khanh Duy Le
duy.le1201@hcmut.edu.vn
Department of Computer Science and Engineering, HCMUT
Ho Chi Minh City, Vietnam

Gionnieve Lim
gionnievelim@gmail.com
School of Computing, KAIST
Daejeon, Korea
SUTD
Singapore

Dae Hyun Kim
dhkim16@kaist.ac.kr
Information & Electronics Research Institute, KAIST
Daejeon, Korea

Yoo Jin Hong
dbwk18@kaist.ac.kr
School of Computing, KAIST
Daejeon, Korea

Juho Kim
juhokim@kaist.ac.kr
School of Computing, KAIST
Daejeon, Korea

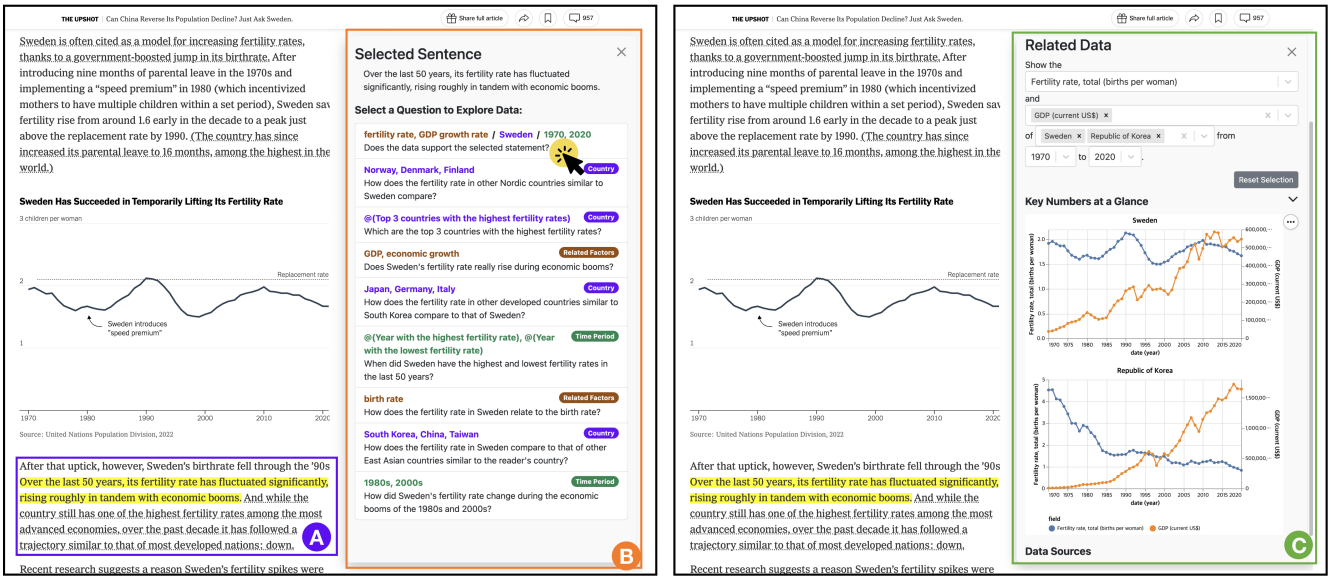

Figure 1: Overall user flow of *DataDive*. (A) While reading, the use select an statistical statement of interest to start exploring contextualization. (B) On the side, *DataDive* generates and presents a set of recommendations suggesting different potential contexts around the statistical statement. (C) Upon selecting a recommendation, *DataDive* matches the most relevant data from its database and visualizes the data. The user can use dropdown widgets to explore data in different contexts.

ABSTRACT

Statistical statements that refer to data to support narratives or claims are commonly used to inform readers about the magnitude of social issues. While contextualizing statistical statements with relevant data supports readers in building their own interpretation of statements, the complexity of finding contextual information on the web and linking statistical statements with it impedes readers' efforts to do so. We present *DataDive*, an interactive tool for contextualizing statistical statements for the readers of online texts. Based on users' selections of statistical statements, our tool uses an LLM-powered pipeline to generate candidates of relevant contexts and poses them as guiding questions to the user as potential contexts for exploration. When the user selects a question, *DataDive* employs visualizations to further help the user compare and explore contextually relevant data. A technical evaluation shows that *DataDive* generates important and diverse questions that facilitate exploration around statistical statements and retrieves relevant data for comparison. Moreover, a user study with 21 participants

IUI '24, March 18–21, 2024, Greenville, SC, USA

ACM ISBN 979-8-4007-0508-3/24/03
https://doi.org/10.1145/3640543.3645155

suggests that *DataDive* facilitates users to explore diverse contexts and to be more aware of how statistical data could relate to the text.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Visualization application domains**.

KEYWORDS

Contextualization, Reader support, Data visualization

ACM Reference Format:

Hyunwoo Kim, Khanh Duy Le, Gionnieve Lim, Dae Hyun Kim, Yoo Jin Hong, and Juho Kim. 2024. *DataDive*: Supporting Readers' Contextualization of Statistical Statements with Data Exploration. In *29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3640543.3645155

1 INTRODUCTION

Statistical information is commonly used to substantiate the magnitude of social issues. For example, data journalists produce news articles based on statistical data analyses, from economic indicators to scientific climate measurements [34]. During the COVID-19 pandemic, public health experts used statistical data to convey the severity of the pandemic, support the implementation of safety measures, and refute fake information [20, 67, 77]. Even laypeople produced and shared their own interpretations of statistical data in online forums to spark discussion on diverse issues [44]. Such texts often contain *statistical statements* (e.g. "Over the last 50 years, its [Sweden] fertility rate has fluctuated significantly, rising roughly in tandem with economic booms" (Figure 1), which describes the current state of the world with the support of statistical information.

The significance of data lies in its ability to derive meaning across various contexts, such as point of references and correlated factors. To facilitate this interpretative process, readers should be able to posit statistical statements within diverse contexts and formulate personalized interpretations of the data [11]. In this paper, we term this approach as "contextualization". Consider a reader reading a news article claiming "In South Korea, the fertility rate — the average number of children born to a woman in her reproductive years — is now 0.78" [1]. The reader could ask questions to interrogate this statement such as comparing the fertility rate to other developed countries or the reader's own country, or checking the temporal change in fertility rate to examine whether there exists a clear trend. Going deeper, they may formulate hypotheses to explain the phenomena, such as attributing the low fertility rate to a shortage of affordable housing, and explore the correlation between the fertility rate and housing prices over the past decade.

Existing research on social data analysis [29] and comments on data-driven articles [34, 44, 57] has shown that readers contribute to the contextualization of data in the public discourse. However, readers often struggle to produce their own contextualization because the statistical statements only include the minimally necessary data to support the authors' messages. Therefore, readers need to search for information to contextualize the statistical statements by themselves, which requires a complex cognitive process for the readers with barriers in each step [11, 82]. While existing work explored interactions for augmenting documents with data to provide greater contexts to statistical statements [24, 33, 55], their support is limited to providing data directly related to the text, thus confining the scope of context that readers can explore.

To support the exploration of relevant contexts beyond those directly related to the text, we propose *DataDive*, an interactive tool for contextualizing statistical statements by exploring relevant data while reading online text (Figure 1). With *DataDive*, the reader can explore various contextualizations by picking a statistical statement of interest in the text and exploring system-generated recommendations of potential contexts to explore. When the reader selects a potential context, *DataDive* provides an interactive visualization of selected contextualization of data, using external data from reliable sources. The reader can explore diverse data relevant to the selected contextualization by browsing the list of data series provided by *DataDive*. We built an LLM-based pipeline to parse statistical statements, which often contain vague and implicit reference to data, for generating recommendations for contextualization and retrieving the most relevant data points from datasets containing a number of data series.

To verify if the pipeline of *DataDive* can generate recommendations for potentially meaningful contexts and retrieve relevant datasets for a given statistical statement, we conducted a technical evaluation with 18 external evaluators on 77 statistical statements on various topics around global social issues and a dataset of global statistical indicators. Our technical evaluation results showed that 96.6% of pipeline-generated recommendations were considered meaningful for contextualization, and the pipeline could retrieve relevant data for 80.5% of the statements. To evaluate whether *DataDive* can support readers in contextualizing the statistical statements from online texts, we conducted a within-subjects study with 21 participants. In the study, participants read two news articles with statistical information on global social issues while using either only a web search engine or a web search engine with *DataDive* to explore contextual information of interest. Findings from observations and interviews showed how *DataDive* supported the participants in exploring more contexts around statistical statements and in better contextualizing the statistical statements in the text, while we did not discover statistically significant differences in participants' knowledge gain from their reading process. We further discuss potential improvements for *DataDive* to support more diverse types of contextualization and design lessons from using LLMs to produce factually reliable yet exploratory results.

This paper provides the following contributions:

- *DataDive*, a system for supporting contextualization of statistical statements while reading online texts.
- A technical pipeline for generating recommendations for contexts and retrieving datasets relevant to statistical statements on the web.
- Findings from a technical evaluation demonstrating the effectiveness of our pipeline in identifying relevant contexts, and from a user evaluation showing that *DataDive* can support readers' contextualization of data while reading online texts.

2 RELATED WORK

In this section, we review the importance of contextualization with data, previous interactive systems and designs for reading text, and existing work on using natural language for interacting with data.

2.1 Contextualization of Data

Data is just a number; to be meaningful, it needs to be interpreted with context [17]. Contextualization refers to interpreting data according to the characteristics and properties of the reader and their environment [86]. Contextualizations are particularly evident in data-driven narratives that aim to make sense and construct stories out of data [11, 21, 25]. When reading text, readers do not solely focus on understanding the text's meaning but also actively connect the content with their prior knowledge [16]. During the process, readers' individual differences, such as personal relationship to the data [61] and prior beliefs [53], affect their interpretation of the data. Studies [34, 57] have shown that readers could suggest different contexts in data-driven articles, albeit rarely.

Several ways have been investigated for the contextualization of data. Narrative visualizations support reader-driven exploration of data [11, 32, 73]. Social data analysis platforms such as sense.us [29], ManyEyes [83], CommentSpace [87], and r/dataisbeautiful from Reddit [44] have supported users to discuss with data. Contextifier [33] supports the understanding of stock charts better by annotating important events for the reader on the chart. In line with existing work, we aim to design and provide support for contextualizing data with online text in general.

2.2 In-text Support for Text Comprehension

Existing research has explored various interaction designs to support comprehension of texts. One common approach was providing an alternative representation of text context in different forms, such as generating visualizations for data-rich texts [4, 56] or linking presentation videos from the author [47]. ScholarPhi [28] and Paper Plain [3] support comprehending academic articles with in-context explanations of symbols and jargon. With unfamiliar numbers in the text, a series of prior work investigated re-expressing them in terms of numbers that the reader is likely to be familiar with [7, 35, 46, 48, 84]. Researchers also commonly explored designs and interactions for coupling data visualizations and texts in documents [9, 14, 18, 45, 51, 52, 64, 81] to support better comprehension of text. Another thread of research explored providing guiding questions to support readers critically reading texts, such as academic papers [62, 89], news articles [13, 49], or press releases [63].

Enriching the text with relevant external information was also commonly explored, such as references in academic papers [42, 66]. Contextifier [33] and NewsViews [24] supported readers by augmenting news articles with data visualization that depicts the data from the articles and provides additional contexts. As Contextifer and NewsViews focused on specific types of data (stock price or geographical data), a couple of approaches focused on developing pipelines that could encompass more general news articles [55, 78]. Building on top of existing work, we aim to develop in-text interactions for exploring related data.

2.3 Natural Language for Interacting with Data

Due to its expressivity and low entry barrier for expressing user intent for analysis, natural language text as a modality for interacting with visualizations has been at the center of interest for many researchers. It is often used as a modality to support exploring data by presenting visualization of data relevant to the user's intent, both in research prototypes [23, 31, 74, 75] and commercial products (e.g., Tableau [80], Microsoft PowerBI [10]). Recent work further extends to support multi-turn conversational interactions for data exploration [19, 79], which could guide users to discover their own insights from the data. For instance, Olio [76] has been proposed to support the exploration of data repositories of pre-created visualizations over diverse topics and datasets. Scoping down for close-ended user intent, question answering from tables [30, 43, 60] and charts [12, 39, 40] have also been widely investigated. With the recent advances with LLMs, recent work [15, 88] has demonstrated that LLMs can achieve high performance and interpretability for fact-checking and question-answering with data tables and handle complex questions requiring multiple reasoning processes. Our work shares the similar technical challenges of mapping natural language onto the most relevant dataset, with additional challenges of resolving vague and implicit references to the data and encompassing multiple topics. Grounding on existing work, we aim to develop a technical pipeline for mapping ambiguous statistical statements from online texts to the relevant datasets.

3 FORMATIVE STUDY

We conducted a formative study to understand in what cases people try to seek related information to contextualize statistical statements in online texts. Based on the findings, we propose design goals for supporting the contextualization of statistical statements.

3.1 Study Setup

We recruited seven participants who read online discussions on social media or online forums daily. The call for participation was posted on an online board of a university in South Korea. The average age of participants was 23 (Min = 21, Max = 25). We screened the participants by asking how often they read and participate in online discussions.

In the study, we chose 'online discussions' as the context, as we expected that online discussions would often refer to statistical data in a less rigorous yet more argumentative manner, which would lead to more natural motivations to better understand the data. We provided four discussion threads with references to statistical information from Reddit's r/changemyview [1], which is well-known as a place for civil discussions [36]:

- We should reward those who are in great shape with yearly bonuses with a tax on junk food [2]
- It is unethical to purchase residential properties for investment purposes [3]

[1] https://reddit.com/r/changemyview
[2] https://www.reddit.com/r/changemyview/comments/13quoqw/cmv_we_should_reward_those_who_are_in_great_shape/
[3] https://www.reddit.com/r/changemyview/comments/12cqfm0/cmv_it_is_unethical_to_purchase_residential

- The gender pay gap is largely caused by differences in goals in life and not by systematic discriminatory work practices [4]
- Nuclear Energy is way better than people think [5]

Participants were asked to choose two threads based on their interests and read them while thinking aloud for 20 minutes per thread, with the minimum requirement of reading the original post and three top-level comments. They were free to explore external information, such as clicking embedded links in the comments or searching for information on the web. After reading the threads, the participants were asked to search for external information of interest on the same two discussion threads for 10 minutes per thread. With a separate stage for searching for external information, we aimed to observe cases where participants had latent informational needs but did not spontaneously search for them. Next, we conducted a semi-structured interview with each participant on their information-searching behaviors and current challenges. Each session lasted 90-120 minutes, and participants were compensated 20,000 KRW (≈ 15 USD). After all sessions, we analyzed our observation notes and interview quotes with affinity diagramming.

3.2 Design Goals

From the interviews and observations, we discovered participants' common behaviors and challenges in finding information to contextualize and understand statistical information. Based on them, we propose three design goals for an interactive system to contextualize statistical statements from online texts with data:

G1. Provide Straightforward Responses to Users' Information Needs with Data. Participants (P1, P5, P7) commonly considered reading as the main goal, therefore expecting only a low cognitive burden for the secondary goal of seeking information. They often wanted to quickly obtain answers to their questions without disrupting their focus on reading discussion threads.

Once they decided to search, participants needed to formulate a search query expressing their information needs. However, when they were unsure about what type and scope of information was appropriate, they felt it challenging to start information exploration (P3, P5, P6). Therefore, we propose that the system be able to intuitively support users' informational needs while minimally disturbing the reading process.

G2. Facilitate Flexible Exploration of Contextual Information Around Data. While participants commonly used quick answer panels from the search engine or top-ranked documents to locate the data of interest, some participants (P1, P5, P6, P7) often wanted to seek data beyond what was available from a single source, such as comparing multiple data points or even different statistical datasets. Such cases were more common when they were verifying comparative statements from the discussion, or wanted to explore contexts involving personal relationships or curiosity. However, retrieving and combining multiple pieces of data required excessive effort of searching multiple times and cognitively processing them, making it challenging to conduct such comparisons by themselves. In such cases, the system should provide relevant data based on the user's diverse information needs and support their exploration of the context surrounding the issue and themselves.

G3. Enable Users to Evaluate the Reliability of Provided Statistical Data. Among the search results, participants sought relevant and credible information that answered their questions. They typically tried to verify suspicious claims within online discussions. Participants often evaluated the reputation of the websites providing the information (P1, P2, P4, P6, P7). However, they rarely looked deeper into the data itself, such as the source of the data and the definition of measurement, to assess the credibility of the data and its relevance to the users' information needs. We propose that the system should provide information on the data to assess its credibility and relevance, such as the producer of the data, the definition of the data, and the link to the raw data.

4 USER INTERACTION

We present *DataDive*, a system for facilitating the contextualization of statistical statements with data, and recommending questions to guide contextualization and interactions for exploration of statistical data surrounding the context of the text being read. *DataDive* is developed as a browser extension, and we also built a standalone version for the controlled experiment. With *DataDive*, the user can first specify the statistical statement of interest. Then, *DataDive* presents a list of recommendations for contextualization which the user can select to explore the data with interactive visualization.

4.1 Step 1. Initiating Context Exploration

To support exploring data with low cognitive burden (G1), *DataDive* provides three ways for users to start data exploration and express their needs for diverse types of data (Figure 2). First, the user can click on an underlined sentence within the text (Figure 2A). To support users to notice statistical statements and facilitate data exploration, *DataDive* pre-processes the text to mark out sentences with statistical information. Second, the user can select any text snippet and click the search button if they want to investigate a specific part of the text (Figure 2B). Third, the user can formulate their own statement of interest with a free-form input (Figure 2C).

4.2 Step 2. Browsing Recommended Contexts

Upon selecting or entering a statistical statement, *DataDive* presents a list of questions to guide the exploration of diverse aspects around the statistical statement (Figure 3), supporting the exploration of contextualization with less burden (G1). The first option is fixed to show the most relevant data for the provided statement (Figure 3A-1). Then, *DataDive* presents a list of pipeline-generated recommendations for contextualizing the statement considering the readers' context, which is self-stated to the system when they first use *DataDive* (Section 5.2). Each recommendation consists of a question (Figure 3A-2) and the associated values of relevant statistical indicators, entities, or time periods (Figure 3A-3) to support users in expecting specific data they would see for each recommendation.

[4] https://www.reddit.com/r/changemyview/comments/w4kbny/cmv_the_gender_pay_gap_is_largely_caused_by/
[5] https://www.reddit.com/r/changemyview/comments/ud5hde/cmv_nuclear_energy_is_way_better_than_people_think/

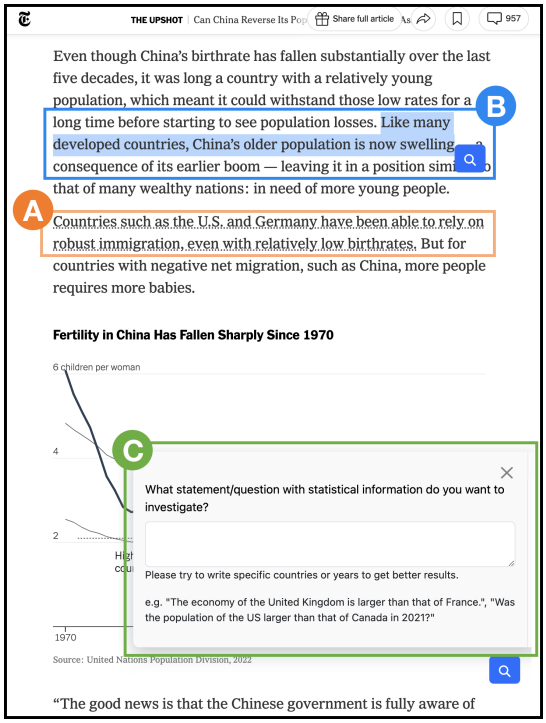

Figure 2: Three ways to start exploring contexts in *DataDive*. The user can (A) click on underlined statistical statements, (B) select the part of the text by themselves, or (C) type in their own statement of interest.

Table 1: The rules of *DataDive* for generating visualization.

Statistical Entities	Selected Indicators	Type of Selected Entities	
		Single Point	Duration
Single	Single	Bar	Single Line
	Multiple	Bar	Multi Line
Multiple	Single	Bar	Multi Line
	Multiple	Grouped Bar	Multi Line / Entity

4.3 Step 3. Investigating and Exploring Data

When the user selects a recommended question, *DataDive* queries the pipeline to retrieve the relevant set of statistical entities, indicators, and time (Section 5.3), fetches the data values, and visualizes them. The data presentation panel consists of dropdown menus, data visualization, and data sources (Figure 3B, C, D).

To support users in flexibly exploring different statistical entities, indicators, and time (G2) as well as correcting the potential errors of the pipeline by themselves, *DataDive* provides dropdowns (Figure 3B). Based on the user's selection for each field, *DataDive* generates either a bar chart or a line chart to present the selected data based on a pre-defined rule considering the number of selected values per field (Table 1). We focused on bar charts and line charts as they are the most common types of charts [8]. For visualizing data, we used the Vega-Lite [72] library.

Existing work on data science and data journalism points out that the provenance of the data is important information for judging the credibility and the potential bias of data [25, 59, 85], which is aligned with our design goals (G3). *DataDive* provides a "Data Source" panel to show the metadata of the selected statistical indicators (Figure 3D), including the data definition and source information.

5 TECHNICAL PIPELINE

DataDive utilizes a two-stage pipeline to recommend and retrieve contexts relevant to a statistical statement (Figure 4) [6].

Our pipeline takes a statistical statement and its surrounding text as input. Based on a context database of datasets that can be used to generate contextual information, *DataDive* (1) generates a ranked list of candidate contexts from the input statistical statement (Section 5.2) and (2) matches the top contexts in the ranked list with the relevant data in the context database (Section 5.3).

5.1 The Context Database

DataDive first requires a database of datasets to map the statistical statements on. To support easier expansion of datasets, the pipeline takes each dataset as two CSV files. The first file contains the numbers, consisting of one column containing the names of entities, one column containing the date information, and the other columns containing the values per each indicator, following a common structure used in previous work [5]. This format simplifies the integration of new datasets, providing a consistent and general structure for the data. The second file contains the metadata about each indicator/feature in the dataset that would be presented to the users, including the unit, the source of the data, and the explanation of the data.

For the evaluation, we prepopulated the database with a variety of datasets on social issues from reputable sources: greenhouse gas emissions and energy consumption data from Our World in Data [69, 70], and the World Development Indicator datasets from the World Bank [6], which encompasses a wide range of topics on global issues (e.g., global health, economics, education, environment, technology).

5.2 Stage 1. Generation of Ranked Candidate Contexts

In the first stage, our pipeline generates a ranked set of candidate contexts through a three-step process: (1) parsing the components of the statistical statement from the input sentence, (2) generating candidate contexts based on the parsed information, and (3) ranking the candidate contexts.

5.2.1 Step 1. Parse Statistical Statement in Input Sentence. First, *DataDive* parses the input sentence to extract the three key components of a statistical statement (Figure 5): (1) *(statistical) entity*, the subject of the statement (e.g., *'Korea'*); (2) the *feature* (e.g., value, trend) and the *indicator* used (e.g., *'fertility rate'*); and (3) *date*, the time point or period of interest (e.g., *'2019'*).

[6]The code for the pipeline is available at [https://github.com/kixlab/ClaimVis]

Figure 3: (A) UI for browsing the list of recommended questions and exploring data. (1) The first question is to verify the sentence with data. Each recommendation consists of (2) a question suggesting a different context and (3) the values of statistical indicators/entities/time periods. (B, C, D) After selecting one of the question from (A), *DataDive* provides (B) dropdown widgets to change selected statistical entities, indicators, and time periods, (C) visualization of selected data, and (D) metadata on data definition and source information.

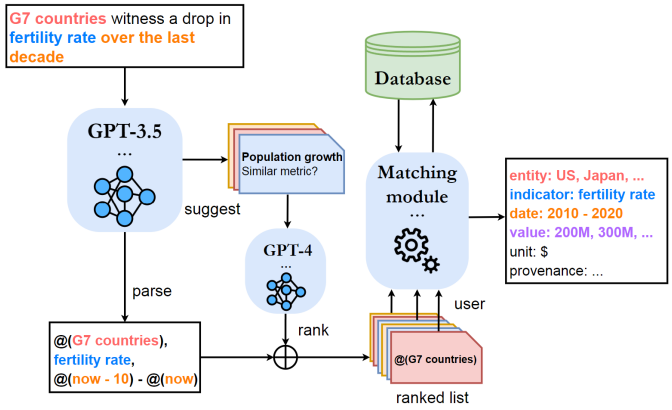

Figure 4: The pipeline starts by parsing a sentence into a tuple (entity, indicator, date) and generating related recommendations. These recommendations are ranked and displayed according to the ranks. User selection of a recommendation triggers a match with relevant data fields in top-k datasets. Finally, the system filters and presents the pertinent data.

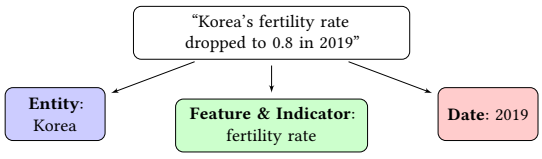

Figure 5: Example of parsing an input sentence into statistical entity, feature and indicator, and date.

While existing work tackled similar challenges of disambiguating natural language queries for data analysis [23, 74], our pipeline faces some unique challenges in handling diverse statistical statements from online texts which often implicitly refer to statistical data.

- **Feature & Indicator:** The value to be extracted might not exactly match the attributes in the tables or is only implicitly inferred from the statement. (e.g. "the number of people" implying "population", "Korean woman gave birth to 0.8 children on average in 2019" implying "fertility rate").
- **Entity:** The statement might refer to a group of entities that cannot be directly assigned to the names of entities (e.g., "Asian countries" or "Countries with a population of over 100 million").
- **Date:** The dates in the statement might rely on implicit inference within the context of the whole text. (e.g. "the last decade" or "10 years after")

Therefore, compared to existing work using semantic similarities with embedding vectors for disambiguation, we decided to leverage the capability of the LLM to handle such implicit references. In this step, such implicit references are first automatically annotated with a fine-tuned GPT-3.5 model, and are resolved in the later part of the pipeline (Section 5.3.2).

5.2.2 Step 2. Generate Candidate Contexts. Our pipeline generates a list of potential contexts based on the components extracted in Step 1. *DataDive* specifically considers four axes (Table 2): (1) *in-text*, which are based on the surrounding sentences; (2) *relational*, which are based on the similarities and relationships (e.g. causal, compositional); (3) *statistical*, which are statistically significant features along the component of interest (in our work, we specifically focus on global extrema); and (4) *personalized*, which are components that a reader is likely to have prior knowledge about due to personal connections or media coverage that can help internalize contents of the statistical statement.

DataDive uses the GPT-3.5 model to generate the candidates of features/indicators, statistical entities, and dates based on the

Table 2: Criteria for generating candidate contexts.

Component	Candidate Context Generation Criteria
Entity	[In-text] Statistical entities referred to in surrounding sentences [Relational] Similar or related statistical entities [Statistical] Statistical entities displaying extrema [Personalized] Statistical entities the reader is likely familiar with
Features & Indicators	[In-text] Features & indicators referred in surrounding sentences [Relational] Similar or related (e.g., compositional, causal) features & indicators [Statistical] Global extrema features
Date	[In-text] Dates & time periods referred to in surrounding sentences [Relational] Dates & time periods with similar or related events (e.g., financial crisis) [Personalized] Dates & periods the reader is likely familiar with

sentence of interest, its surrounding paragraph, and the readers' personal contexts. Our pipeline also generates a teaser question for each context, as shown in Section 4.2. We recombine the generated candidates and the original components to generate the (entity, feature, date) tuples that we use as candidate contexts.

5.2.3 Step 3. Rank Candidate Contexts. Next, *DataDive* ranks the candidate contexts by these four criteria: (1) helpfulness in the broader context of the statistical statement, (2) interestingness to the reader, (3) novelty of the perspective, and (4) likelihood of discovering supporting (or refuting) data. Following a practice from a previous work [65], we used the GPT-4 model to run pairwise comparisons between each context considering the four criteria and combined the comparison results to rank the contexts.

5.3 Stage 2. Extraction of Relevant Data

When the user selects one of the candidate contexts from Stage 1, *DataDive* retrieves the relevant datasets in our context database and extracts the relevant data from the selected datasets through a two-step process.

5.3.1 Step 1. Retrieve Relevant Datasets. DataDive begins by retrieving relevant datasets from the context database to narrow the search space for relevant indicators. To do so, we compute the semantic similarity of indicators in each dataset and the components extracted from the statistical statement of interest.

The pipeline uses *all-MiniLM-L6-v2* Sentence-BERT model [68] to produce sentence embedding vectors. For the given context set, the pipeline computes embedding vectors for three components extracted from the statistical statement (Section 5.2.1), along with the values for the ranked candidate contexts. The embedding vectors are collectively used as a representative set for the original statement, which we will refer to as the *statement set*. Similarly, the pipeline gathers *dataset set* for each dataset consisting of sentence embedding vectors for each indicator name. To improve computational efficiency, these *dataset sets* are pre-computed and cached.

Our pipeline selects the top 7 datasets (empirically set) with the highest maximum attribute similarity scores. To reduce the dataset size for later steps, we only keep the top 15 attributes per dataset (empirically set) and merge the datasets into a single data table.

5.3.2 Step 2. Locate Context in Data Table. When the user selects a candidate context or the original context from the recommendation panel, *DataDive* locates the context within the data table. The pipeline then specifically tries to match the (entity, feature & indicator, date) tuples with the data table from Step 1. It first attempts to match each component directly with the table attributes. When the components do not match the attributes exactly, such as variations (e.g., "GDP" vs. "G.D.P."), synonyms (e.g., "birth rate" vs. "fertility rate"), and annotated implicit references (Section 5.2) (e.g., "Companies with high revenue"), the pipeline attempts to resolve them with the following strategies:

- **Feature & Indicator**: We compute the semantic similarity between the indicator and the attributes from the data table using the *all-MiniLM-L6-v2* sentence BERT model and select the attribute with the highest similarity score.
- **Entity**: We apply semantic matching to entities without implicit references. Implicit references are resolved following the approach from existing work [15]. First, we use GPT-3.5 to convert them into SQL queries, which are then executed to retrieve the corresponding data. If SQL translation fails, we use the LLM's comprehensive knowledge base to use an end-to-end question-answering process.
- **Date**: We first evaluate the expression using Python's `eval()` function, setting *now* to the current year. For more complex date expressions, we employ techniques akin to those used for Entity resolution.

After accurately mapping the `(Feature & Indicator, Entity, Date)` tuple to the table's corresponding attributes, we extract specific values. For instance, given the tuple `(GDP, United States, 2010)`, we filter the table to exclusively present the GDP figures for the United States in 2010.

5.4 Statistical Check-worthiness

Apart from the two-stage pipeline, *DataDive* also supports identifying data-related and check-worthy sentences within the text the user is reading. This function is particularly useful for highlighting statistical statements in the UI, as detailed in Section 4.1. For assessing the check-worthiness of these statements, we use the ClaimBuster API [27], a web-based tool for live automated fact checking. Furthermore, to determine the relevance of the data, *DataDive* tailors a prompt for the GPT-4 model, following the approach of Liang et al. [54]. This prompt enables GPT-4 to provide a binary judgment on each sentence, 'Y' for relevance and 'N' otherwise. A statement is deemed highlight-worthy if it surpasses a threshold of 0.5 (empirically set) in check-worthiness and obtains a 'Y' in data relevance.

6 PIPELINE EVALUATION

To understand whether the pipeline of *DataDive* can produce meaningful questions for exploring contexts surrounding the statistical statements and to match the statistical statements to the relevant datasets capably, we conducted a technical evaluation of the pipeline. For the evaluation, we focused on statistical statements from online texts on global social issues. Therefore, we only had cases where statistical entities were countries and regions. We used the World Bank's World Development Indicators [6] and Our World In Data's Greenhouse Gas Emissions [69] and Energy Consumptions [70] dataset.

6.1 Materials

In our study, we compiled 77 statistical statements and their contextual paragraphs from various online sources, including data journalism sections of The New York Times and The Economist, and social media platforms like Twitter, Reddit (r/changemyview, r/economics, r/news, r/dataisbeautiful), Quora, as well as blogs and non-governmental organization publications. We manually reviewed recent social media posts and news articles and searched for news articles and social media posts on global population and climate change. From the articles and posts, we extracted statistical statements related to the World Bank's World Development Indicators.

6.2 Participants

We recruited 18 external evaluators from university communities in South Korea, ensuring that they had no previous involvement or knowledge in the study. These evaluators were chosen based on their interest in social issues, experience searching statistical information, and English proficiency to understand English texts on social issues. The evaluators were aged between 18 and 24 (Mean = 20.4), and were compensated with 40,000 KRW (≈30 USD).

6.3 Procedure

We randomly divided the statistical statements into three sets and assigned evaluators to each set. Each evaluator rated 25-26 statistical statements in total. Due to the dropouts, the number of evaluators per statement differed from five to seven.

We provided an evaluation system for the evaluators to read the statistical statement and its surrounding paragraph and examine the recommended questions and the set of indicators-year-country triplet matched to the statistical statement.

Then, the evaluator rated the quality of the pipeline results with 5-point Likert scale questions (1 = Strongly Disagree to 5 = Strongly Agree) on the (a) meaningfulness and the relevance of individual recommendations, (b) meaningfulness, diversity, and interestingness of the recommendation set, and (c) relevance of matched statistical entities, indicators, and time. The detailed questions are available on the appendix.

6.4 Results

The evaluation results showed that the pipeline could generate meaningful context recommendations for statistical statements and match them with relevant data sets.

6.4.1 Meaningfulness of Individual Recommendation Item. Figure 6A shows the evaluators' evaluation results of the top 5 recommendations. In general, participants considered that the recommendations generated by the pipeline would help readers gain meaningful insights into the statement's context. The average rating for meaningfulness for all recommendation questions was 4.12 (SD = .55). The rating was consistent for the top 5 questions, with average ratings between 4.08 and 4.20. We observed 13 instances where the average rating was lower than 3. Five cases involved recommending other countries with extreme values for statements on trends of an explicit set of countries. Three cases involved recommending years with extreme values of a statistical indicator for a single country for statements on comparing multiple countries. Other cases involved recommending unrelated statistical indicators or countries.

Also, the evaluators considered that the statistical entities mapped to each recommendation were relevant. Out of 385 recommendations with individual evaluations, only one case had a relevance score of less than 3.

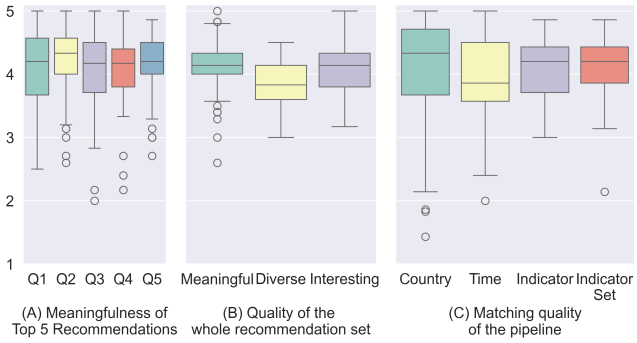

Figure 6: Boxplot of the average rating of recommendations and the matching quality from the pipeline. (A, B) The results show that evaluators considered the recommendations meaningful, diverse, and interesting. (C) While evaluators were generally favorable to the matching quality, matching for the country had the most errors.

6.4.2 Quality of Recommendation Set. Figure 6B shows the evaluators' evaluation results of the whole recommendation set. The average rating for the meaningfulness (4.12 / 5, SD = .43), diversity (3.86 / 5, SD = .35), and interestingness (4.07 / 5, SD = .40) of the recommendation set suggests that participants considered the recommendation set as useful.

6.4.3 Relevance of matched statistical entities/indicators/time. Figure 6C shows the participants' evaluation results of the matching quality. Participants responded that the pipeline could get a set of data points relevant to the given statistical statement. The pipeline retrieved relevant countries and regions for 84.4% (65 out of 77) with an average rating of 4.02 / 5 (SD = .90), relevant time periods for 92.2% (71 out of 77) with an average rating of 3.92 / 5 (SD = .69), and relevant sets of statistical indicators for 100% (77 out of 77) of the cases with an average rating of 4.12 (SD = .48). A total of 80.5% (62 / 77) of the statements had relevant matching for all three components. The list of potentially relevant statistical indicators was rated as relevant for 98.7% (76 out of 77) of the cases.

We observed more errors for countries and regions, with two common types. The first type was when the statement did not directly refer to countries or regions, and the pipeline needed to infer the country from the context. The second type was when the statement involved extreme values and explicit references to countries, such as "Out of the large emitters the USA has the highest emissions per capita." The pipeline retrieved the list of the top 5 countries in this case rather than the explicitly referred entity (USA).

7 USER EVALUATION

We conducted a user evaluation of *DataDive* to assess how *DataDive* supports the contextualization of statistical statements for users while they are reading texts and how it benefits the reading process. Specifically, we aimed to explore the following research questions:

- RQ1 Does using *DataDive* support readers to explore more context around statistical statements from the text?
- RQ2 Does using *DataDive* support the readers to gain more knowledge on the topic of the text?
- RQ3 How do people cognitively/emotionally feel about using *DataDive*?

7.1 Participants

We recruited 21 participants with interests in and experiences with reading texts on social issues from universities in South Korea. For the screening, participants responded with their prior interests in global social issues on a 5-point Likert scale as well as open-ended responses on the specific topics of interest, and we filtered out participants without a high level of interest. Also, to ensure that the participants were able to understand English articles, we gave two short reading comprehension tasks in English and filtered participants with wrong answers. Participants (13 male, 8 female) were aged from 19 to 31 (M = 22.9) with diverse nationalities including South Korea (N=12), Kazakhstan (N=3), Bangladesh, Indonesia, Kyrgyzstan, Poland, Russia, and Spain (each with N=1), Each participant was compensated with 30,000 KRW (≈ 22 USD).

7.2 Conditions

We conducted a within-subjects experiment with two conditions:

- Baseline (BL): Participants could use the web search freely, such as making queries to better contextualize content from the text they read. We selected the baseline to be close to the current behaviors.
- *DataDive* (Ours): In addition to using the web search, participants were able to use the full features of the system, including highlights on statistical claims, suggestions for contextualization, and data exploration.

7.3 Procedure

Participants read two texts for the *DataDive* condition and the Baseline condition, respectively. Participants were instructed to read the text thoroughly and encouraged to consider what information they wanted to learn more, especially related to their personal contexts. Before using the system, we briefly introduced it and let the participant use it for 10 minutes to get used to it.

7.3.1 Materials. We took two articles from the New York Times about global issues with statistics. The articles were on carbon emissions [7] (Carbon Emission) and decreasing fertility rate [8] (Fertility Rate). To manage the reading time, the articles were reduced to around 500 words by removing various portions of the text.

7.3.2 Task. The participants repeated the task twice, once with *DataDive* and once with only Baseline. The order of the texts and conditions were counterbalanced. In each task, participants were given 5 minutes to skim through the text and another 20 minutes to read it in detail. While reading the text in detail, they could search the web for external information to aid their understanding. Participants could also use *DataDive* depending on the experimental condition. During the session, we observed the participants' behaviors while reading the text. Thereafter, they responded to a post-task survey. We first asked open-ended questions about what they recalled from the text, the external information exploration, and what questions they had while reading the text. We further asked 7-point Likert-scale questions on their reading experience and information search experience. Finally, we asked for the task load using NASA-TLX [26] and the usability of the system using SUS [38] (for *DataDive* condition).

7.3.3 Post-Task. After both tasks, we interviewed each participant. We asked participants about their workflow when conducting searches, understanding what made them search, and when they used the system. We also asked about their experience with the system and how it may have supported them in contextualizing the text. Participants were asked to reflect on how their reading experience with the system differed from how they might have read similar content in the past. Following this, we delved into their perceptions of the system regarding what they liked or disliked about it, the issues they faced, and suggestions for improvement. We also asked about their willingness to use the system in the future and for what scenarios they would.

7.3.4 Data Analysis. To analyze the user behaviors, one of the authors counted the number of web searches and the instances of using *DataDive* in each session.

To analyze the interview responses, we first transcribed all interview recordings with ClovaNote [9], and the two authors collaboratively extracted common themes from the interview transcripts.

To analyze the open-ended responses to participants' recalled facts after the task and the recalled questions, we invited three external evaluators who were capable of analyzing English and Korean responses. One evaluator analyzed both recalled facts and the recalled questions, while the other two participated in either one of the tasks. During the analysis process, two evaluators worked together to establish shared criteria. The evaluators created a list of recalled facts and a list of recalled questions per participant. Then, they annotated ones that were (a) about the data itself (e.g. specific values or trends) and (b) about the interpretation of data (e.g. data definitions, causal factors related to the trend, participants' own insight). For recalled facts, evaluators additionally annotated ones that were from out of the text.

8 USER EVALUATION RESULTS

In this section, we present how *DataDive* supported the contextualization of statistical statements and how it affected participants' comprehension of data articles with statistical statements.

[7] https://www.nytimes.com/2016/04/06/upshot/promising-signs-that-economies-can-rise-as-carbon-emissions-decline.html
[8] https://www.nytimes.com/2023/02/09/upshot/china-population-decline.html
[9] https://clovanote.naver.com

Table 3: The average behavior counts to explore external information. The difference was significant for the Fertility Rate article and for the total.

	DataDive	Baseline	*p*-value	Stats
Fertility Rate	7.20	5.00	0.018	U = 21.5
Carbon Emission	7.36	6.00	0.195	U = 36.5
Total	7.29	5.48	0.017	W = 119

8.1 Assessment of RQ1

Our analysis of behavior logs and survey suggested that *DataDive* led the participants to explore more external information, while the number of questions participants had did not show a significant difference across the conditions.

From the behavior logs, we observed that the participants attempted to search for more external information when using *DataDive* compared to the Baseline condition. (Wilcoxon Signed-Rank Test, p = 0.017, W = 119). The between-subject comparison showed that participants with *DataDive* searched more in the case of *Fertility Rate* article (Mann-Whitney U test, p = 0.018, U = 21.5), while the difference was not significant for the *Carbon Emission* article (Mann-Whitney U test, p = 0.195, U = 36.5) (Table 3).

From the survey, we compared the number of self-recalled questions for each condition (*DataDive* vs. Baseline) (Table 4). From the within-subject comparisons, we did not observe a statistically significant difference in the number of questions (Wilcoxon Signed-Rank Test, p = 0.070, W = 102.5 for total question count, p = 0.903, W = 79.5 for data-related question count, p = 0.750, W = 54.0 for interpretation-related question count), suggesting that *DataDive* supported the participants to explore more context per question. However, the self-perception of participants' reading behaviors did not show significant differences between *DataDive* and the baseline condition (Table 5).

From interviews and observations, we found several instances of *DataDive* facilitating to explore more contexts around the statistical statements. We first introduce the common types of external information participants wanted and describe how *DataDive*'s underlines on statistical statements and recommendations of potential contexts supported exploring contexts around data.

During the tasks, participants had diverse motivations to search for external information in both the *DataDive* and baseline conditions. The most common case was searching for supporting evidence for statistical statements (10/21 participants). The simplest cases were checking whether the numbers or trends shown in the text aligned with the raw data they encountered. Participants also sought whether the trends from the data and their explanations were persistent and generalizable to other contexts, such as whether the decoupling between carbon emission and economic growth described in the *Carbon Emission* article was actually a prolonged trend or not (8/21 participants). Additionally, searches to gain more knowledge about the topic were common, such as the meaning of words (e.g., "Baby bonus" and "Speed premium" policy for boosting birth rates, "Paris Agreement" for reducing carbon emissions), and the definitions of the statistics mentioned in the text (8/21 participants).

8.1.1 In-text Underlines Nudging Exploration. Among the three ways to initiate exploration (underlined sentences, self-selected phrases, and entering into a search box), the underlined sentences were perceived to be more convenient and likely more reliable in ensuring that the system would provide high-quality results. 12 participants mentioned that *DataDive*'s underlined sentences on statistical statements nudged them to pay greater attention to the statement and be more interested in exploring the recommended questions for them. Out of 232 instances of using *DataDive* to explore statistical information, 141 instances (60.8%) started from clicking the underlined sentences and 61 instances from the free-form selections. P19 said, "As [underlined] highlights picked statistical statements, I could pay more attention to them, which I might have missed if I read [the article] without the system."

8.1.2 Question Recommendations Encouraging Further Exploration. Participants commonly considered that the recommendation feature of *DataDive* encouraged them to explore the datasets. Out of 240 clicks on the recommended questions, 165 instances (68.8%) were clicks on the questions generated by the pipeline and 75 (31.2%) instances were clicks on the default question ("Does the data support this statement?"), showing that participants often clicked the recommended questions. Six participants valued the recommended questions as they provided novel aspects of the issue. Eight participants thought recommended questions on statistical indicators helped them expand their thoughts by considering novel aspects.

Among the recommended questions, personally relevant questions drew some participants' interest to explore further. This was particularly prominent for questions mentioning the participants' home and neighboring countries where they were interested to see what the situation reported in the text would be like in their own country (5/21 participants). The questions took the scope of their interpretation beyond the text to consider how it relates to themselves or other contexts they were aware of.

8.1.3 Nudging Features Potentially Leading to Passive Reading. However, at the same time, some participants (P3, P18) were concerned that the underlining and recommended questions would discourage them from thinking about their own perspectives and instead rely on the system's guidance. P3 commented that "The list of questions made me read the text more passively. I could see the questions when I clicked the sentence, so I didn't really think questions by myself. However, the provided questions were not really good ones touching the core of the issue, so I felt I was a bit trapped in some frame." Participants also pointed out cases where the recommended questions were shallow and repetitive and where a greater variety would be appreciated (5/21 participants). P3 said that "The provided questions were too simple and specific. They were too focused on the facts themselves. I think the more important part is about the narrative with the numbers, such as cases where the fertility rate dropped even when policies for fertility rates were in action."

8.2 Assessment of RQ2

Our analysis of the number of recalled facts and the self-evaluation on the knowledge gain did not show statistically significant results from using *DataDive*, but participants shared some beneficial cases from using *DataDive*.

Table 4: The number of recalled questions per each article and condition. We did not observe statistically significant differences.

	Carbon Emission (Mann-Whitney U test)				Fertility Rate (Mann-Whitney U test)				Total (Wilcoxon Signed-Rank test)			
	DataDive	Baseline	p-value	U	*DataDive*	Baseline	p-value	U	*DataDive*	Baseline	p-value	W
Questions	3.64	5.10	0.830	51.5	4.10	4.36	0.190	36.5	3.86	4.71	0.070	102.5
Data-related Question	2.09	2.10	0.971	54.0	1.70	1.73	0.942	53.5	1.90	1.90	0.903	79.5
Interpretation-related Question	2.45	2.00	0.457	44.5	1.30	1.36	0.853	52.0	1.90	1.67	0.750	54.0

Table 5: Participants' self-rating on exploring external information. There was no statistical significance between the *DataDive* (Ours) and baseline (BL) conditions. The detailed questions are available from the appendix.

	Carbon Emission		Fertility Rate		Total	
	Ours	BL	Ours	BL	Ours	BL
Motivated for searching	6.09	5.90	6.20	6.27	6.05	6.10
Success of finding information	4.64	5.30	4.70	5.45	4.71	5.38
Ease of finding information	4.91	4.90	4.90	5.18	4.90	5.05
Ease of understanding the information found	5.00	5.30	5.30	6.00	5.05	5.67
Trustworthiness of information found	5.55	5.70	5.70	4.91	5.62	5.29
Novelty of information found	5.09	5.00	4.90	4.45	4.81	4.71

From the post-task survey, we discovered that participants recalled more facts from the text and the information search in the Baseline condition than in *DataDive* condition (Wilcoxon Signed-Rank test, p = 0.011, W = 172.5) (Table 6). However, the number of recalled facts on data or data interpretation did not show significant differences across conditions. This suggests that using *DataDive* led participants to recall relatively more data-related facts.

Participants' self-evaluation on their perceived knowledge gain from the task did not show significant differences between the conditions (Table 7).

Still, we discovered notable cases where *DataDive* supported more discovery on the topic of the text from the interviews and observation. Participants generally felt that *DataDive* provided easy and flexible ways to explore data.

8.2.1 Increasing Awareness on Statistical Indicators. Some participants (P5, P12) valued that *DataDive* guided the use of data to observe social issues. P5 shared an example of how *DataDive* helped them decide which indicators to use, "When the text says 'Countries A and B emit [carbon dioxide] a lot', then I would have vague thoughts on seeing the carbon emissions of these countries and drag on the country names. *DataDive* suggested 'CO2 emissions per capita', and I realized that emission per capita would be a more relevant indicator in this case." Similarly, P12 said, "Statistical statements often refer to the data abstractly, like 'economic growth rate'. There are multiple different metrics for economic growth, and it's hard to recall them even though I already know them. The system provides a list of relevant indicators, which helped me recall them."

Also, with the variety of statistical indicators being available to the participants, participants had new discoveries by swapping and comparing data (7/21 participants). From the Fertility Rate article, P6 clicked a statement on Sweden's parental leave to discover a recommended question "What impact does the longer parental leave duration have on the labor force participation rate in Sweden?" From the matched datasets, P6 unexpectedly discovered the data on the ratio between female and male participation in the labor force, which was almost equal to parity. P6 commented that the data supported what they found about gender equality in Sweden from their previous web search, leading to more impressive knowledge.

8.2.2 Critical Examination of Text. The participants commonly used *DataDive* to verify the trends and values of data within the text (12/21 participants). They reflected that such verification behaviors were not common in their everyday reading, yet was helpful for reading the text more critically (9/21 participants).

Participants also mentioned that *DataDive* supported the identification of statistical fallacies by providing a larger context. P18 gave a specific example, "There are statistical fallacies, like the trend could be felt very different by how the data was normalized, or the data was relative or absolute. To be aware of such [fallacies] from the articles I have to read the text in detail and think critically. But with the system, I can discover such differences with a couple of clicks, which helped me critically read the statistics." During the tutorial, P11 read a short piece of text claiming that automation did not cause the loss of jobs and by exploring the system, commented that "After I saw the actual data, I realized how the author was cherry-picking the data. The unemployment rate in the past was high due to the subprime mortgage crisis, and I thought the text was distorting the statistics. I wouldn't have noticed it if I didn't check the past data beyond what was written in the text."

8.3 Assessment of RQ3

While the usability scores from standardized metrics suggest that the usability of *DataDive* needs more improvement, participants generally valued the system for supporting to focus on browsing and interacting with data to contextualize the statement.

The average SUS score for the system was 74 (Min = 40, Max = 97.5, SD = 15.61), suggesting that participants had mixed opinions on the usability of the system. Task load between *DataDive* and the baseline condition did not show a significant difference, showing that using *DataDive* did not incur additional task load (Table 8).

Here, we list some commonly mentioned factors that affected participants' usability and usefulness of the system.

8.3.1 Credibility of Provided Information. During the user study, *DataDive* showed data only from authoritative sources, with metadata on the data sources (Section 5.1). Participants typically considered that the provided data by *DataDive* was credible (9/21 participants).

8.3.2 Conciseness of Results. Where web search results tended to contain a variety of media and data formation from a diverse range

Table 6: The number of recalled facts per each article and condition. We did not observe statistically significant differences between conditions, except for the number of recalled facts from out of the text in the Carbon Emission article and in total.

	Carbon Emission (Mann-Whitney U test)				Fertility Rate (Mann-Whitney U test)				Total (Wilcoxon Signed-Rank test)			
	Ours	BL	p-value	U	Ours	BL	p-value	U	Ours	BL	p-value	W
Recalled facts	5.00	7.20	0.092	31.0	6.90	6.91	0.972	54.0	5.95	7.05	0.068	92.5
Recalled facts from out of the text	2.64	4.10	0.008	18.0	2.50	3.45	0.178	36.0	2.57	3.76	0.011	172.5
Recalled facts on data	3.09	4.30	0.144	34.5	4.80	4.36	0.614	47.5	3.90	4.33	0.282	99.5
Recalled facts on data from out of the text	2.50	2.00	0.344	41.5	2.18	1.70	0.347	41.5	1.86	2.33	0.240	112.5
Recalled facts on data interpretation	3.82	4.30	0.825	51.5	4.60	4.09	0.564	46.5	4.19	4.29	0.791	70.5
Recalled facts on data interpretation from out of the text	2.36	2.20	0.938	53.5	1.70	1.91	0.589	47.0	2.05	2.05	1.000	76.0

Table 7: Participants' self-rating on their gain after the experiment. There was no statistical significance between *DataDive* (Ours) and the baseline (BL) condition. The detailed question is available from the appendix.

	Carbon Emission		Fertility Rate		Total	
	Ours	BL	Ours	BL	Ours	BL
Knowledge gain	6.27	5.50	5.80	6.09	6.05	5.81
Critical attitude	5.73	6.30	5.30	6.09	5.52	6.19
Enjoyment	6.00	6.00	5.60	6.00	5.81	6.00
Personal relevance	5.09	4.80	4.40	5.36	4.76	5.10
Exposure to greater context	5.55	5.30	5.40	5.82	5.48	5.57

Table 8: Participants' task load for each condition with the NASA-TLX questionnaire. There was no statistical significance between *DataDive* (Ours) and baseline (BL) conditions.

	Carbon Emission		Fertility Rate		Total	
	Ours	BL	Ours	BL	Ours	BL
Mental Demand	2.73	2.80	2.70	3.09	2.81	2.95
Physical Demand	1.45	1.80	1.60	1.64	1.52	1.71
Temporal Demand	1.91	2.10	1.80	2.09	1.86	2.10
Performance	3.45	3.90	3.80	3.73	3.67	3.81
Effort	3.55	3.80	3.80	3.73	3.76	3.76
Frustruation	2.09	2.10	2.40	2.36	2.33	2.24

of sources, the focused use case of *DataDive* was appreciated in removing the bloat of unnecessary content while still providing vital and meaningful information (4/21 participants). Furthermore, *DataDive* was appreciated for being more precise and efficient for statistical searches as participants can gain access to and wrangle with the data immediately to fit their purpose rather than having to sift through several web pages until they get to the relevant result or perhaps even fail to do so (4/21 participants).

8.3.3 Focused search. The structured nature of *DataDive* also allowed participants to be focused on diving into their exploration of the system as compared to web search where the breadth of information may distract them and lead them to searches that deviate from their original intent (3/21 participants).

8.3.4 Easier Comparisons Between Data. Participants valued that *DataDive* provided easier interactions and options for freely comparing the data (11/21 participants). When searching for statistical data from the web, they commonly relied on image search or charts provided by the search engine, but felt limited as pre-generated charts show only a static set of indicators over pre-selected countries and time (5/21 participants).

9 DISCUSSION

In this section, we discuss potential directions for future research in contextualizing statements with statistical data. First, we discuss additional application contexts that *DataDive* can support. Then, we discuss the design decisions of our LLM-powered pipeline to produce factually reliable yet diverse outputs. To generalize our findings, we discuss considerations for organically expanding the database of datasets to encompass diverse data sources. Lastly, we discuss the limitations of our study and future directions.

9.1 Incorporation of Contextual Information Beyond Structured Data

While *DataDive* focuses on supporting the discovery and exploration of relevant contexts based on statistical data in structured forms, contextual information in other forms (e.g., natural language text, visualizations, images) can introduce additional valuable information. In fact, a third of our user study participants mentioned that they would have appreciated contextual information beyond just the statistics in *DataDive*, such as governmental reports (P8) and news articles (P11). For example, information in Wikipedia [10] or informational videos can provide general background knowledge on the topics [2]. Temporal information, such as important events related to statistical entities, can be augmented with data from news media, similar to the approach employed in Contextifier [33]. Furthermore, official reports from governments and global organizations as well as research articles can provide credible interpretations of data. Such additional information can be delivered to users through graphical overlays [50] and text annotations. For example, in Figure 1C, *DataDive* could provide annotations of important economic events in Sweden on the chart of GDP.

Including contextual information beyond structured data can also enable improvements in the recommendations generated by *DataDive* and its visualization interface by leveraging knowledge for generating recommendations of related entities or dates. For instance, in the example in Figure 1B, *DataDive* could suggest potential mediating variables between economic growth and fertility rate from external reports [22].

[10] https://en.wikipedia.org

9.2 Considerations for Factually Reliable LLM-Powered System

In this work, we leveraged an LLM-based pipeline for two features: recommending potential contexts around statistical statements and mapping implicit data references and recommended contexts in statistical statements to existing statistical entities, indicators, and time periods. To design the two features, we utilized the generative capacity of LLMs in different ways.

For recommending potential contexts, we focused on diversity, using the LLM's generation capability to create a range of potential contexts. On the other hand, for mapping the contexts to statistical entities and indicators, we prioritized relevance so that the pipeline would provide trustworthy information. Therefore, we avoided solely relying on LLMs for mapping data references and generated the names of entities using LLMs as a final resort. Instead, we matched references with semantic similarities first, and we used LLMs to generate a method for resolving entities, such as creating SQL statements to resolve the entities, to ensure more trustworthy responses. Since LLMs can hallucinate, we designed the system to produce a list of potential alternatives and allowed the users to make the final selection. To support users in making the decision, we also provided extra information such as teaser questions to explain how the recommended contexts relate to the text and the explanation of statistical indicators to describe how the contexts are related to the indicators.

However, for mapping implicit data references, we found that providing multiple candidates can cause more challenges. Errors related to countries and dates were easily noticed and recoverable, but participants had trouble recovering from incorrect statistical indicators as there were often many similar statistical indicators. With false positives, participants felt that it was difficult to determine whether the matched indicators were actually related to the statement, which, in the worst case, could have led to false interpretations. For false negatives, participants had difficulty deciding how to recover from the error as they were unsure whether the relevant indicators were present in the system or if their input statement was wrong. To address this, providing more accessible and concise information for the users to understand and verify *DataDive*'s output could better support them in handling errors. By communicating the errors and uncertain cases to users, for example by providing visual warnings with uncertain cases, users' mental models of *DataDive* can also be improved.

9.3 Considerations for an Organically Growing Context Database

While we prepopulated the context database with datasets from Our World in Data [69, 70] and the World Bank [6], we envision that the support for contextualization by *DataDive* would apply to texts in other domains as well, such as local data [37]. To achieve this, users can potentially upload or provide links to their data based on their needs to *DataDive*, which would organically grow the database. We expect that this feature would attract a broader pool of users interested in domains not necessarily covered by our seed datasets.

However, we foresee several challenges in supporting the organic expansion of the system. Ensuring the credibility of the data [25, 59, 85] would be one challenge. Another issue would be ensuring the integrity and consistency of the data across different datasets, which is known to be one of the most effortful and time-consuming processes from data science literature [41, 71, 90]. Also, as *DataDive* encompasses more data, *DataDive* would need to be able to disambiguate a large set of indicators or entities with similar names. One potential solution would be having moderators to manage the expanding datasets and build rules from data definitions and formats of existing centralized sources, such as Our World In Data [11], Data Commons [12], or FRED [13]. With the established rules, moderators can manage the new datasets added to the system by validating against seed datasets and enforcing the matching format. We envision that such a collaborative ecosystem for organizing trustworthy and coherent datasets and supporting the collaborative analysis of data [58] will further enrich *DataDive*.

9.4 Limitations and Future Work

Generalizing Evaluation & User Study Findings. We evaluated the pipeline of *DataDive* with only text on global social issues and a database on global social issues. To verify the generalizability of the pipeline, future work can evaluate the pipeline with different sets of indicators and statistical entities from various contexts.

For the user evaluation, we only tested the system with news articles on social issues in a controlled setting. We expect that the motivations and behaviors of using *DataDive* would differ according to users' goals when reading text and the type of text. Also, in this study, we evaluated the system only with college students. We expect that the level of expertise and experience with data and the topic of the article would affect the value of *DataDive*. In future work, we expect that longitudinal studies in real-life settings with diverse participants, such as a diary study, could capture the usefulness of *DataDive* in diverse settings.

Improving Runtime Efficiency. Due to the multiple calls to the LLM, the delay in responding to user input was often noticeable, which hampered the user experience (P4, P5, P6, P7, P9, P20). We believe that the runtime efficiency of the pipeline can be further improved by parallelizing calls to the LLM or leveraging more efficient LLM models with fine-tuning. Furthermore, preprocessing sentences while the user is reading through the text could reduce wait time.

10 CONCLUSION

The meaning of statistical statements differs by what context it was interpreted in. While readers may consider their own or relevant contexts when reading statistical statements to develop their own understanding, it is challenging to do so due to the barriers in searching for contextual information around them. We built *DataDive*, an interactive system to support the contextualization of statistical statements in texts by recommending diverse potential contexts around the statements and providing relevant data, powered by an LLM-based pipeline to parse statistical statements, generate potentially relevant contexts, and map statements to the datasets. Our

[11] https://www.ourworldindata.org
[12] https://www.datacommons.org/
[13] https://fred.stlouisfed.org/

evaluation results showed that the pipeline could produce meaningful recommendations and retrieve relevant datasets, and that the users of *DataDive* could easily explore contextualizations around statistical statements.

ACKNOWLEDGMENTS

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-01347, Video Interaction Technologies Using Object-Oriented Video Modeling). Dae Hyun Kim is partly supported by the G-CORE Research Project grant at KAIST.

REFERENCES

[1] Ashley Ahn. 2023. South Korea has the world's lowest fertility rate, a struggle with lessons for us all. *NPR* (March 2023). https://www.npr.org/2023/03/19/1163341684/south-korea-fertility-rate

[2] Tarfah Alrashed, Lea Verou, and David Karger. 2022. Wikxhibit: Using HTML and Wikidata to Author Applications that Link Data Across the Web. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3526113.3545706

[3] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Transactions on Computer-Human Interaction* (2023). https://doi.org/10.1145/3589955 Just Accepted.

[4] Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. 2019. Elastic Documents: Coupling Text and Tables through Contextual Visualizations for Enhanced Document Reading. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 661–671. https://doi.org/10.1109/TVCG.2018.2865119 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[5] Oana Balalau, Simon Ebel, Théo Galizzi, Ioana Manolescu, Quentin Massonnat, Antoine Deiana, Emilie Gautreau, Antoine Krempf, Thomas Pontillon, Gérald Roux, and Joanna Yakin. 2022. Statistical Claim Checking: StatCheck in Action. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 4798–4802. https://doi.org/10.1145/3511808.3557198

[6] The World Bank. 2023. World Development Indicators. Retrieved Oct 06, 2023 from https://datatopics.worldbank.org/world-development-indicators/

[7] Pablo J. Barrio, Daniel G. Goldstein, and Jake M. Hofman. 2016. Improving Comprehension of Numbers in the News. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2729–2739. https://doi.org/10.1145/2858036.2858510

[8] Leilani Battle, Peitong Duan, Zachery Miranda, Dana Mukusheva, Remco Chang, and Michael Stonebraker. 2018. Beagle: Automated Extraction and Interpretation of Visualizations from the Web. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–8. https://doi.org/10.1145/3173574.3174168

[9] Fabian Beck and Daniel Weiskopf. 2017. Word-Sized Graphics for Scientific Texts. *IEEE Transactions on Visualization and Computer Graphics* 23, 6 (June 2017), 1576–1587. https://doi.org/10.1109/TVCG.2017.2674958 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[10] Power BI. 2024. Retrieved Jan 25, 2024 from https://powerbi.microsoft.com/.

[11] Jeremy Boy, Francoise Detienne, and Jean-Daniel Fekete. 2015. Storytelling in Information Visualizations: Does it Engage Users to Explore Data?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 1449–1458. https://doi.org/10.1145/2702123.2702452

[12] Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3512–3521.

[13] Xiang 'Anthony' Chen, Chien-Sheng Wu, Lidiya Murakhovs'ka, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2023. Marvista: Exploring the Design of a Human-AI Collaborative News Reading Tool. *ACM Transactions on Computer-Human Interaction* (2023). https://doi.org/10.1145/3609331 Just Accepted.

[14] Zhutian Chen and Haijun Xia. 2022. CrossData: Leveraging Text-Data Connections for Authoring Data Documents. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3517485

[15] Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding Language Models in Symbolic Languages. *ICLR* (2023).

[16] Byeong-Young Cho, Peter Afflerbach, and Hyeju Han. 2018. Strategic processing in accessing, comprehending, and using multiple sources online. *Handbook of multiple source use* (2018), 133–150.

[17] George W. Cobb and David S. Moore. 1997. Mathematics, Statistics, and Teaching. *The American Mathematical Monthly* 104, 9 (1997), 801–823. http://www.jstor.org/stable/2975286

[18] Matthew Conlen. 2021. Idyll Studio: A Structured Editor forAuthoring Interactive & Data-Driven Articles. (2021), 12.

[19] Kedar Dhamdhere, Kevin S. McCurley, Ralfi Nahmias, Mukund Sundararajan, and Qiqi Yan. 2017. Analyza: Exploring Data with Conversation. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. Association for Computing Machinery, New York, NY, USA, 493–504. https://doi.org/10.1145/3025171.3025227

[20] Sara Doan. 2021. Misrepresenting COVID-19: Lying With Charts During the Second Golden Age of Data Design. *Journal of Business and Technical Communication* 35, 1 (Jan. 2021), 73–79. https://doi.org/10.1177/1050651920958392 Publisher: SAGE Publications Inc.

[21] Sheena Erete, Emily Ryou, Geoff Smith, Khristina Marie Fassett, and Sarah Duda. 2016. Storytelling with Data: Examining the Use of Data by Non-Profit Organizations. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1273–1283. https://doi.org/10.1145/2818048.2820068

[22] International Monetary Fund. 2022. The New Economics of Fertility. Retrieved Oct 10, 2023 from https://www.imf.org/en/Publications/fandd/issues/Series/Analytical-Series/new-economics-of-fertility-doepke-hannusch-kindermann-tertilt

[23] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*. Association for Computing Machinery, New York, NY, USA, 489–500. https://doi.org/10.1145/2807442.2807478

[24] Tong Gao, Jessica R. Hullman, Eytan Adar, Brent Hecht, and Nicholas Diakopoulos. 2014. NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3005–3014. https://doi.org/10.1145/2556288.2557228

[25] Jonathan Gray, Lucy Chambers, and Liliana Bounegru. 2012. *The Data Journalism Handbook*. O'Reilly Media, Inc.

[26] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[27] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: The First-Ever End-to-End Fact-Checking System. *Proc. VLDB Endow.* 10, 12 (aug 2017), 1945–1948. https://doi.org/10.14778/3137765.3137815

[28] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–18. https://doi.org/10.1145/3411764.3445648

[29] Jeffrey Heer, Fernanda B. Viégas, and Martin Wattenberg. 2007. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*. ACM Press, San Jose, California, USA, 1029–1038. https://doi.org/10.1145/1240624.1240781

[30] Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. TAPAS: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4320–4333. https://doi.org/10.18653/v1/2020.acl-main.398 arXiv:2004.02349 [cs].

[31] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2018. Applying Pragmatics Principles for Interaction with Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan. 2018), 309–318. https://doi.org/10.1109/TVCG.2017.2744684 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[32] J. Hullman and N. Diakopoulos. 2011. Visualization Rhetoric: Framing Effects in Narrative Visualization. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec. 2011), 2231–2240. https://doi.org/10.1109/TVCG.2011.255 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[33] Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. 2013. Contextifier: Automatic Generation of Annotated Stock Visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI*

'13). Association for Computing Machinery, New York, NY, USA, 2707–2716. https://doi.org/10.1145/2470654.2481374

[34] Jessica Hullman, Nicholas Diakopoulos, Elaheh Momeni, and Eytan Adar. 2015. Content, Context, and Critique: Commenting on a Data Visualization Blog. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, Vancouver BC Canada, 1170–1175. https://doi.org/10.1145/2675133.2675207

[35] Jessica Hullman, Yea-Seul Kim, Francis Nguyen, Lauren Speers, and Maneesh Agrawala. 2018. Improving Comprehension of Measurements Using Concrete Re-expression Strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173608

[36] Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. Designing for Civil Conversations: Lessons Learned from ChangeMyView. (2017), 11.

[37] Ian G. Johnson, Aare Puussaar, Jennifer Manuel, and Peter Wright. 2018. Neighbourhood Data: Exploring the Role of Open Data in Locally Devolved Policymaking Processes. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–20. https://doi.org/10.1145/3274352

[38] Patrick W Jordan, Bruce Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. 1996. *Usability evaluation in industry*. CRC Press.

[39] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5648–5656.

[40] Kushal Kafle, Robik Shrestha, Scott Cohen, Brian Price, and Christopher Kanan. 2020. Answering questions about data visualizations using efficient bimodal fusion. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*. 1498–1507.

[41] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. 2011. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization* 10, 4 (Oct. 2011), 271–288. https://doi.org/10.1177/1473871611415994 Publisher: SAGE Publications.

[42] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-based Exploration and Organization of Scientific Literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3526113.3545660

[43] Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. 2020. Scrutinizer: A Mixed-Initiative Approach to Large-Scale, Data-Driven Claim Verification. http://arxiv.org/abs/2003.06708 arXiv:2003.06708 [cs].

[44] Tobias Kauer, Marian Dörk, Arran L. Ridley, and Benjamin Bach. 2021. The Public Life of Data: Investigating Reactions to Visualizations on Reddit. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3411764.3445720

[45] Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. 2018. Facilitating Document Reading by Linking Text and Tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, Berlin Germany, 423–434. https://doi.org/10.1145/3242587.3242617

[46] Juho Kim, Eun-Young Ko, Jonghyuk Jung, Chang Won Lee, Nam Wook Kim, and Jihee Kim. 2015. Factful: Engaging Taxpayers in the Public Discussion of a Government Budget. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 2843–2852. https://doi.org/10.1145/2702123.2702352

[47] Tae Soo Kim, Matt Latzke, Jonathan Bragg, Amy X. Zhang, and Joseph Chee Chang. 2023. Papeos: Augmenting Research Papers with Talk Videos. https://doi.org/10.1145/3586183.3606770 arXiv:2308.15224 [cs].

[48] Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. 2016. Generating Personalized Spatial Analogies for Distances and Areas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 38–48. https://doi.org/10.1145/2858036.2858440

[49] Eunyoung Ko, Yeonsu Kim, and Juho Kim. 2022. ReviewAid: A Scaffolded Approach to Supporting Readers' Evaluation of Health News. In *Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022, pp. 313-320*. International Society of the Learning Sciences.

[50] Nicholas Kong and Maneesh Agrawala. 2012. Graphical overlays: Using layered elements to aid chart reading. *IEEE transactions on visualization and computer graphics* 18, 12 (2012), 2631–2638.

[51] Nicholas Kong, Marti A. Hearst, and Maneesh Agrawala. 2014. Extracting references between text and charts via crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 31–40. https://doi.org/10.1145/2556288.2557241

[52] Shahid Latif, Zheng Zhou, Yoon Kim, Fabian Beck, and Nam Wook Kim. 2021. Kori: Interactive Synthesis of Text and Charts in Data Documents. https://doi.org/10.48550/arXiv.2108.04203 arXiv:2108.04203 [cs].

[53] Crystal Lee, Tanya Yang, Gabrielle D Inchoco, Graham M. Jones, and Arvind Satyanarayan. 2021. Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445211

[54] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=iO4LZibEqW Featured Certification, Expert Certification.

[55] Allen Yilun Lin, Joshua Ford, Eytan Adar, and Brent Hecht. 2018. VizByWiki: Mining Data Visualizations from the Web to Enrich News Articles. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 873–882. https://doi.org/10.1145/3178876.3186135

[56] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2023. Charagraph: Interactive Generation of Charts for Realtime Annotation of Data-Rich Paragraphs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581091

[57] Brian James McInnis, Lu Sun, Jungwon Shin, and Steven P. Dow. 2020. Rare, but Valuable: Understanding Data-centered Talk in News Website Comment Sections. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–27. https://doi.org/10.1145/3415245

[58] Kristi Morton, Magdalena Balazinska, Dan Grossman, and Jock Mackinlay. 2014. Support the data enthusiast: challenges for next-generation data-analysis systems. *Proceedings of the VLDB Endowment* 7, 6 (Feb. 2014), 453–456. https://doi.org/10.14778/2732279.2732282

[59] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300356

[60] Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir Radev. 2021. FeTaQA: Free-form Table Question Answering. https://doi.org/10.48550/arXiv.2104.00369 arXiv:2104.00369 [cs].

[61] Evan M. Peck, Sofia E. Ayuso, and Omar El-Etr. 2019. Data is Personal: Attitudes and Perceptions of Data Visualization in Rural Pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300474

[62] Zhenhui Peng, Yuzhi Liu, Hanqi Zhou, Zuyu Xu, and Xiaojuan Ma. 2022. CReBot: Exploring interactive question prompts for critical paper reading. *International Journal of Human-Computer Studies* 167 (Nov. 2022), 102898. https://doi.org/10.1016/j.ijhcs.2022.102898

[63] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3580907

[64] Joao Pinheiro and Jorge Poco. 2022. ChartText: Linking Text with Charts in Documents. http://arxiv.org/abs/2201.05043 Number: arXiv:2201.05043 arXiv:2201.05043 [cs].

[65] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2023. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. *arXiv e-prints*, Article arXiv:2306.17563 (June 2023), arXiv:2306.17563 pages. https://doi.org/10.48550/arXiv.2306.17563 arXiv:2306.17563 [cs.IR]

[66] Napol Rachatasumrit, Jonathan Bragg, Amy X. Zhang, and Daniel S Weld. 2022. CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 707–719. https://doi.org/10.1145/3490099.3511162

[67] Josh Radinsky and Iris Tabak. 2022. Data practices during COVID: Everyday sensemaking in a high-stakes information ecology. *British Journal of Educational*

Technology 53, 5 (2022), 1221–1243. https://doi.org/10.1111/bjet.13252 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13252.
[68] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410
[69] Hannah Ritchie, Max Roser, and Pablo Rosado. 2020. CO_2 and Greenhouse Gas Emissions. *Our World in Data* (2020). https://ourworldindata.org/co2-and-greenhouse-gas-emissions.
[70] Hannah Ritchie, Max Roser, and Pablo Rosado. 2022. Energy. *Our World in Data* (2022). https://ourworldindata.org/energy.
[71] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[72] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2017. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan. 2017), 341–350. https://doi.org/10.1109/TVCG.2016.2599030 Conference Name: IEEE Transactions on Visualization and Computer Graphics.
[73] Edward Segel and Jeffrey Heer. 2010. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1139–1148. https://doi.org/10.1109/TVCG.2010.179
[74] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. 2016. Eviza: A Natural Language Interface for Visual Analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. Association for Computing Machinery, New York, NY, USA, 365–377. https://doi.org/10.1145/2984511.2984588
[75] Vidya Setlur, Enamul Hoque, Dae Hyun Kim, and Angel X. Chang. 2020. Sneak Pique: Exploring Autocompletion as a Data Discovery Scaffold for Supporting Visual Analysis. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 966–978. https://doi.org/10.1145/3379337.3415813
[76] Vidya Setlur, Andriy Kanyuka, and Arjun Srinivasan. 2023. Olio: A Semantic Search Interface for Data Repositories. http://arxiv.org/abs/2307.16396 arXiv:2307.16396 [cs].
[77] Ben Shneiderman. 2020. Data Visualization's Breakthrough Moment in the COVID-19 Crisis. https://medium.com/nightingale/data-visualizations-breakthrough-moment-in-the-covid-19-crisis-ce46627c7db5
[78] Levy Silva and Luciano Barbosa. 2022. Matching news articles and wikipedia tables for news augmentation. *Knowledge and Information Systems* (Dec. 2022). https://doi.org/10.1007/s10115-022-01815-0
[79] Arjun Srinivasan and Vidya Setlur. 2021. Snowy: Recommending Utterances for Conversational Visual Analysis. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 864–880. https://doi.org/10.1145/3472749.3474792
[80] Tableau. 2024. Retrieved Jan 25, 2024 from https://www.tableau.com/.
[81] Edward R Tufte. 2001. *The visual display of quantitative information*. Vol. 2. Graphics press Cheshire, CT.
[82] Pertti Vakkari. 2016. Searching as learning: A systematization based on literature. *Journal of Information Science* 42, 1 (2016), 7–18. https://doi.org/10.1177/0165551515615833 arXiv:https://doi.org/10.1177/0165551515615833
[83] Fernanda B. Viegas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. 2007. ManyEyes: a Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov. 2007), 1121–1128. https://doi.org/10.1109/TVCG.2007.70577
[84] Yanan Wang and Yea-Seul Kim. 2023. Making Data-Driven Articles more Accessible: An Active Preference Learning Approach to Data Fact Personalization. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. ACM, Pittsburgh PA USA, 1353–1366. https://doi.org/10.1145/3563657.3595986
[85] Wibke Weber, Martin Engebretsen, and Helen Kennedy. 2018. Data stories. Rethinking journalistic storytelling in the context of data journalism. *Studies in Communication Sciences* 18, 1 (Nov. 2018), 191–206–191–206. https://doi.org/10.24434/j.scoms.2018.01.013 Number: 1.
[86] C. J. Wild and M. Pfannkuch. 1999. Statistical Thinking in Empirical Enquiry. *International Statistical Review* 67, 3 (1999), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-5823.1999.tb00442.x
[87] Wesley Willett, Jeffrey Heer, Joseph Hellerstein, and Maneesh Agrawala. 2011. CommentSpace: structured support for collaborative visual analysis. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. ACM Press, Vancouver, BC, Canada, 3131. https://doi.org/10.1145/1978942.1979407
[88] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large Language Models Are Versatile Decomposers: Decomposing Evidence and Questions for Table-Based Reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (Taipei, Taiwan) (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 174–184. https://doi.org/10.1145/3539618.3591708
[89] Kangyu Yuan, Hehai Lin, Shilei Cao, Zhenhui Peng, Qingyu Guo, and Xiaojuan Ma. 2023. CriTrainer: An Adaptive Training Tool for Critical Paper Reading. (2023).
[90] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *arXiv:2001.06684 [cs, stat]* (April 2020). http://arxiv.org/abs/2001.06684 arXiv: 2001.06684.

A APPENDIX

A.1 Details of Questions for Technical Evaluation

Here, we describe the exact questions used for the technical evaluation.

- Quality of individual recommendations
  - The question would help readers **gain meaningful insights into the context** surrounding the statement.
  - The **suggested choices of statistical indicators, countries, or time frames are relevant** to the suggested question.
- Quality of recommendation set
  - The set of questions would help readers **gain meaningful insights into the context** surrounding the statement.
  - The set of questions would help readers **consider diverse aspects of the context** surrounding the statement.
  - The readers will be **interested in clicking the questions** to see the relevant context around the statement.
- Relevance of matched statistical entities/indicators/time
  - The **selected set of countries and regions is relevant** to the given statistical statement.
  - The **selected time period is relevant** to the given statistical statement.
  - The **selected set of statistical indicators** is relevant to the given statistical statement.
  - I was able to **find statistical indicators relevant to the statement from the list of statistical indicators** in the dropdown menu.

A.2 User Survey Questions

Here, we describe the exact questions used for the user evaluation.

A.2.1 Self-evaluation of context exploration behavior.

- **Motivated to search**: I was motivated to search for external information on the questions I had while reading the text.
- **Success of finding information**: I was able to find the information that I wanted while exploring.
- **Ease of finding information**: I felt it easy to find the information that I wanted while exploring.
- **Ease of understanding the information found**: The information I found while exploring can be easily understood.
- **Trustworthiness of information found**: The information that I found while exploring was trustworthy.
- **Novelty of information found**: The information that I found while exploring was new and beyond my expectations.

A.2.2 Self-evaluation of positive effects from context exploration.

- **Knowledge gain**: I gained more knowledge about the topic of the text through this task.

- **Critical attitude**: I became more critical of the content of the text while reading and exploring external information.
- **Enjoyment**: I enjoyed reading the text and exploring external information.
- **Personal relevance**: I felt that (some of) the content I came across while reading the text and exploring external information was personally relevant to me.
- **Exposure to greater context**: I felt that I was exposed to the greater context surrounding the text while reading the text and exploring external information.