“The Data Says Otherwise” – Towards Automated Fact-checking and Communication of Data Claims

Yu Fu
fuyu@gatech.edu
Georgia Tech
Atlanta, United States

Shunan Guo
sguo@adobe.com
Adobe Research
San Jose, United States

Victor S.Bursztyn
soaresbu@adobe.com
Adobe Research
San Jose, United States

Jane Hoffswell
jhoffs@adobe.com
Adobe Research
San Jose, United States

Ryan Rossi
ryrossi@adobe.com
Adobe Research
San Jose, United States

John Stasko
john.stasko@cc.gatech.edu
Georgia Tech
Atlanta, GA, United States

ABSTRACT

Fact-checking data claims requires data evidence retrieval and analysis, which can become tedious and intractable when done manually. This work presents *Aletheia*, an automated fact-checking prototype designed to facilitate data claims *verification* and enhance data evidence *communication*. For verification, we utilize a pre-trained LLM to parse the semantics for evidence retrieval. To effectively communicate the data evidence, we design representations in two forms: data tables and visualizations, tailored to various data fact types. Additionally, we design interactions that showcase a real-world application of these techniques. We evaluate the performance of two core NLP tasks with a curated dataset comprising 400 data claims and compare the two representation forms regarding viewers' assessment time, confidence, and preference via a user study with 20 participants. The evaluation offers insights into the feasibility and bottlenecks of using LLMs for data fact-checking tasks, potential advantages and disadvantages of using visualizations over data tables, and design recommendations for presenting data evidence.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; **Empirical studies in visualization**; • **Information systems** → *Information retrieval query processing*.

KEYWORDS

automated fact-checking, information visualization, data-driven storytelling

ACM Reference Format:
Yu Fu, Shunan Guo, Victor S.Bursztyn, Jane Hoffswell, Ryan Rossi, and John Stasko. 2024. “The Data Says Otherwise” – Towards Automated Fact-checking and Communication of Data Claims. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3654777.3676359

1 INTRODUCTION

Imagine skimming through an ESPN analysis comparing two athletes' performances, reading a New York Times article about COVID trends, or browsing Fox News coverage on the upcoming U.S. election. These articles likely contain many data claims. However, data claims can contain inaccuracies from various sources. Errors and omissions might arise from oversights during the composition phase, where analysts engage in data analysis and manually transcribe insights into textual form [21]. Additionally, the frequent updates in data can exacerbate discrepancies between the data and the claims made. More alarmingly, bad actors may deliberately manipulate or fabricate data facts to advance specific agendas or propaganda. Regardless of intent, such flawed data claims contribute to a flood of misinformation that inundates and contaminates our information ecosystem, potentially misleading the public.

A standard practice to mitigate misinformation is through fact-checking, a process of assessing the veracity of textual claims based on authoritative or trusted evidence. Typically undertaken by professional fact-checkers within news organizations, fact-checking has long held a vital role in upholding the accuracy and integrity of information [30]. However, with manual fact-checking challenged by the increasing column of information production and dissemination, both practitioners and researchers are turning to advanced technology, notably *automated fact-checking (AFC)* [1, 31].

Automated fact-checking applies to many scenarios, especially in news platforms and social media. Researchers in computational journalism [22, 27] have advocated for sophisticated technologies to enhance and facilitate fact-checking tasks. Such tools can serve multiple roles for fact-checkers and journalists alike, and empower news readers to critically audit the content they consume [28]. Recently, the data mining and natural language processing (NLP) communities have contributed to a growing body of research to address this demand [101, 102], with a particular emphasis on downstream tasks [53, 55, 108], domain-specific requirements [54, 92], and the provision of annotated claim-evidence datasets [32, 89] for model training. The majority of prior work has mainly concentrated on text-based evidence, wherein claims are verified by cross-referencing them with a textual corpus of established facts, including sources like Wikipedia pages [78] and scientific articles [93].

Data claims, also known as numeric or statistical claims [17], use natural language to describe facts/insights derived from structured data and/or statistics. We posit that the veracity of data claims

UIST '24, October 13–16, 2024, Pittsburgh, PA, USA

is intrinsically tied to specific quantitative datasets, leading to a divergence in tasks from the conventional text-based fact-checking pipeline. This divergence can manifest at various stages, primarily evidence retrieval and presentation. While some existing solutions for automated fact-checking translate natural language into SQL queries [43, 47], these methods often struggle to address more complex insights (e.g., anomalies, trends, associations, etc.). These methods also require a pre-established knowledge base or set of SQL query candidates tailored to the dataset, limiting their applicability to new datasets. Furthermore, despite recognizing the importance of providing corroborating justification [31, 90], research on data evidence presentation and optimal representation forms for diverse data claims is scarce. Effective presentation of data evidence can not only bolster persuasiveness but also empower viewers to pinpoint inconsistencies between evidence and outcomes.

Thus, this work tackles two primary research questions: **Q1**. *"how do we enable out-of-the-box automated fact-checking for data claims?"* and **Q2**. *"how do we effectively represent and communicate data evidence?"* Building upon an established fact-checking framework [31], we first propose a modified automated data fact-checking framework (Figure 1) comprising six components: *data claim detection* (C1), *text-to-data mapping* (C2), *data evidence retrieval* (C3), *verdict determination & presentation* (C4), *data evidence presentation* (C5), and *end user interaction* (C6). We design and develop a prototype fact-checking system, *Aletheia*, to integrate these components, serving as both a proof-of-concept for our proposed framework and a design prototype that showcases its feasibility for practical applications.

This pipeline uses GPT models for downstream NLP tasks, leveraging its innate semantic parsing abilities [77, 107]. Informed by a content analysis of data claims in sixteen real-world articles, our prompting pipeline uses seven steps (Figure 2) to transform claims into data fact specifications [25, 80, 97]. This transformation enhances the transparency and interpretability of the connections between claims, the pertinent data subsets, and the derived insights.

For more effective data evidence communication, we introduce twenty-six data evidence representations across both data tables and visualizations for thirteen data fact types. To improve *Aletheia*'s practical utility, we incorporate interactions to facilitate stakeholders' fact-checking needs, such as overriding AI-induced mistakes.

We evaluate *Aletheia*'s two key components: the performance of the core steps in our *claim-to-data* transformation pipeline (**Q1**), and the effectiveness of the data evidence presentation when comparing the data table and visualization (**Q2**). In particular, we assess the backend pipeline on a manually curated dataset of 400 claims across various types (Section 5), demonstrating the LLM's promising capabilities in classifying data facts and converting natural language data claims to data fact specifications. Regarding **Q2**, we conducted a mixed-method user study with 20 participants tasked with reviewing data claims (Section 6). Our findings indicate that visualization charts outperform data tables in terms of the reviewing time for most data fact types (7 out of 13), enhance participant confidence across all data fact types, and are preferred in the majority. Drawing from our findings, we ultimately put forth four general design recommendations for effectively presenting data evidence.

2 RELATED WORK

This work is driven by prior research in automated fact-checking, existing strategies on justification presentation, and techniques for visually linking between text, tables, visualizations, and data.

2.1 Automated Fact-checking

Automated fact-checking has garnered significant attention in the NLP community to help counter misinformation and disinformation. Extensive research effort has been dedicated towards downstream tasks, including claim detection [34, 35], evidence retrieval [33, 55], verdict prediction [72, 86], justification production [41, 68], and more. Guo et al. [31] consolidate these tasks into a cohesive framework outlining the essential components for automated fact-checking systems. We refined this framework specifically for data claims, leading to the creation of *Aletheia*.

Traditionally, automated fact-checking systems was grounded on knowledge bases, verifying claims against a textual corpus of accumulated facts (e.g., Wikipedia pages [78], scientific articles [93], or knowledge graphs [87]). These systems rely on pre-established, reliable information sources to identify related supporting claims as evidence and determine the veracity based on the coherence with the evidence. In contrast, our work focuses on data claims that are not explicitly in the knowledge base but can be inferred from structured data tables. Data tables are a ubiquitous medium for storing information across various applications, and practitioners(e.g., data analysts, business analysts, etc.) often create text reports to summarize insightful statistics.

Previous research in text-to-data matching has tackled similar challenges, linking entities in text paragraphs to data tables through semantic parsing [36, 60, 65, 88, 105]. Additionally, chart reasoning techniques have been applied to fact-checking applications, focusing on verifying the correctness of data statements with a given chart image. For instance, Akhtar et al. introduced two baseline datasets [3, 4] for generating explainable fact-checking results over chart images. Most of these methods operate within supervised settings, which require expensive training on extensive documents, data tables, and chart images. Alternatively, unsupervised solutions [2] often suffer from limited scalability and unstable performance [95]. Another line of research addresses this problem by translating natural language claims into SQL queries and validating the claimed values against the queried results. For example, AggChecker [43] maps data claims to a probability distribution over a set of candidate SQL queries. While this method operates in an unsupervised manner, expanding the system to accommodate new datasets requires updates to the query candidates and probabilistic models to account for new table schemas. Similarly, Scrutinizer [47] employs an NL-to-SQL translation strategy but integrates an additional mixed-initiative pipeline that permits input from human experts to guide the translation process. However, the initial translation model relies on machine learning classifiers trained exclusively on the schema of the input data table, thereby constraining its adaptability to new datasets.

Considering the broad accessibility of pretrained LLMs (e.g., GPTs) and the proven ability in initial fact-checking trials [16, 38, 84], this study delves into a fact-checking solution harnessing the integrated capability of pretrained LLM. Our primary goal is to explore an "out-of-the-box" fact-checking solution designed for non-expert practitioners, such as journalists and business analysts, who often lack the resources for model training and may not possess an in-depth understanding of complex fact-checking models.

2.2 Verdict and Justification Presentation

When assessing verdicts through automated approaches, it is crucial to communicate its fact-checking decisions to the fact reviewers with comprehensible justifications [24, 29, 31, 53, 90]. Consequently, effectively presenting fact-checking results emerges as a vital research aspect that culminates at the end of the fact-checking process [82]. The primary method of conveying verdicts involves the use of veracity indicators [7], i.e., graphical elements succinctly encapsulating the veracity of claims on truth scales. For instance, color codings are commonly applied to present fact-checking outcomes in a variety of fact-checking research endeavors [43, 75, 89]. Public-facing fact-checking platforms, such as PolitiFact and Snopes, often include more comprehensive fact-check ratings that encompass not only varying levels of claim truthfulness but also categories such as scam, outdated, or research-in-progress to enhance credibility. ClaimViz [76] presents a visual analytics system that supports journalists in reviewing large amounts of factual claims and identifying check-worthy ones.In an effort to provide guidance on effective presentation of verified information to fact-checking report readers, Hettiachchi et al. [37] identified six critical design elements in fact-checking reports and studied their impact on improving the credibility and presentation of the reports with crowd-sourced experiments.

In the pursuit of enhancing the explainability of automated fact-checking systems, automated fact-checking research has employed different approaches [24, 31, 90], including summarization (extractive and abstractive) [11, 53], logic-based [18], attention-based [73, 81], and counterfactual [103] methods. Vallayil et al. [90] specifically examines the application of explainable AI (XAI) to automated fact-checking, highlighting significant challenges existing in multiple aspects, including the current lack of datasets that facilitate the explanations production and the ambiguity surrounding different concepts and taxonomy (e.g., global vs. local explainability). More recent studies (e.g.[5]) aim to provide datasets for explainable fact-checking. However, it is worth noting existing explainable fact-checking research also predominately resolves around claims and evidence presented in unstructured text, whereas our work centers on data-driven claims and evidence rooted in structured quantitative information that requires distinct forms of presentation.

2.3 Linking Data to Visual Representations

When reviewing data-driven claims, identifying relevant data sources serves as the cornerstone for assessing the veracity [46, 67]. Consequently, effectively communicating the underlying data to viewers during the process is pivotal in developing data fact-checking systems. The HCI community has made substantial contributions in advancing the realm of efficient data communication and content consumption within data documents using other visual representations, including data tables and visualization charts. For instance, Kong et al. [52] developed an interactive document viewer with the reference among text, tables, and visualization charts established by crowdsourced workers. Kim et al. [49] automated the association between text and table cells using NLP techniques, enabling the interactive highlighting of relevant table cells in response to user-selected sentences. Badam et al. [13] proposed to connect text and tables through the generated contextual visualizations to enhance the reading experience. Latif et al. introduced Kori [56], a mixed-initiative interface designed to facilitate the authoring process of interactive data documents by offering both recommendations for linking text with charts and manual construction of references. These techniques effectively link textual content with predefined tables or visualizations embedded in the same document. In our fact-checking context, we consider the entire dataset behind the scenes, with the audience exposed solely to the textual content.

Within this context, Chen and Xia developed CrossData [21], an authoring assistance system that retrieves backend data and presents it in table or visualization form during the document authoring process. CrossData primarily focuses on providing rich interactions to facilitate author-driven associations between text, tables, and visualizations during document creation. Conversely, our work is centered on devising means to effectively communicate data evidence relevant to claims. Zhi et al. [106] highlighted the positive impact of linking visualizations and text in storytelling on aspects such as comprehension, engagement, and recall. However, the objectives differ regarding the communication of data evidence, encompassing efficiency, data consistency, and user confidence in the verdict. In our work, we explore two visual representations — data tables and visualization charts, as means to effectively present data evidence. We assess their impact on efficiency, user confidence, and preference during the claim review tasks through user studies.

3 A FRAMEWORK FOR DATA CLAIM FACT-CHECKING AND COMMUNICATION

Although the four-part NLP framework formulated by Guo et al. [31] encapsulates the essence of fact-checking tasks, the downstream tasks are rooted in knowledge-based fact-checking research [78, 87, 93] focusing on claims sourced from qualitative evidence. Our work, in contrast, focuses on fact-checking **data claims**: natural language sentences with one or more facts from *quantitative* information.

Fact/knowledge-based claims and data claims differ in the nature of their evidence. The former, like the statement "*The director of the movie 'Oppenheimer' also directed 'Interstellar',*" are typically verifiable through historical records, direct evidence, or established knowledge. In contrast, a data claim, such as "*The total gross of 'Oppenheimer' accounts for 20% of the worldwide box office gross of all films directed by Christopher Nolan,*" relies on retrieving and aggregating a collection of data points across specific measurements (e.g., gross), aligning with the claimed insight type (i.e., proportion) and matching the asserted value against the gathered data.

While sharing parallel goals with conventional automated fact-checking endeavors, our focus on data claims necessitates a distinct approach to processing claims and conveying data evidence. To align closely with our focus, we have adapted the NLP framework proposed by Guo et al. [31] to model the automated fact-checking and communication process for data claims. This modified framework (Figure 1), comprises the following six components:

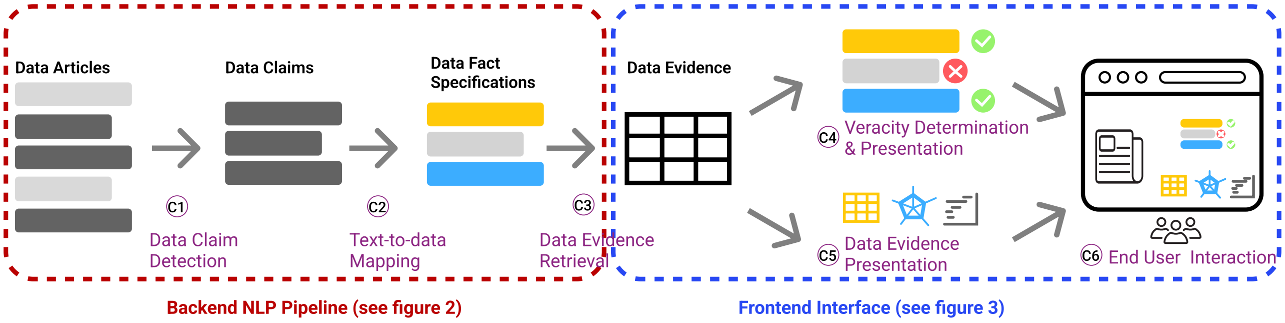


Figure 1: An overview of our modified framework for automated data claim fact-checking and communication, based on Guo et al.'s framework [31]. The process begins by extracting data claims from data articles. These claims are mapped into data fact specifications designed to fetch pertinent evidence. This evidence not only aids in determining the veracity of the associated data claim but also serves as the justification for the verdict. The initial three components constitute a pipeline dedicated to the NLP tasks. This NLP pipeline (Figure 2) underpins *Aletheia*'s backend. The last three components connect to *Aletheia*'s interface (Figure 3).

C1. Data Claim Detection. Our modified framework starts with the extraction of individual data claims from data articles or reports. This focus diverges from the conventional NLP approach to claim identification, which primarily concentrates on assessing the 'check-worthiness' and 'checkability' of claims [10, 24, 31, 66, 76].

C2. Text-to-data Mapping. The extracted data claims are transformed into corresponding data fact specifications, with *Text-to-data Mapping* accommodating the distinct evidence retrieval process and data aggregations involved in data claims. Researchers have developed frameworks and taxonomies to encapsulate the diversity and characteristics of such 'data facts/insights' [6, 20, 57, 69, 104]. We use the term ***data fact*** as the granular representation of the data insight extracted from the textual claim. Particularly, we adopt the specifications of data fact from previous research [80, 97] to define the core data and insight within the claim descriptors such as data fact types, subspace, value, aggregation, measure, etc.

C3. Data Evidence Retrieval. The data fact specifications are used to retrieve pertinent ***data evidence***, i.e., the subset of data directly related to the claim. This process differs significantly from conventional NLP-focused evidence retrieval that searches for credible information from large text corpus/knowledge bases [78, 87, 93] or incorporating additional metadata [96]. In our work, data evidence consists of structured formats of numeric information. For instance, fact-checking a data claim about COVID-19 trends can utilize datasets from official/authoritative sources (e.g., WHO [70]). Thus, our work assumes the availability of a specific dataset to check against and concentrates on retrieving the relevant data evidence.

C4. Veracity Determination & Presentation. Unlike knowledge-based fact-checking approaches that rely on pre-existing credible text excerpts, our method employs the procured data evidence and the associated data operations. We assess the veracity of a data claim by comparing the computed values/statistics with those claimed in the text. Note that veracity assessment can depend on three fundamental dimensions: *clarity* at the linguistic level, *consistency* with the data at the factual level, and *conformity* with the logic at the reasoning level. While our work encompasses steps to disambiguate the linguistic expressions of data claims, our scope remains on the factual level — ensuring that the textual description aligns with the actual data but not considering veracity at the reasoning level.

C5. Data Evidence Presentation. The data evidence and operations need to be communicated to users for verdict justification. Unlike presentations used for qualitative evidence [24, 31, 90], our work involves quantitative evidence, such as the subset of data and the statistical logic/rules associated with claimed data insights. This departure leads us into the realm of HCI and data visualization, which has received limited attention in previous studies [37]. In this work, we aim to investigate innovative methods for more effective communication of data evidence (introduced in Subsection 4.2).

C6. End User Interaction. Previous fact-checking frameworks have often overlooked human involvement, which can improve fact-checking outcomes by clarifying semantics, correcting AI errors, and making the fact-checking outcomes more actionable. With diverse end-users, their interaction needs can vary greatly; for instance, authors or editors may revise a data article to ensure accuracy, whereas fact-checkers aim to explain problematic data claims to a broader audience. Our work addresses these needs by proposing user interactions with AI-driven fact-checking tools that can operationalize preceding components effectively.

4 ALETHEIA

We have developed a prototype, *Aletheia*, to encapsulate these six components in our framework. *Aletheia* is an interactive system with an LLM-based backend and a web-based frontend interface.

4.1 Backend LLM-based Pipeline

A pivotal component in *Aletheia*'s framework is extracting data claims and retrieving the pertinent data evidence (C1–C3). This process converts a data-rich article into a series of data fact specifications that can be readily employed for data retrieval.

4.1.1 Design Implications for Prompt Pipeline. We conducted a qualitative content analysis of real-world data claims to guide us in defining subtasks for our prompt design. We examined sixteen real-world data articles from various topics and sources, manually extracting 108 data claims for thematic analysis. This step involved

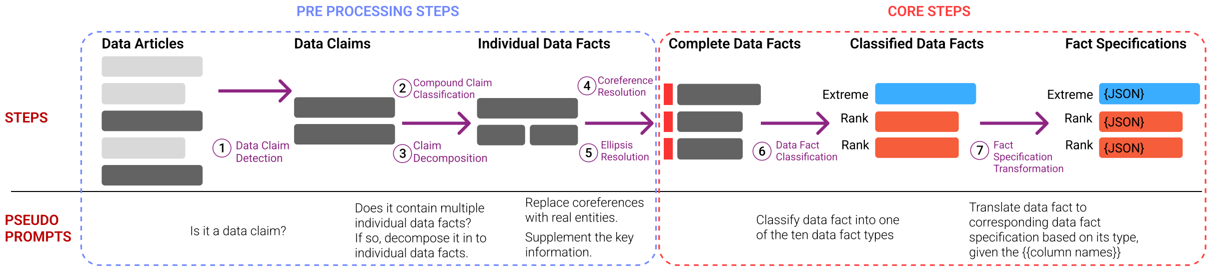

PRE PROCESSING STEPS
CORE STEPS
Data Articles
Data Claims
Individual Data Facts
Complete Data Facts
Classified Data Facts
Fact Specifications
STEPS
1 Data Claim Detection
2 Compound Claim Classification
3 Claim Decomposition
4 Coreference Resolution
5 Ellipsis Resolution
6 Data Fact Classification
7 Fact Specification Transformation
Extreme
Rank
Rank
Extreme {JSON}
Rank {JSON}
Rank {JSON}
PSEUDO PROMPTS
Is it a data claim?
Does it contain multiple individual data facts? If so, decompose it in to individual data facts.
Replace coreferences with real entities.
Supplement the key information.
Classify data fact into one of the ten data fact types
Translate data fact to corresponding data fact specification based on its type, given the {{column names}}

Figure 2: Overview of our LLM-based pipeline, which takes in a data article and outputs JSON specifications used to retrieve data evidence for each data claim. Seven steps are chained. The first five steps form the “pre-processing phase”, which transforms the input text first into individual data claims (S1) with compound claim identification (S2), then into distinct data facts (S3) with coreference resolution (S4) and ellipsis resolution (S5). These data facts are further processed in the last two steps, the “core steps” of our prompt pipeline, which classify the types of data facts (S6) before converting them into data fact specifications used for retrieving the pertinent data evidence (S7).

identifying key themes, refining initial codes, and engaging in iterative discussions to agree on the final codes and themes. The data claim corpus and the established codes are accessible in the supplemental materials. Our analysis yielded four main findings, each leads to an implication for our prompt pipeline design:

Findings	Implications
F1. Data claims often include more than one data fact.	I1. Decompose compound data claims into multiple independent, individual data facts.
F2. Inconsistent/ambiguous terms may be used to describe attributes.	I2. Provide the reference dataset’s attribute list to improve identification accuracy.
F3. The contextual information determining the subspace may be omitted.	I3. Infer and supplement key information based on context to determine subspace.
F4. Coreferences are generally used to represent real-world entities.	I4. Replace coreferences with actual entities based on context.

Table 1: Findings and implications for prompt design

Summary: The complexity and ambiguity inherent in real-world data claims require that they be decomposed into simpler, more verifiable facts, and disambiguated through more precise referencing and scoping. For example, to verify the claim, “*Prices tumbled 1.1% year-on-year, logging their first annual decline since June 2020*” from a news article [98], the process involves ***I1*** to dissect the claim into two distinct data facts (i.e., *‘tumbled 1.1%’* and *‘first annual decline’*); ***I2*** and ***I3*** to specify what ‘prices’ refer to, in this case, ‘housing prices,’ and to define their geographical and temporal scope; and ***I4*** to ensure that ‘their’ refers to ‘housing prices.’

4.1.2 Developing an LLM-based Pipeline. Our content analysis indicates that fact-checking data claims require several NLP tasks. Training models from scratch for each task requires immense amounts of data, computational resources, and time. Consequently, we employed a pre-trained LLM (i.e., GPT-3.5) to address this multifaceted NLP challenge because it permits flexible pipeline assembly through iterative prompt engineering [15, 62, 99]. Such flexibility aligns well with our primary objective: not necessarily to attain peak performance but to investigate the viability of a fully–automated fact-checking pipeline, formulate logical, coherent steps, and garner insights that can inform subsequent end-to-end optimization.

Pipeline Components. Our proposed NLP pipeline consists of seven chained steps. Due to the uncertainty in both our pipeline structure and the LLM’s ability to handle each task during our initial trial-and-error phase [23], we adopted the notion of LLM-chaining proposed by Wu et al. [100]. Rather than burdening the LLM with the end task of generating data fact specifications directly from data articles, we split this complex task into smaller, more tractable sub-tasks. This approach enabled us to build the pipeline with greater control and transparency of its intermediate steps. We also use NLTK [63] for sentence tokenization pre-processing. Figure 2 illustrates the pipeline and briefly explains each step within it.

S1. Data Claim Detection. This step corresponds to C1 in our framework. We use GPT-3.5 to classify tokenized sentences as either a ‘data claim’ or not, guided by our task description and examples.

S2 & S3. Compound Claim Classification and Decomposition. These two steps address the potential presence of compound claims (F1). We first ask the model to distinguish if a data claim is a ‘single’ or ‘compound’ claim (S2). The compound claims are subsequently decomposed into distinct data facts while the single claims are kept intact, ultimately producing a list of decomposed data fact dictionaries (S3). These dictionaries contain both the original sentence strings and the decomposed data fact strings.

S4 & S5. Coreference and Ellipsis Resolution. Addressing F3&4, for each data fact string, we instruct GPT to conduct coreference resolution (i.e., replacing pronouns with entities in the dataset) and ellipsis resolution (identifying and restoring omitted information, e.g., year=2023, based on the context of the input data document.

S6. Data Fact Classification. This step classifies each single claim as one of the ten data fact types. For the prompt, we provide a task description for classification along with our definitions of each data fact type, supplementing definitions with specific examples.

S7. Fact Specification Transformation. Next, we instruct GPT to convert the data fact strings into a type-specific JSON specification (Table 2). The JSON specifications are designed to capture the necessary key-value pairs for fact-checking claims that belong to the matching type. In this step, the attribute names of the reference

Fact Types	Claim Examples	JSON specifications
Value	The average IMDB score for horror movies released in 2020 is 6.7.	{"measure": “IMDB score”, "value": 6.7, "aggregation": “average”, "subspace": [{"genre"="horror"},{"year"=2020}], "identifier_key": “movies”}
Proportion	In 2013, Christopher Nolan's films comprised 34.8% of the total gross for movies with an IMDb score over 7.	{"measure": “gross”, "value": “34.8%”, "focus": [{"director" = "Christopher Nolan"}], "subspace": [{"year" = 2013}, {"IMDb_score" > 7}], "identifier_key": “movies”}
Trend	From March 2020 to March 2021, the number of COVID-19 cases in the US showed an increase.	{"measure": “case”, "value": “increase”, "subspace": [{"date" >= "March 2020"}, {"date" <= "March 2021"}, {"country" = "US"}]}
Extreme	Glenlivet 18 has the highest rating among whiskies originating from Scotland.	{"measure": “rating”, "value": “max”, "focus": [{"brand" = "Glenlivet 18"}], "subspace": [{"origin" = "Scotland"}], "identifier_key": “whiskies”}
Rank	Among players in point guards position who played more than 60 games in 2023, Trae Young is ranked 4th in three-point attempts.	{"measure": “3PA”, "value": 4, "focus": [{"player"="Trae Young"}], "subspace": [{"position" = "PG"}, {"games_played" > 60}, {"year" = 2023}], "identifier_key": “players”}
Association	There's a positive correlation between a movie's budget and its gross earnings.	{"measure_x": “budget”, "measure_y": “gross”, "value": “positive”, "identifier_key": “movies”}
Difference	During the 2019 NBA season, James Harden outscored Stephen Curry by 6.1 points.	{"measure": “points”, "value": 6.1, "focus_x": {"player" = "James Harden"}, "focus_y": {"player" = "Stephen Curry"}, "subspace": [{"season" = "2019"}]}
Categorization	There are seven movies that have an IMDb score over 9 and a gross of more than 300 million.	{"value": 7, "subspace": [{"IMDb_score" > 9}, {"gross">"300,000,000"}], "identifier_key": “movies”}
Distribution	The acceptance rates of colleges follow a right-skew distribution.	{"measure": “acceptance rates”, "value": “right-skew distribution”, "identifier_key": “colleges”}
Outlier	The movie “Oppenheimer” has a gross that's quite the outlier among historical biopic.	{"measure": “gross”, "focus": {"movie = "Oppenheimer"}, "subspace": ["genre" = "historical biopic"], "identifier_key": “movies”}

Table 2: Examples of the 10 data fact types along with the corresponding JSON output from our LLM pipeline.

dataset are provided in the prompt as contextual information to improve *Aletheia*'s ability to accurately parse the semantics.

4.1.3 *Veracity Determination.*

Utilizing the JSON specifications derived from **S7**, *Aletheia* retrieves relevant data evidence from the provided reference dataset and computes precise values based on the specified data aggregation and operations. Specifically, for objective value-based fact types, including *value, rank, proportion, extreme, difference*, and *categorization*, *Aletheia* directly compares the *claimed value* to the *actual value*. For *trend*, *Aletheia* is restricted to comparing only the two end values within a given timeframe. For other fact types (i.e., *outlier, association, distribution*), *Aletheia* employs established mathematical calculations and rules to evaluate veracity. For *outlier* detection, the interquartile range is applied to identify univariate outliers, while covariance matrix is used to detect bivariate outliers. The Pearson correlation coefficient is used to assess *association*, and the skewness formula is applied to determine whether a *distribution* is left- or right-skewed.

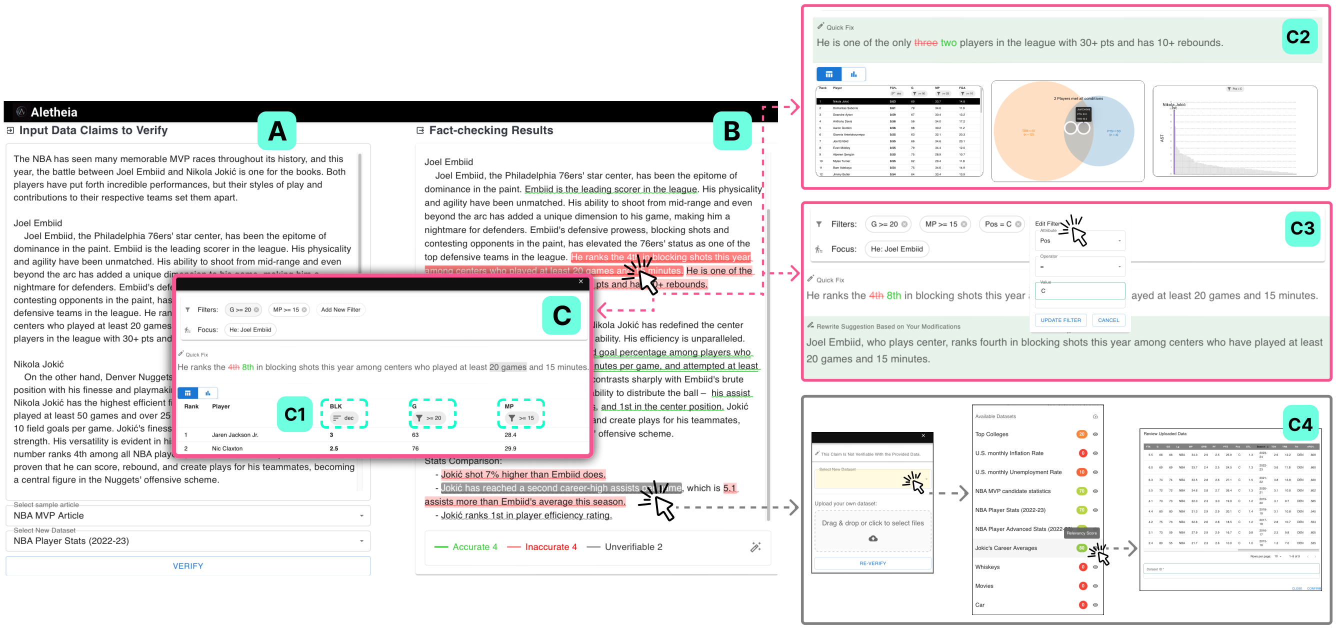
4.2 Designing *Aletheia*'s Interface

Once the veracity is determined and the supporting data evidence is obtained, the results must be effectively communicated to the audience, substantiating the verdict with interpretable evidence and explanations, and empowering practitioners to act upon these insights. This section addresses this challenge by designing visual representations for demonstrating data evidence (C5) and interactions to support actions from practitioners (C6).

4.2.1 *Designing Data Evidence Representations.*

Existing research lacks comprehensive guidance on effectively presenting data evidence. Thus, our main objective is to explore the design space for representing data evidence. We specifically focus on two common approaches: data tables and visualization charts. Data tables, as a conventional and widely accessible form of data representation in the data fact-checking workflow, serve as a baseline. Visualization charts, known for their capacity to handle larger datasets, improve readability [45], and enhance human cognition through rapid perceptual inference and pattern recognition [26], may offer a more efficient means of presenting data evidence. We propose designs for data table and visualization chart representations tailored to 13 subcategories derived from 10 data fact types: *value (mean), value (median), value (sum), proportion, trend, extreme, rank, association, difference, categorization, distribution, 1-D outlier*, and *2-D outlier.*

To develop these designs for data fact-checking, we identified the following three design goals. Our designs are also informed by literature on foundational visualization values [26] and strategies counter cognitive biases (e.g. [39]). We engaged in an iterative design process with active participation from two authors and feedback from a senior visualization researcher. Details of our design choices and the corresponding illustrations for both the data tables and the visualizations are available in the Appendix A. Further



Input Data Claims to Verify

The NBA has seen many memorable MVP races throughout its history, and this year, the battle between Joel Embiid and Nikola Jokić is one for the books. Both players have put forth incredible performances, but their styles of play and contributions to their respective teams set them apart.

Joel Embiid
Joel Embiid, the Philadelphia 76ers' star center, has been the epitome of dominance in the paint. Embiid is the leading scorer in the league. His physicality and agility have been unmatched. His ability to shoot from mid-range and even beyond the arc has added a unique di... nightmare for defenders. Embiid's defe... contesting opponents in the paint, has... defensive teams in the league. He ran... centers who played at least 20 games... players in the league with 30+ pts and...

Nikola Jokić
On the other hand, Denver Nuggets... position with his finesse and playmaki... Nikola Jokić has the highest efficient fi... played at least 50 games and over 25... 10 field goals per game. Jokić's finess... strength. His versatility is evident in hi... number ranks 4th among all NBA play... proven that he can score, rebound, and create plays for his teammates, becoming a central figure in the Nuggets' offensive scheme.

Select sample article: NBA MVP Article

Select New Dataset: NBA Player Stats (2022-23)

VERIFY

Fact-checking Results

Joel Embiid
Joel Embiid, the Philadelphia 76ers' star center, has been the epitome of dominance in the paint. Embiid is the leading scorer in the league. His physicality and agility have been unmatched. His ability to shoot from mid-range and even beyond the arc has added a unique dimension to his game, making him a nightmare for defenders. Embiid's defensive prowess, blocking shots and contesting opponents in the paint, has elevated the 76ers' status as one of the top defensive teams in the league. He ranks the 4th in blocking shots this year among centers who played at least 20 games and 15 minutes. He is one of the ... pts and ha... 0+ rebounds.

...ikola Jokić has redefined the center ... ability. His efficiency is unparalleled. ...d goal percentage among players who ...nutes per game, and attempted at least ...contrasts sharply with Embiid's brute ...bility to distribute the ball – his assist ..., and 1st in the center position. Jokić ...and create plays for his teammates, ...' offensive scheme.

Stats Comparison:
- Jokić shot 7% higher than Embiid does.
- Jokić has reached a second career-high assists per game, which is 5.1 assists more than Embiid's average this season.
- Jokić ranks 1st in player efficiency rating.

Accurate 4 Inaccurate 4 Unverifiable 2

Filters: G >= 20 MP >= 15 Add New Filter
Focus: He: Joel Embiid

Quick Fix
He ranks the ~~4th~~ 8th in blocking shots this year among centers who played at least 20 games and 15 minutes.

Rank	Player	BLK dec	G >= 20	MP >= 15
1	Jaren Jackson Jr.	3	63	28.4
2	Nic Claxton	2.5	76	29.9

Quick Fix
He is one of the only ~~three~~ two players in the league with 30+ pts and has 10+ rebounds.

Rank	Player	FG% dec	G >= 50	MP >= 25	FGA >= 10
1	Nikola Jokić	0.63	69	33.7	14.8
2	Domantas Sabonis	0.61	79	34.6	11.9
3	Deandre Ayton	0.59	67	30.4	13.2
4	Anthony Davis	0.56	56	34.0	17.2
5	Aaron Gordon	0.56	68	30.2	11.2
6	Giannis Antetokounmpo	0.55	63	32.1	20.3
7	Joel Embiid	0.55	66	34.6	20.1
8	Evan Mobley	0.55	79	34.4	12.0
9	Alperen Şengün	0.55	75	28.9	10.7
10	Myles Turner	0.55	62	29.4	11.8
11	Bam Adebayo	0.54	75	34.6	14.9
12	Jimmy Butler	0.54	64	33.4	13.9

Filters: G >= 20 MP >= 15 Pos = C
Focus: He: Joel Embiid

Edit Filter
Attribute: Pos
Operator: =
Value: C
UPDATE FILTER CANCEL

Quick Fix
He ranks the ~~4th~~ 8th in blocking shots this year ... ayed at least 20 games and 15 minutes.

Rewrite Suggestion Based on Your Modifications
Joel Embiid, who plays center, ranks fourth in blocking shots this year among centers who have played at least 20 games and 15 minutes.

This Claim Is Not Verifiable With the Provided Data.
Select New Dataset
Upload your own dataset:
Drag & drop or click to select files
RE-VERIFY

Available Datasets

Top Colleges	20	
U.S. monthly Inflation Rate	0	
U.S. monthly Unemployment Rate	10	
NBA MVP candidate statistics	70	
NBA Player Stats (2022-23)	70	
NBA Player Advanced Stats (2022-...		Relevancy Score
Jokic's Career Averages	80	
Whiskeys	0	
Movies	0	
Car	0	

Review Uploaded Data

FTA	G	GS	Lg	MP	ORB	PF	PTS	Pos	STL	Season	TOV	TRB	Tm	eFG%
5.5	66	66	NBA	34.3	2.9	2.5	25.9	C	1.3	2023-24	2.9	12.2	DEN	.609
6.0	69	69	NBA	33.7	2.4	2.5	24.5	C	1.3	2022-23	3.6	11.8	DEN	.660
6.3	74	74	NBA	33.5	2.8	2.6	27.1	C	1.5	2021-22	3.8	13.8	DEN	.620
5.5	72	72	NBA	34.6	2.8	2.7	26.4	C	1.3	2020-21	3.1	10.8	DEN	.602
4.1	73	73	NBA	32.0	2.3	3.0	19.9	C	1.2	2019-20	3.1	9.7	DEN	.565
4.4	80	80	NBA	31.3	2.9	2.9	20.1	C	1.4	2018-19	3.1	10.8	DEN	.545
4.2	75	73	NBA	32.6	2.6	2.8	18.5	C	1.2	2017-18	2.8	10.7	DEN	.554
3.1	73	59	NBA	27.9	2.9	2.9	16.7	C	0.8	2016-17	2.3	9.8	DEN	.605
2.4	80	55	NBA	21.7	2.3	2.6	10.0	C	1.0	2015-16	1.3	7.0	DEN	.535

Rows per page: 10 1-9 of 9
Dataset ID *
CLOSE CONFIRM

Figure 3: *Aletheia*’s interface. Users enter textual content and select/upload a reference dataset in *Input View (A)*. The backend then detects data claims, retrieves corresponding data evidence, and verifies them. The fact-checking results are presented in *Result View (B)*, utilizing color codings to signify their verdicts: *accurate*, *inaccurate*, and *unverifiable*. Users click on the highlighted data claims to access the *Evidence View (C)*. This view contains the designed data evidence presentation and interactions.

design details and enlarged versions are available in the supplementary materials.

DG1. Ensure alignment with the original data insights conveyed by the data claim. Each data fact type inherently provides a unique insight. For example, *Rank* (Figure 8 V7) highlights an entity’s position within a group. Effective verification of the accuracy necessitates that the evidence representation not only corresponds with the insight (e.g., the particular rank, *‘8th’*) but also encapsulates the scope of the relevant data (e.g., a sorted list that includes the rank). Similarly, a *Categorization* fact (Figure 8 V10) indicates an entity’s affiliation with a subset of entities. Thus, the design should display the entity, the intersection, and the inclusion criteria. The variety of insight types emphasizes the need for customized representations tailored to specific data fact types or subtypes.

DG2. Streamline viewers’ analysis of data evidence, facilitating quicker judgments. Efficiency is key in designing data evidence presentations for fact-checking, as it could enhance professionals’ productivity and increase general readers’ adoption. To achieve this goal, we embraced three key design approaches. First, we display only the relevant data segments, eliminating the need to sift through the full dataset. Second, we highlight and annotate salient elements, especially data points mentioned in the claim. Third, for claims involving derived statistics, we automate data operations and computation to directly show the summary statistics.

DG3. Bolster viewers’ confidence in their assessments. Another critical element we emphasize is viewers’ confidence in their judgment. Given that both the verdict and the representation provided are rooted in the same data evidence, we posit that a representation that boosts viewers’ confidence can also enhance their trust in the system’s ability to retrieve accurate evidence and process it appropriately. We first ensure the transparency of data operations either by incorporating *operation widgets* (Figure 3 C1) or directly through visual encodings. Next, we emphasize the presentation of individual data points, aiming to provide an overview of the data distribution that allows for a ‘sanity check.’

4.2.2 *Interface Design.*

In addition to designing evidence representations, we explore the feasibility of integrating them into an interactive application along with our proposed fact-checking framework. We envision a scenario where authors must verify and rectify inaccurate data claims, and thus propose three key design requirements:

DR1. Facilitate rapid reviewing and correction of erroneous claims. Upon receiving fact-checking results, authors naturally want to review the verdicts and supporting data evidence. This review allows them to decide whether to trust the results and take action, e.g., to correct inaccuracies in the data claim. Thus, an interactive system should streamline the review and revision process.

DR2. Enable user intervention for AI mistakes. AI mistakes may occur when GPT incorrectly infers the data subspace or coreferences (S4&5), especially when keywords are missing or ambiguous in the data claim. Also, GPT may incorrectly parse a focused attribute (S7). Under these circumstances, it is critical to incorporate human supervision and intervention for enhanced reliability.

DR3. Support integration of additional reference data. Real-world articles can simultaneously rely on multiple datasets with varying contexts, thereby posing a challenge when attempting to fully automate the fact-checking process. Enabling users to add additional reference datasets to evaluate claims that are initially unverifiable could enhance *Aletheia*'s utility. Furthermore, considering users' potential unfamiliarity with the 'suitability ' [91] of reference data, the system should provide support for users to assess whether a dataset is appropriate for verifying certain data claims.

4.3 *Aletheia* Workflow and Usage Scenario

We integrated our LLM-based pipeline with the interactive user interface, resulting in a functional prototype, *Aletheia*. The backend of *Aletheia* is built on Python Flask, leveraging OpenAI (GPT-4) for its NLP tasks. The frontend is developed using React, with the visualizations implemented in D3.js. As shown in Figure 3, *Aletheia*'s interface has three main views: an input view (A), a result view (B), and an evidence view (C). To illustrate the utility of *Aletheia*, consider the scenario of a sports editor, Jordan, tasked with reviewing the accuracy of a written article containing various data claims.

Jordan first loads the draft article of NBA MVPs into *Aletheia*'s *Input View*, and selects a reference dataset of players' average statistics from *Basketball-Reference*, an authoritative sports data platform. Upon requesting *Aletheia* to verify the claims, the *Result View* generates a fact-checking report with the data claims color-coded. Jordan navigates through these data claims and toggles between *table* and *visualization* forms to examine the data evidence. This interaction helps him further assess the verification results and determine if he should apply a quick correction suggested by *Aletheia*. For instance, as shown in Figure 3 C2, *Aletheia* recommends correcting a *ranking value* error from *'4th'* to *'8th'* based on the computed results. *Aletheia* also supports a 'quick rectify' action, which can be useful when many instances of text-data misalignment occur (e.g., during data updates). These system functions support **DR1**.

To mitigate the risks associated with inferential mistakes (**DR2**), *Aletheia* provides an interactive widget (Figure 3 C3). This widget displays the key AI inferences, with filters for tuning the subspace, coreferences, and focused attributes. Jordan hovers through these 'chips' to examine associated text segments, and identifies that the AI overlooked a filter (Position=Center), which should have been applied to the phrase 'among all centers' in the claim. *Aletheia* enables him to directly edit these 'chips' to override the AI's inferred elements, i.e., to add missing filters. This action prompts *Aletheia* to reassess the associated claim based on Jordan's modifications and simultaneously propose a text revision reflecting these adjustments.

After addressing the discrepancies flagged in red, Jordan encounters two claims marked as 'unverifiable' (Figure 3 C4). To resolve these, he imports new datasets for targeted evaluation. These extra datasets are bound to individual claims and do not impact other verified claims. Employing *Aletheia*'s *data relevance evaluation* function, which leverages the GPT model to assign a 'suitability score' based on the attribute names and the claim, Jordan can quickly compare and identify the most pertinent datasets (**DR3**). This prompts *Aletheia* to reassess the accuracy of the targeted claim.

5 PROMPTING FRAMEWORK EVALUATION

In this section, we evaluate the feasibility of *Aletheia*'s LLM-based pipeline, with a particular focus on two core steps: **data fact classification** (*S6*) and **data facts specification transformation** (*S7*) in Figure 2. We focus on these core steps for three reasons. First, conducting a comprehensive evaluation requires viable testing datasets of data documents with corresponding reference datasets, which are currently unavailable and expensive to curate. Second, our initial experiments with real-world data articles (e.g., [58, 64]) indicate that GPT is competent in identifying data claims from passages/articles; our unoptimized prompt achieved an accuracy of 87.2% and 93.1%, respectively (see the supplementary materials). Third, although we utilize GPT to perform the pre-processing steps, the downstream tasks (e.g., coreference, ellipsis resolution) have been extensively explored in NLP research, with pre-trained statistical models showing increasing capabilities [8, 44].

5.1 Data Curation

Existing benchmarking datasets for automated fact-checking (e.g., FEVER [89], LIAR [96], MultiFC [12], ClaimBuster [9], etc.) primarily focus on text-sourced (i.e., knowledge-based) claims. For example, LIAR [96], based on human-labeled shorts claims from PolitiFact [71], includes statements like "*Newly elected Republican senators sign pledge to eliminate food stamp program in 2015.*" These text-sourced datasets do not align with the focus of this work, and there is a notable absence of open-sourced benchmarking datasets tailored for data claims. In the limited body of work specifically addressing data/statistical claims (e.g., [47, 75]), training/testing datasets are often synthesized with templates due to the scarcity of benchmarking datasets. In this work, we employ a similar template-based data curation approach, focusing more on diverse types of data insights [80, 83, 97] to cover a range of insight categories.

We programmatically curated ground truth claims for 10 aggregated data fact types, generating 40 template-based claims per type along with their corresponding data fact JSON specifications, following Table 2. To better represent the language variation in real-world claims, we employed GPT to produce a paraphrased version of each claim. We compiled a dataset of 400 test claims, each featuring data fact specifications, a claim generated from type-specific templates, and a paraphrased version. We manually reviewed the paraphrased claims to ensure they preserved the original data facts.

5.2 Evaluation Results

Data Fact Classification. We tested our data fact type classifier on a balanced dataset of 400 paraphrased examples using GPT-4. GPT-4 achieved perfect classification against natural language variation. The results indicate that ***an LLM can robustly classify data facts into types following the data fact taxonomy,*** simplifying previous approaches that required multiple algorithms [94].

Fact Specification Transformation. Next, we test our fact specification transformation step, taking the preprocessed, GPT-paraphrased claims as input and outputting a JSON specification for each claim. We present matching accuracy for each data fact type in Figure 4, with the green area representing complete matches (i.e., the generated JSON *exactly* matches the ground truth) and the red area the partial matches (i.e., *parts* of the generated JSON match those in the ground truth). There are no cases in which the ground truth is entirely missed, and the ratio of partial matches is also represented below each red box.

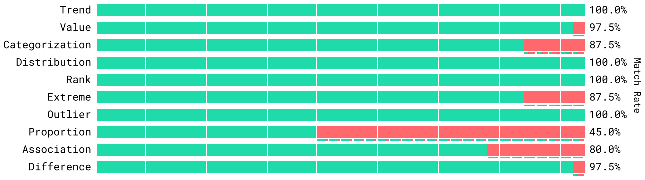


Figure 4: Success rate of data fact specification transformation. Each row corresponds to a distinct data fact type. Boxes within the rows represent individual examples. Green boxes indicate successful transformations, where all transformed attributes and values match the ground truth. Red boxes represent examples with incomplete or incorrect conversions. The small rectangles below the red boxes represent partial match performance with the same color code. The average rate of complete matches is 89.5%.

Across all data fact types, the average rate of complete matches is 89.5% (σ = 16.27), which means that nearly 90% of the generated JSONs are fully usable for data evidence retrieval. For comparison, state-of-the-art LLMs for code generation, when given a single trial, produce executable code around 20-40% of the time (see "Pass Rate" or "Pass@1" in [19, 59]), showing how much more challenging the generation of executable representations is when compared to classification tasks such as in Section 5.2. We consider the current average rate of complete matches (89.5%) positive evidence that ***an LLM can be used to generate well-formed data fact specifications***, allowing for automated systems such as *Aletheia*.

5.3 Failure Case Analysis

Nevertheless, there is more room for improvement in fact specification transformation: the fact type "Proportion" has the lowest rates (45%) of complete matches. The primary cause of failure can be traced back to the parsing of data filters — achieving a complete match in the subspace requires identifying all applied filters. Consider a *Proportion* data claim: "*In 2013, 75.59% of the budget for Italian movies was spent on movies with an IMDB score higher than 6.*" A full match in parsing requires the detection of a focus space filter *(IMDB_score ≥ 6)* and two subspace filters *(released_year = 2013 and country = Italy)*. This complexity explains why *Proportion* type has the lowest *complete matching* score — it requires not only the identification of all filters but also a clear distinction between *focus_space* and *subspace.* This bottleneck could justify adding new steps to our pipeline to better perform this task for this specific set of fact types. It is important to note that data claims can become more complex regarding fact type, semantics, and compoundness, posing greater challenges. We further discuss *Aletheia*'s limitations and future opportunities in Section 9.

6 USER EVALUATION OF DATA EVIDENCE REPRESENTATIONS (CHARTS & TABLES)

We conducted a mixed-method user study with 20 participants, using our 26 data evidence representations (Subsection 4.2) as probes to gather quantitative behavioral insights and qualitative feedback.

Participants Demographics. We recruited a total of 20 participants from our institution for the study, comprising 12 males (60%) and 8 females (40%). The majority, 18 participants, are aged between 25 and 34, with 2 aged 18 to 24. All are either graduate degree holders or candidates, with diverse backgrounds in statistics, data analysis, data visualization, and varying experiences with data articles.

Data Curation. To maintain a consistent and reliable study environment, we chose a manually curated, predetermined test dataset over real-time API calls to GPT, due to the potential for the model's variable runtime to disrupt controlled conditions. We curated four data claims — two accurate and two inaccurate ones — for each of the 13 data fact types. To reduce participant workload, we split the 52 test data claims into two datasets. Each dataset contains an accurate and an inaccurate claim for every fact type, resulting in 26 tasks per participant per study phase (2*13).

User Study Interface. Given the interactive features of our data evidence representations and the need to monitor assessment duration, we devised a study-tailored interface (Figure 9) based on *Aletheia*. This study interface consists of five components: a tutorial page (C), two study session pages (A & B), a page for visualizing participants' results (D), and an "Exploration Page (E)" that helps participants review the representations and answer interview questions.

6.1 User Study Procedure

Our study consists of three phases. *Phase I* and *Phase II* focus on collecting quantitative data on *assessment time* (tracking), *confidence shift* (self-reporting), and *preference* (self-reporting). *Phase III* is a post-study interview. A detailed interview protocol is available in the supplemental material. The entire study process was video-recorded, and the audio was transcribed for qualitative analysis.

Phase I. Participants were randomly given one of the two datasets in a counter-balanced manner and assigned to either the *Table Group* (Group A) or the *Visualization Group* (Group B). Each group used the respective data evidence representations to review the data claims. Prior to assessing the claims, participants underwent a five-minute tutorial session to familiarize themselves with the data encodings for various data fact types. We also showcased the interface for Phase I using a demo dataset.

During the main part of *Phase I* (shown in Figure 9 (A)), participants were sequentially presented with individual claims. They were instructed to read the claim carefully and then click the "Verify" button to view the model's verdict. The system then fetched the corresponding data evidence and displayed a verdict stating "*The AI determines this claim to be accurate/inaccurate.*" Participants were informed beforehand that the *"AI prediction"* might be inaccurate and should only be considered as a reference. After viewing the verdict, participants were instructed to assess the veracity of the claim based on the data evidence revealed after clicking the "Show Data Evidence" button. A timer started upon the appearance of the visual representations. After reviewing the data evidence, participants indicated their decision by selecting either "Accurate" or "Inaccurate", which simultaneously stopped the timer. Participants also self-reported their confidence level on a scale from 1 to 5.

Phase II. At the beginning of *Phase II*, participants spent five minutes familiarizing themselves with the counterpart representations (i.e., participants in the Table Group were presented with visualizations in this phase and vice versa). We again showcased the interface for Phase II using a demo dataset. Participants performed the same fact-checking task as in Phase I, but with counterpart representations. After making a judgment, participants' confidence level from Phase I was disclosed, along with the corresponding Phase I representations (as shown in Figure 9 (B)). Participants were instructed to compare their confidence with their confidence Phase I ratings. Additionally, we asked participants to self-report their preference between the table and visualization representation for fact-checking the current claim using a five-point scale.

Phase III. Concluding the study, we presented a summarized visualization illustrating the participant's confidence shift and preferences between the two study parts (Figure 9 (D)), followed by a semi-structured interview to gather insights about their thinking process, perceived advantages/disadvantages, and feedback on both representation methods. During the interview, participants could use the "Exploration Page" (Figure 9 (E)) to navigate the reviewed claims and the two corresponding visual representations while answering questions.

6.2 Quantitative results

We report four measurements: **(A)** ***time-spent*** assessing the data evidence in *Phase I*, **(B)** ***confidence shift*** between the data evidence representation forms in *Phase I* and *Phase II*, **(C)** ***preference*** and **(D)** ***accuracy*** between the two data evidence representations. The results are shown in Figure 5.

Time-spent. We applied the Mann-Whitney U test at a significance level of 0.05 to determine whether there existed any statistically significant differences between the *table* and *vis* representation types across various data fact types. In general, participants who used visualization charts to assess the claim spent less time ($M = 7.9, SD = 6.1$) than those who used data tables ($M = 15.0, SD = 10.7$) in *Phase I*. This time efficiency was consistent across all fact types, with statistical significance identified in the majority of fact types (8 out of 13), including *Distribution* ($U = 370.0, p < 0.001$), *Trend* ($U = 326.5, p < 0.001$), *Rank* ($U = 287.5, p < 0.03$), *Association* ($U = 375.0, p < 0.001$), *Outlier (Univariate)* ($U = 362.0, p < 0.001$), *Outlier (Bi-variant)* ($U = 373.5, p < 0.001$), *Value (Median)* ($U = 289.5, p < 0.03$), and *Extreme* ($U = 276.0, p < 0.05$). No significant differences were observed in the rest of the data fact types. More detailed mean, U-value, and p-values for each data fact type under two representations are reported in Figure 5, with the statistically significant conditions marked in green.

Average confidence shift. Confidence shift is quantified using a five-point scale: -2 (much more confident with table), -1 (more confident with table), 0 (about the same), 1 (more confident with visualization), and 2 (much more confident with visualization). Positive values indicate higher confidence on average with visualizations for fact-checking, whereas negative values suggest the opposite. In general, *vis* has slight advantage over *table* in enhancing user confidence ($M = 0.56, SD = 0.87$). While this advantage is consistently observed across all fact types, it is most pronounced ($M > 1$) for *Association* ($M = 1.48, SD = 0.74$), *Trend* ($M = 1.33, SD = 0.66$), and *Outlier (Bi-variant)* ($M = 1.28, SD = 0.85$). On the contrary, the least advantage was observed in types including *Value (Sum)*, *Proportion*, *Categorization*, *Difference*, *Rank*, *Extreme*, and *Value (Mean)* with average confidence shifts less than 0.2.

Preference. Preference is measured using the same five-point scale: [-2 (strongly favor table), -1 (favor table), 0 (neutral), 1 (favor visualization), 2 (strongly favor visualization)]. Positive values indicate that participants prefer visualizations over tables for specific fact types, whereas negative values suggest a preference for tables. Overall, participants showed preference on *vis* over *table* ($M = 0.81, SD = 1.17$). This observation is consistent for the majority of the data fact types (11 out of 13), except *Proportion* ($M = -0.08, SD = 1.12$) and *Value (Sum)* ($M = -0.28, SD = 1.04$). Notably, participants exhibited the most pronounced preference on *vis* for fact types including *Association*, *Trend*, *Outlier (Univariate)*, *Outlier (Univariate)*, *Outlier (Bi-variant)* and *Distribution* with an average preference score over 1.7.

Accuracy. Both groups exhibit high accuracy when determining the veracity of the data statements. The overall accuracy (under time pressure) is 90.19% ($\sigma = 0.3$). The table group achieved 89.62% accuracy, while the visualization group achieved 90.77%. The accuracy was higher for objective value-based claims, including *Rank (100%), Proportion (100%), Difference (97.5%), Categorization (95%), Value(Mean) (97.5%), Value(Median) (95%), Value(Sum) (100%), Trend (87.5%), Extreme (100%)*, and lower for subjective ones, including *Outlier (Univariate) (67.5%), Outlier (Bivariate) (65%), Association (78%), Distribution (93.3%).* While the overall difference in detection accuracy is not significant, the visualization group achieved notably better accuracy in *Association* (88% vs. 68%) but lower accuracy in *Outlier* facts (60% vs. 72.5%). The lower detection accuracy in *Outlier* facts can be attributed to the subjective nature of visual interpretation and varying algorithms and thresholds used to determine outliers. Nevertheless, the overall high detection accuracy indicates that our designed representations can effectively assist users in identifying inaccurate claims under time pressure.

6.3 Findings and Takeaways

T1. Visualizations inherently offer advantages when fact-checking data claims related to patterns or distributions across numerous data points. Visualizations exhibit pronounced advantages over tables when verifying *association*, *distribution*, *outlier*, and *trend* data fact types, which is evident across all measurements: time-spent (Figure 5A), participants' confidence (Figure 5B) and preference (Figure 5C). Participants commonly expressed that visualizations are significantly more helpful for determining the veracity of data claims that necessitate an 'overview of the data.' Particularly, P12 stated that "*Visualizations is just a lot clearer than looking at a data table, especially if the table has a lot of rows you have to scroll through and process all of those information*". This

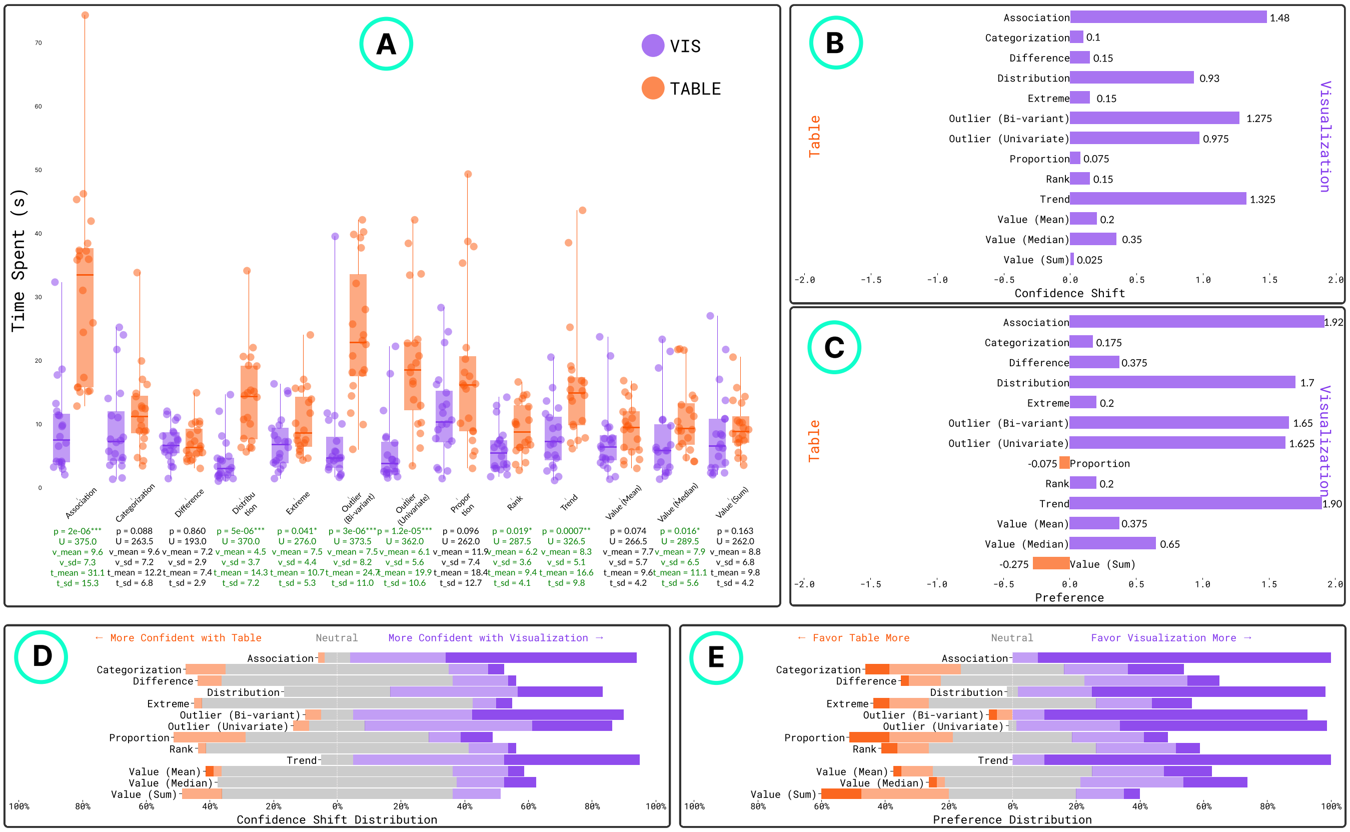


Figure 5: Quantitative results from our user study with 20 participants comparing visualizations and tables as different data evidence presentation forms. (A) The distribution (box) and individual (point) time taken to assess the accuracy of thirteen distinct data facts. The x-axis represents the data fact types, while the y-axis indicates the duration in seconds. The two diverging bar charts show the average shift in (B) the viewers’ confidence and (C) their preferences across the thirteen data fact types. Right-pointing bars signify that participants have greater confidence in their assessment when using the visualization, or they prefer to use visualizations for fact-checking the respective data facts. Conversely, left-pointing bars indicate greater confidence or preference for tables. Figures (D) and (E) display the percentage distribution for each response option regarding confidence shift (D) and preference (E). The length of the bars represents the percentage of each selection on the five-point scale. Gray bars represent *Neutral*. Orange bars represent *Table* while purple bars represent *Visualization*. Darker **red** and **purple** signify greater intensity (i.e., much more confident/strongly favor).

advantage can become more prominent for bigger datasets. Interestingly, *extreme*, *rank*, and *value (median)* also exhibited statistically significant time improvements in visualizations over tables. Given that statistics are readily available in the data table, verifying these claims only requires participants to process a single number. However, statistical significance was not found in *categorization*, *difference*, *proportion*, and *value (mean&sum)*, five data types with similar settings. We anticipated that the visualization representation of the ranked bar chart for these three data types, offering visual confirmation of accurate sorting, could potentially reduce the time needed for viewers to be persuaded.

T2. Displaying data operation widgets accelerates assessment and boosts confidence. We included operation widgets (Figure 3 C1) to indicate data operations in both *table* and *visualization* representations. 19 participants agreed that these widgets bolstered their confidence in the system retrieving the correct data evidence. Participants noted that their initial mental task upon seeing the data evidence was to align the keywords with the filtering widgets. P6 stated that “*it [showing widgets] is big, because I need to know, especially when a data claim is being made over a subset*”. P4 concurred that “*these [filter widgets] are the ones that really affect my confidence. . . it gives me an understanding of what subset of data the person was trying to analyze.*” Seeing these widgets also expedited their assessment, as it obviated the need to scrutinize individual values to confirm their presence in the subspace. In particular, P20 emphasized that “*[not showing filters] would impact visualization a lot more*” because “*[in table], I can quickly see it [the relevant information]*”. Though only occasionally checked the widgets, P10 recognized the importance of them:“*[knowing] what they present matters a lot, and I can take a look any time I want.*” However, there was one outlier (P8), who assumed the data evidence was correctly retrieved, therefore, examining the widgets “*made me spend a little bit more time and had no effect on my confidence.*” P2 also indicated that once they established trust in *Aletheia*'s ability to apply the correct filters, they paid less attention to the widgets.

T3. Unit representations enhance confidence, even when not scrutinized. During the interview phase, participants reflected on their attention to unit representations and how their absence might affect their confidence regarding fact types linked to specific summary statistics. While all participants agreed that they did not scrutinize the visual elements representing individual values, 15 out of 20 participants agreed that displaying summary statistics alone would reduce their confidence compared to pairing them with unit representations because the unit representations allow them to verify that the data distributions align with the aggregated statistics. For example, P2 stated that "*I want to see raw data to make sure that the thing that I'm consuming is accurate.*" P7 elaborated on verifying computed values based on the underlying raw data: "*This is one way of confirming that the average is correct… if the [average] bar is somewhere in the middle.*" P17 expressed that "*without any individuals, you don't have a global understanding about the data points distribution.*" Three participants (P1, P5, P18) emphasized the need for visualizations where the distribution of units offers greater value for a "sanity-check" of the provided statistics than mere "numbers" in the table.

T4. Contextual information can be both reassuring and distracting in customized visualization. For the data fact types requiring only single summary statistics to verify the claim, we provide contextual information using different visual techniques. Three particular visuals — *value (sum)*, *proportion*, and *categorization* — are particularly customized, but received low preference on average. A common reason cited by our participants during the interview phase was that these fact types solely require aggregated statistics to assess their veracity after confirming the subspace and focus, and it can take additional effort to process the additional contextual information. Participants also mentioned their struggle with unfamiliar chart types, leading to slower comprehension. For example, P10 explained that "*… it took me more time to understand the mapping… so I feel a little bit distracted when trying to extract useful, relevant information to do the fact-check.*" Participants also appreciated the assurance of contextual information provided. P10 particularly liked the proportional Venn diagram, "*it's just nicer… you get more information… for fact-checking, the context helps because it gives you an assurance that the data is valid and there's no arbitrary thing.*" P15 pointed out that when it comes to *sum values*, "*using visualization, there is no concern about the total [sum operation] because the height should be the [total] height of each one.*"

T5. Highlighting salient information streamlines the fact-checking process. All participants concurred that a crucial mental step in fact-checking data claims involves extracting salient information from both the text and data evidence and then verifying their alignment. Participants also acknowledged that our design decision to *highlight and annotate salient visual elements* aids in accelerating this process. Participants who preferred tables over visualizations for "one-number" fact-checking noted that tables provided them a clear location to focus on, typically the last sticky row we highlighted in the table. P5 expressed that"*I know where to expect to see it,*". However, it is not as consistent with visualizations, even though the visualizations included annotated labels with the same information. Participants in favor of the visualizations emphasized the intrinsic value of visualization as a form of abstract information. For example, P7 stated that the "*it [visualization] just abstracts away all the information that I don't need to know.*" P17 added that "*abstracted information that corresponds to the statement is way more efficient*" when it comes to fact-checking.

Design Recommendations for Presenting Data Evidence. Drawing on our quantitative results, qualitative findings, and reflections on design, we derive four general design recommendations:

(1) **Display data operations, especially the filters**: Regardless of the presentation format, our findings suggest the importance of consistently displaying the data operations executed to retrieve relevant data evidence and calculate aggregated statistics, particularly the filters used to determine the subspace (**T2**). Such transparency enhances viewers' trust in data validity.
(2) **Make salient information visually predominant**: After viewers establish trust in data validity, they search for the *salient* information crucial to determining veracity (observed in **T5**). This salient information typically includes elements like *value*, *measure*, and *focus* depending on the fact type. Making these elements visually predominant in the evidence presentation will streamline the verification process.
(3) **Key information first, contextual information on-demand**: Contextual information can enhance understanding and bolster trust, but an excess may overwhelm viewers since context is not essential for assessing fact-level veracity (as suggested in **T4**). Therefore, we recommend abstracting such information and enabling viewers to access it on demand, i.e., via interactions.
(4) **Enhance readability with visual aids**: While *Aletheia* offers computed statistics, comparing raw numbers becomes challenging when the values are large (e.g., movie grosses). As favored by participants (**T5**), we suggest incorporating visual aids in visualization, like an additional line indicating the "claimed value", to ease comparison. Visual aids (e.g., colored underlines) can also be utilized to highlight the mapping of entities, attributes, and filters between textual references in claims and visual counterparts in representations, facilitating more efficient "sanity check".

7 APPLICATION SCENARIOS FOR *ALETHEIA*

We design *Aletheia* specifically to assist authors and editors in the task of ensuring accuracy within data-rich articles. Given that many news organizations have already established their data infrastructure, inserting a system like *Aletheia* in their editorial workflow could markedly enhance their efficiency. Additionally, the methodologies underpinning *Aletheia* can be seamlessly integrated with data article authoring tools, such as CrossData [21], DataTales [85], reinforcing the accuracy of content produced. We envision *Aletheia*'s adaptability across diverse user groups, with potential applications ranging from browser extensions with reference data plugins for general news readers to fact-checking social media posts, financial reports and data-driven claims in other domains. Given these diverse applications, we propose the following two recommendations for tailoring *Aletheia* to specific domains or contexts:

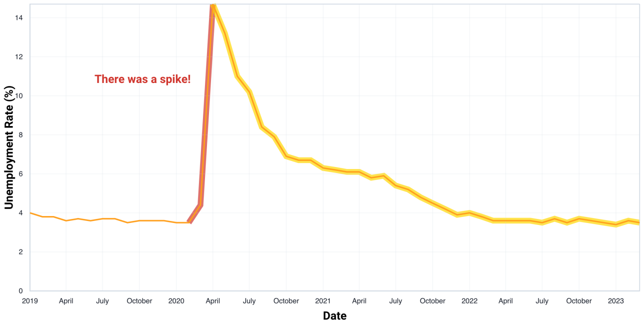


Figure 6: An example of showing contextual information. The data claim is “*The unemployment rate experienced a decrease between April 2020 and March 2023.*” While the claim is technically accurate, it might be suspected of “cherry-picking the timeframe”. Displaying the context can help inform audiences about the potential ‘pitfall.’

Leverage in-context examples in LLM prompts to enhance NLP task performance and accommodate domain-specific requirements. We provide in-context examples in prompts for multiple NLP tasks in *Aletheia*. The LLM’s flexibility enables easy customization of the pipeline to meet domain-specific requirements through mere modifications of the in-context examples and their associated definition and reasoning. Consider the case of a data fact not present in the data fact taxonomy, such as the *“prominent streak”* [42]. If this type of data claim is prevalent in a domain (e.g., sports), one can integrate its definition, examples, and desired output specifications in the chained prompts, thereby enabling LLM to retrieve the corresponding data evidence.

Tailor data evidence representations for specific user groups, associated tasks, and domain conventions. While our 13 pairs of type-specific data evidence representations cover common data fact types, they are not exhaustive. Additional data fact types will require customized evidence presentations. For instance, a sequential bar chart that highlights instances meeting (or not meeting) specific conditions would be appropriate for representing a *prominent streak* data fact. Furthermore, we recommend using our data evidence designs as a baseline, tailoring them to the needs of target users and domain characteristics/conventions while adhering to our general design guidance. For instance, in contexts where “cherry-picking data” [61, 94] is a concern, it is recommended to add visual elements that present more contextual information; for example, supplementing a line graph that depicts the *trend* with an *overview* (example shown in Figure 6) can provide added context to “cherry-picking timeframes”, potentially assisting audiences in better judgment about the original claim.

8 LIMITATIONS

Sole focus on fact-level assessment. While our method prioritizes verifying the underlying data facts, it is important to acknowledge the potential logical fallacies or misinterpretations that could undermine the plausibility of the resulting conclusions. Consider extending the example presented in Figure 6 to “*the unemployment rate experienced a decrease between April 2020 and March 2023, indicating a strong job market.*” Although the initial part of this data fact may be technically accurate, the concluding inference might not hold true. The claim omits the ‘spike’ in March 2020, coinciding with the COVID-19 pandemic lockdowns in many U.S. states.

Require manual selection of reference dataset. Another limitation of *Aletheia* is the need for users to manually select or upload suitable reference data for fact-checking. In real-world scenarios, however, data articles or claims may derive from various data sources. Furthermore, the users may have limited access to, or knowledge of, such datasets, which restricts their ability to select appropriate reference data. Although *Aletheia* currently incorporates features that facilitate this process (i.e., allowing for incorporating additional reference dataset and leveraging GPT to assess the suitability of a dataset for fact-checking), there is room for further enhancement through the adoption of more sophisticated tools. This includes developing more reliable mechanisms to evaluate the suitability of available datasets and automated methods to extract reference data from extensive data infrastructures and reconcile synergies and conflicts among multiple datasets.

Lack of comprehensive optimization and evaluation of the LLM-based pipeline. We acknowledge that there is potential for further optimization of our NLP pipeline (Figure 2) to enhance its effectiveness and undergo a more comprehensive evaluation. It’s noteworthy that the modular structure of our chained LLM-based pipeline offers significant flexibility in the choice of models for downstream tasks. Future work could explore substituting or comparing the LLM-based approach with more established NLP methods for specific steps (e.g., statistical models for coreferences and ellipses resolution [8, 44]) and experimenting with different sequences of steps.

User study designs. One study limitation pertains to the demographics of our study participants. While our participants exhibit diverse interactions with and trust in data articles/reports, they are uniformly graduate students who are likely more proficient with data compared to other potential user groups (e.g., fact-check professionals and general news readers). Therefore, we consider our quantitative results to be more relevant to ‘data-savvy’ individuals. Nevertheless, we acknowledge the need for future experiments with other user groups to further broaden the scope of our findings.

Limitation and risk in adopting LLM. The limitations and inherent risks associated with LLMs (e.g., hallucination, inconsistent accuracy) can impact automated fact-checking systems that rely on them. This is particularly notable for tasks like information search [79] and veracity prediction, especially when contextual information is scarce [74]. Our approach, in contrast, does not rely on LLM for direct determination of claims’ veracity. Instead, the verification is through computing the pertinent data retrieved by data fact specification. While inaccuracies may occur at other stages of the pipeline, such as retrieving an incorrect subset of data or misinterpreting focused attributes, *Aletheia* offers features that assist in identifying such errors and manually correcting them.

9 FUTURE RESEARCH

Assessing and Communicating Plausibility of Data Claims On Reasoning Level. Our first limitation highlights the need for developing automated/semi-automated technology that assesses and communicates the plausibility of data claims at both the *factual level* and the *reasoning level.* To achieve this, similarly, a series of NLP tasks are required to determine the linguistic relationship between *data facts* and corresponding *conclusions*, as well as retrieve the data evidence. A reasoning error taxonomy, similar to the visual error typology [61] recently proposed by Lisnic et al., is essential for determining the text-data reasoning error and subsequent evidence communication. Given the intricacy of reasoning errors compared to factual ones, we anticipate that visualization — with its capability to provide context and communicate uncertainty [40] — will assume a larger role. Future HCI/Visualization research can delve into and broaden the design space of data evidence presentation, addressing not just factual errors but also reasoning flaws. Further, evaluating data statements' reasoning validity often involves external contextual information, including facts/knowledge or additional quantitative datasets. Kim et al. recently developed an LLM-based interactive tool [51] to retrieve relevant data associated with data claims. Such tools can potentially be integrated with *Aletheia*, enabling a more comprehensive evaluation of data claims.

Fact-checking More Complex Data Claim. There is substantial room for further enhancing *Aletheia*'s capability to effectively handle more complex data claims. A data claim can become more complex when it 1) involves more complicated data operations (e.g., *multiple filters*), 2) encompasses nuanced semantics, and 3) is compound. Our evaluation (Section 5.2) reveals that parsing data filters is a primary cause of failures in text-to-data mapping. This poses a substantial challenge for handling data facts involving different filters, such as *proportion* type, where not only must all filters be identified but also be categorized (i.e., *focus* or *subspace*) correctly. We hypothesize that a constructive improvement might involve the introduction of a specific step dedicated to subspace information extraction. Nuanced semantics encompasses a broader range of language descriptors. For example, while *Aletheia* currently supports two basic types of trends (i.e., increase/decrease), the real-world descriptors for trends can be more diverse and nuanced, considering various adjective/verb pairings (e.g., peak, tanking, spike, etc.) [14]. Recent studies [14, 48, 50] on automatic labeling and detecting such visual-text interplay and mismatch regarding temporal data offer more linguistic flexibility and sophisticated computational approaches to quantify such semantic nuances. These studies and approaches could be incorporated into *Aletheia*, potentially enhancing its capability to verify a wider range of data claims. As for compound claims, our unoptimized decomposition step can effectively separate simple compound claims. However, when faced with compound data claims coupled with information omission or co-references, *Aletheia* struggles to decompose them into distinct and accurate data facts, resulting in incomplete data claims or occasional hallucinations. We encourage future research to explore and optimize an end-to-end solution or curate ground-truth datasets for potential task-specific fine-tuning and evaluation.

10 CONCLUSION

Under the backdrop of escalating challenges posed by misinformation, our research delves into data claims — textual descriptions of facts/insight derived from structured, quantitative data sources. Specifically, we concentrate on two critical problems of automated fact-checking: (1) retrieving pertinent data evidence to *verify* data claims and (2) designing effective presentations to *communicate* the data evidence. We developed a prototype, *Aletheia*, to operationalize our proposed framework and tackle the multi-faceted challenge. We utilize a pretrained LLM to address the NLP tasks that decompose a data article and transform the data claims into data fact specifications. We explored the design space of data evidence by designing and implementing two representation formats, data table and visualization, across 13 types of data facts. Additionally, we equipped *Aletheia* with various interactions to enhance its utility and demonstrate its practical potential. Through a performance analysis with a manually curated dataset, we showcased LLM's robust capability both in classifying data facts and in translating textual claims into data fact specifications. We subsequently conducted a mixed-method user study with 20 participants, utilizing our designs as probes to gather insights into *assessment time*, *confidence*, and *preference.* Our findings revealed that our visualization designs are advantageous for 7 out of 13 data fact types regarding *assessment time.* Furthermore, based on participants' feedback and our reflection on the design process, we provided four general design recommendations for presenting data evidence. Ultimately, we discuss the limitations of our study and suggest avenues for future work to adapt and extend our work to accommodate more intricate real-world scenarios and thereby benefit broader audiences.

REFERENCES

[1] Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress Toward "the Holy Grail" : The Continued Quest to Automate Fact-Checking. https://api.semanticscholar.org/CorpusID:4648724

[2] Naser Ahmadi, Hansjorg Sand, and Paolo Papotti. 2022. Unsupervised Matching of Data and Text. In *2022 IEEE 38th International Conference on Data Engineering (ICDE).* IEEE Computer Society, 1058–1070. https://doi.org/10.1109/ICDE53745.2022.00084

[3] Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and reasoning over chart images for evidence-based automated fact-checking. *arXiv preprint arXiv:2301.11843* (2023).

[4] Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2023. Chartcheck: An evidence-based fact-checking dataset over real-world chart images. *arXiv preprint arXiv:2311.07453* (2023).

[5] Saud Althabiti, Mohammad Ammar Alsalka, and Eric Atwell. 2023. Generative AI for Explainable Automated Fact Checking on the FactEx: A New Benchmark Dataset. In *Disinformation in Open Online Media*, Davide Ceolin, Tommaso Caselli, and Marina Tulin (Eds.). Springer Nature Switzerland, Cham, 1–13.

[6] R. Amar, J. Eagan, and J. Stasko. 2005. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* 111–117. https://doi.org/10.1109/INFVIS.2005.1532136

[7] Michelle A Amazeen, Emily Thorson, Ashley Muddiman, and Lucas Graves. 2018. Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism & Mass Communication Quarterly* 95, 1 (2018), 28–48. https://doi.org/10.1177/1077699016678186

[8] Rahul Aralikatte, Matthew Lamm, Daniel Hardt, and Anders Søgaard. 2021. Ellipsis Resolution as Question Answering: An Evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.* Association for Computational Linguistics, Online, 810–817. https://doi.org/10.18653/v1/2021.eacl-main.68

[9] Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. *ClaimBuster: A Benchmark Dataset of Check-worthy Factual Claims.* https://doi.org/10.5281/zenodo.3836810

[10] Pepa Atanasova, Lluís Màrquez i Villodre, Alberto Barrón-Cedeño, T. Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino,

and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 1: Check-Worthiness. *ArXiv* abs/1808.05542 (2018). https://api.semanticscholar.org/CorpusID:51809121

[11] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7352–7364. https://doi.org/10.18653/v1/2020.acl-main.656

[12] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *EMNLP*. Association for Computational Linguistics.

[13] Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. 2018. Elastic documents: Coupling text and tables through contextual visualizations for enhanced document reading. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 661–671.

[14] Dennis Bromley and Vidya Setlur. 2023. What Is the Difference Between a Mountain and a Molehill? Quantifying Semantic Labeling of Visual Features in Line Charts. (2023). arXiv:2308.01370 [cs.HC]

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[16] Mars Gokturk Buchholz. 2023. Assessing the Effectiveness of GPT-3 in Detecting False Political Statements: A Case Study on the LIAR Dataset. *arXiv preprint arXiv:2306.08190* (2023).

[17] Tien Duc Cao, Ioana Manolescu, and Xavier Tannier. 2019. Extracting Statistical Mentions from Textual Claims to Provide Trusted Content. In *Transactions on Petri Nets and Other Models of Concurrency XVII*. Transactions on Petri Nets and Other Models of Concurrency XVII, 402–408. https://doi.org/10.1007/978-3-030-23281-8_36

[18] Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and Lei Li. 2022. LOREN: Logic-Regularized Reasoning for Interpretable Fact Verification. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 10 (Jun. 2022), 10482–10491. https://doi.org/10.1609/aaai.v36i10.21291

[19] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[20] Yang Chen, Jing Yang, and William Ribarsky. 2009. Toward effective insight management in visual analytics systems. In *2009 IEEE Pacific Visualization Symposium*. 49–56. https://doi.org/10.1109/PACIFICVIS.2009.4906837

[21] Zhutian Chen and Haijun Xia. 2022. CrossData: Leveraging Text-Data Connections for Authoring Data Documents *(CHI '22)*. Article 95, 15 pages. https://doi.org/10.1145/3491102.3517485

[22] Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational journalism: A call to arms to database researchers. 148–151. 5th Biennial Conference on Innovative Data Systems Research, CIDR 2011.

[23] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. (2022). arXiv:2209.01390 [cs.HC]

[24] Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing & Management* 60, 2 (2023), 103219. https://doi.org/10.1016/j.ipm.2022.103219

[25] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, and Dongmei Zhang. 2019. Quickinsights: Quick and automatic discovery of insights from multi-dimensional data. In *Proceedings of the 2019 International Conference on Management of Data*. 317–332.

[26] Jean Daniel Fekete, Jarke J. Van Wijk, John T. Stasko, and Chris North. 2008. The value of information visualization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 4950 LNCS. 1–18. https://doi.org/10.1007/978-3-540-70956-5_1

[27] Terry Flew, Christina Spurgeon, Anna Daniel, and Adam Swift. 2012. THE PROMISE OF COMPUTATIONAL JOURNALISM. *Journalism Practice* 6, 2 (2012), 157–171. https://doi.org/10.1080/17512786.2011.616655

[28] Yu Fu and John Stasko. 2023. More Than Data Stories: Broadening the Role of Visualization in Contemporary Journalism. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–20. https://doi.org/10.1109/TVCG.2023.3287585

[29] D Graves. 2018. Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism* (2018).

[30] Lucas Graves. 2016. *Deciding What's True: The Rise of Political Fact-Checking in American Journalism.* Columbia University Press. https://books.google.com/books?id=VcGlDAAAQBAJ

[31] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206. https://doi.org/10.1162/tacl_a_00454

[32] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. (2019). arXiv:1911.01214

[33] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A Survey on Stance Detection for Mis-and Disinformation Identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 1259–1277.

[34] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1803–1812.

[35] Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th acm international on conference on information and knowledge management*. 1835–1838.

[36] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Mueller, Francesco Piccinno, and Julian 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.

[37] Danula Hettiachchi, Kaixin Ji, Jenny Kennedy, Anthony McCosker, Flora Dylis Salim, Mark Sanderson, Falk Scholer, and Damiano Spina. 2023. Designing and Evaluating Presentation Strategies for Fact-Checked Content. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. ACM. https://doi.org/10.1145/3583780.3614841

[38] Emma Hoes, Sacha Altay, and Juan Bermeo. [n. d.]. Using ChatGPT to Fight Misinformation: ChatGPT Nails 72% of 12,000 Verified Claims. ([n. d.]).

[39] Eli Holder and Cindy Xiong. 2023. Dispersion vs Disparity: Hiding Variability Can Encourage Stereotyping When Visualizing Social Outcomes. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 624–634. https://doi.org/10.1109/TVCG.2022.3209377

[40] Jessica Hullman. 2020. Why Authors Don't Visualize Uncertainty. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 130–139. https://doi.org/10.1109/TVCG.2019.2934287

[41] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4198–4205.

[42] Xiao Jiang, Chengkai Li, Ping Luo, Min Wang, and Yong Yu. 2011. Prominent Streak Discovery in Sequence Data. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. ACM, 1280—-1288. https://doi.org/10.1145/2020408.2020601

[43] Saehan Jo, Immanuel Trummer, Weicheng Yu, Xuezhi Wang, Cong Yu, Daniel Liu, and Niyati Mehta. 2019. Verifying text summaries of relational data sets. In *Proceedings of the 2019 International Conference on Management of Data*. 299–316.

[44] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics* 8 (2020), 64–77. https://doi.org/10.1162/tacl_a_00300

[45] Sam C Joyce, Grace Guo, Bianchi Dy, Nazim Ibrahim, and Ate Poorthuis. 2019. Seeing numbers: Considering the effect of presentation of engineering data in design. In *Proceedings of IASS Annual Symposia*, Vol. 2019. International Association for Shell and Spatial Structures (IASS), 1–8.

[46] Prerna Juneja and Tanushree Mitra. 2022. Human and technological infrastructures of fact-checking. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–36.

[47] Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. 2020. Scrutinizer: A Mixed-Initiative Approach to Large-Scale, Data-Driven Claim Verification. *Proc. VLDB Endow.* 13, 12 (jul 2020), 2508–2521. https://doi.org/10.14778/3407790.3407841

[48] Dae Hyun Kim, Seulgi Choi, Juho Kim, Vidya Setlur, and Maneesh Agrawala. 2023. EMPHASISCHECKER: A Tool for Guiding Chart and Caption Emphasis. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–11. https://doi.org/10.1109/TVCG.2023.3327150

[49] Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. 2018. Facilitating Document Reading by Linking Text and Tables *(UIST '18)*. ACM, 423–434. https://doi.org/10.1145/3242587.3242617

[50] Dae Hyun Kim, Vidya Setlur, and Maneesh Agrawala. 2021. Towards Understanding How Readers Integrate Charts and Captions: A Case Study with Line Charts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, Article 610, 11 pages. https://doi.org/10.1145/3411764.3445443

[51] 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. *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI)* (2024). https://doi.org/10.1145/3640543.3645155

[52] 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*. 31–40.

[53] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking: A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5430–5443. https://doi.org/10.18653/v1/2020.coling-main.474

[54] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7740–7754. https://doi.org/10.18653/v1/2020.emnlp-main.623

[55] Dilek Küçük and Fazli Can. 2020. Stance Detection: A Survey. *ACM Comput. Surv.* 53, 1, Article 12 (feb 2020), 37 pages. https://doi.org/10.1145/3369026

[56] Shahid Latif, Zheng Zhou, Yoon Kim, Fabian Beck, and Nam Wook Kim. 2022. Kori: Interactive Synthesis of Text and Charts in Data Documents. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 184–194. https://doi.org/10.1109/TVCG.2021.3114802

[57] Po-Ming Law, Alex Endert, and John Stasko. 2020. Characterizing Automated Data Insights. In *2020 IEEE Visualization Conference (VIS)*. 171–175. https://doi.org/10.1109/VIS47514.2020.00041

[58] Ian Levy. 2023. *The Whiteboard: What pieces should the Bulls keep for a rebuild?* Retrieved Dec 12, 2023 from https://fansided.com/posts/bulls-pieces-rebuild-coby-white-patrick-williams

[59] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. StarCoder: may the source be with you! *arXiv preprint arXiv:2305.06161* (2023).

[60] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep entity matching with pre-trained language models. *Proceedings of the VLDB Endowment* 14, 1 (2020), 50–60.

[61] Maxim Lisnic, Cole Polychronis, Alexander Lex, and Marina Kogan. 2023. Misleading Beyond Visual Tricks: How People Actually Lie with Charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, Article 817, 21 pages. https://doi.org/10.1145/3544548.3580910

[62] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9, Article 195 (jan 2023), 35 pages. https://doi.org/10.1145/3560815

[63] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. *CoRR* cs.CL/0205028 (2002). https://arxiv.org/abs/cs/0205028

[64] Bobbi-Jean MacKinnon. 2023. . Retrieved Dec 12, 2023 from https://www.msn.com/en-ca/health/other/nb-reports-2-more-covid-19-deaths-2-children-hospitalized-rise-in-icu-admissions/ar-AA1loFGD

[65] Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 International Conference on Management of Data*. 19–34.

[66] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. (2021). arXiv:2103.07769 [cs.AI]

[67] An Nguyen and Jairo Lugo-Ocando. 2016. The state of data and statistics in journalism and journalism education: Issues and debates. *Journalism* 17, 1 (2016), 3–17. https://doi.org/10.1177/1464884915593234

[68] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 189–199.

[69] C. North. 2006. Toward measuring visualization insight. *IEEE Computer Graphics and Applications* 26, 3 (2006), 6–9. https://doi.org/10.1109/MCG.2006.70

[70] World Health Organization. 2019. *COVID-19 Data.* Retrieved Nov 18, 2023 from https://covid19.who.int/data

[71] PolitiFact. 2023. *PolitiFact.* Retrieved Nov 23, 2023 from https://www.politifact.com

[72] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 22–32.

[73] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 22–32. https://doi.org/10.18653/v1/D18-1003

[74] Dorian Quelle and Alexandre Bovet. 2023. The Perils & Promises of Fact-checking with Large Language Models. *arXiv preprint arXiv:2310.13549* (2023).

[75] Rayhane Rezgui, Mohammed Saeed, and Paolo Papotti. 2021. Automatic Verification of Data Summaries. In *Proceedings of the 14th International Conference on Natural Language Generation*. Association for Computational Linguistics, Aberdeen, Scotland, UK, 271–275. https://aclanthology.org/2021.inlg-1.27

[76] Md Main Uddin Rony, Enamul Hoque, and Naeemul Hassan. 2020. ClaimViz: Visual analytics for identifying and verifying factual claims. In *2020 IEEE Visualization Conference (VIS)*. IEEE, 246–250. https://doi.org/10.1109/VIS47514.2020.00056

[77] Nathan Schucher, Siva Reddy, and Harm de Vries. 2022. The Power of Prompt Tuning for Low-Resource Semantic Parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 148–156.

[78] Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 624–643. https://doi.org/10.18653/v1/2021.naacl-main.52

[79] Chirag Shah and Emily M. Bender. 2022. Situating Search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval (CHIIR '22)*. ACM, 221–232. https://doi.org/10.1145/3498366.3505816

[80] Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. 2021. Calliope: Automatic Visual Data Story Generation from a Spreadsheet. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 453–463. https://doi.org/10.1109/TVCG.2020.3030403

[81] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. DEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. ACM, 395–405. https://doi.org/10.1145/3292500.3330935

[82] Damiano Spina, Mark Sanderson, Daniel Angus, Gianluca Demartini, Dana Mckay, Lauren L Saling, and Ryen W White. 2023. Human-AI Cooperation to Tackle Misinformation and Polarization. *Commun. ACM* 66, 7 (2023), 40–45.

[83] Arjun Srinivasan, Steven M. Drucker, Alex Endert, and John Stasko. 2019. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 672–681. https://doi.org/10.1109/TVCG.2018.2865145

[84] Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)* (2020), 32–43.

[85] Nicole Sultanum and Arjun Srinivasan. 2023. DATATALES: Investigating the use of Large Language Models for Authoring Data-Driven Articles. In *2023 IEEE Visualization and Visual Analytics (VIS)*. 231–235. https://doi.org/10.1109/VIS54172.2023.00055

[86] Neset Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan, and Michael J Witbrock. 2023. Multi2Claim: Generating Scientific Claims from Multi-Choice Questions for Scientific Fact-Checking. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2644–2656.

[87] Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. ClaimsKG: A knowledge graph of fact-checked claims. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*. Springer, 309–324.

[88] James Thorne and Andreas Vlachos. 2017. An Extensible Framework for Verification of Numerical Claims. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, 37–40. https://aclanthology.org/E17-3010

[89] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

[90] Manju Vallayil, Parma Nand, Wei Qi Yan, and Héctor Allende-Cid. 2023. Explainability of Automated Fact Verification Systems: A Comprehensive Review. *Applied Sciences* 13, 23 (2023). https://doi.org/10.3390/app132312608

[91] Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Association for Computational Linguistics, Baltimore, MD, USA, 18–22. https://doi.org/10.3115/v1/W14-2508

[92] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7534–7550.

[93] David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 61–76.

[94] Brett Walenz, Y Wu, S Song, Emre Sonmez, Eric Wu, Kevin Wu, Pankaj K Agarwal, Jun Yang, Naeemul Hassan, Afroza Sultana, Gensheng Zhang, Chengkai Li, and Cong Yu. 2014. Finding, monitoring, and checking claims computationally based on structured data. In *Computation + Journalism Symposium*.

[95] Pengfei Wang, Xiaocan Zeng, Lu Chen, Fan Ye, Yuren Mao, Junhao Zhu, and Yunjun Gao. 2022. PromptEM: Prompt-Tuning for Low-Resource Generalized Entity Matching. 16, 2 (oct 2022), 369–378. https://doi.org/10.14778/3565816.3565836

[96] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[97] Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. 2020. DataShot: Automatic Generation of Fact Sheets from Tabular Data. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 895–905. https://doi.org/10.1109/TVCG.2019.2934398

[98] Hannah Ward-Glenton. 2023. *HSBC pulls some UK mortgage deals as fears of rising rates hits home buyers once more.* Retrieved Nov 18, 2023 from https://www.cnbc.com/2023/06/09/hsbc-pulls-some-uk-mortgage-deals-as-fears-of-rising-rates-hits-home-buyers-once-more.html

[99] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. (02 2023). https://doi.org/10.48550/arXiv.2302.11382 arXiv:2302.11382

[100] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, Article 385, 22 pages. https://doi.org/10.1145/3491102.3517582

[101] You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2017. Computational Fact Checking through Query Perturbations. *ACM Trans. Database Syst.* 42, 1, Article 4 (jan 2017), 41 pages. https://doi.org/10.1145/2996453

[102] You Wu, Brett Walenz, Peggy Li, Andrew Shim, Emre Sonmez, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. ICheck: Computationally Combating "Lies, d–Ned Lies, and Statistics". In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD '14)*. ACM, 1063–1066. https://doi.org/10.1145/2588555.2594522

[103] Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual Debiasing for Fact Verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 6777–6789. https://doi.org/10.18653/v1/2023.acl-long.374

[104] Ji Soo Yi, Youn-ah Kang, John T. Stasko, and Julie A. Jacko. 2008. Understanding and Characterizing Insights: How Do People Gain Insights Using Information Visualization?. In *Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization (BELIV '08)*. ACM, Article 4, 6 pages. https://doi.org/10.1145/1377966.1377971

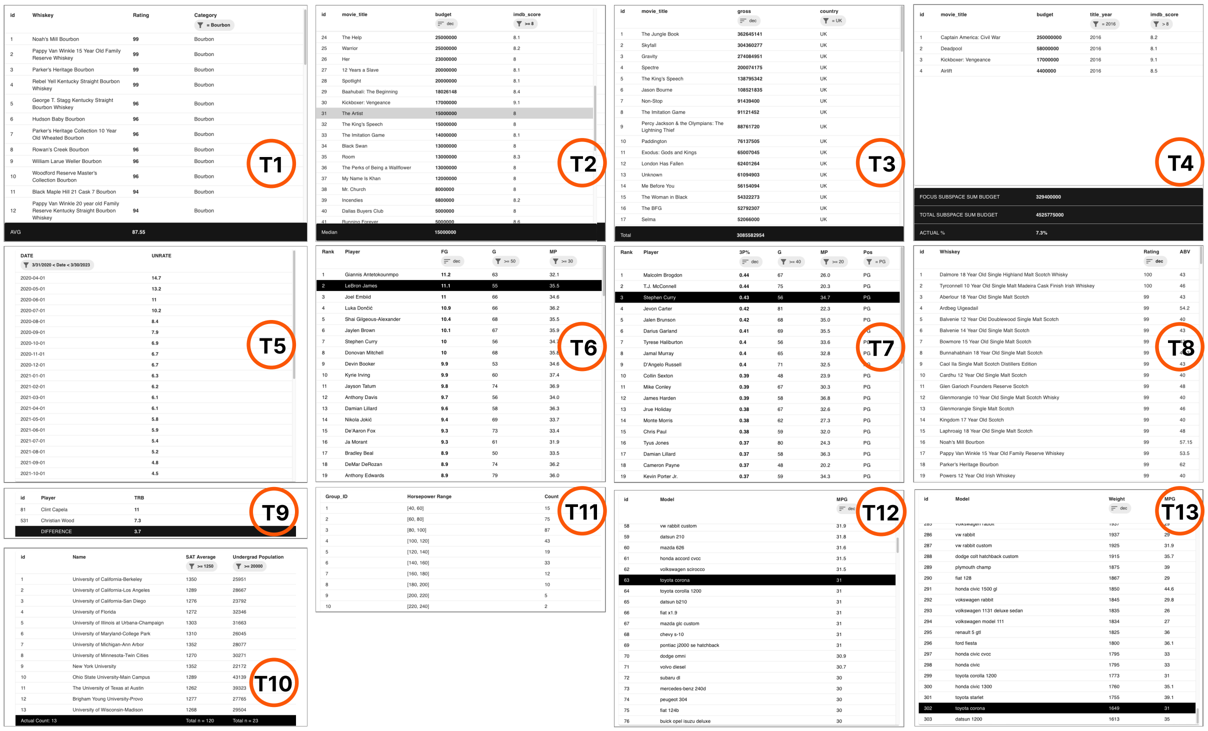
[105] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8413–8426. https://doi.org/10.18653/v1/2020.acl-main.745

[106] Qiyu Zhi, Alvitta Ottley, and Ronald Metoyer. 2019. Linking and Layout: Exploring the Integration of Text and Visualization in Storytelling. *Computer Graphics Forum* 38, 3, 675–685. https://doi.org/10.1111/cgf.13719

[107] Terry Yue Zhuo, Zhuang Li, Yujin Huang, Fatemeh Shiri, Weiqing Wang, Gholamreza Haffari, and Yuan-Fang Li. 2023. On Robustness of Prompt-based Semantic Parsing with Large Pre-trained Language Model: An Empirical Study on Codex. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1090–1102. https://doi.org/10.18653/v1/2023.eacl-main.77

[108] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. 51, 2, Article 32 (feb 2018), 36 pages. https://doi.org/10.1145/3161603

A APPENDIX



T1

id	Whiskey	Rating	Category = Bourbon
1	Noah's Mill Bourbon	99	Bourbon
2	Pappy Van Winkle 15 Year Old Family Reserve Whiskey	99	Bourbon
3	Parker's Heritage Bourbon	99	Bourbon
4	Rebel Yell Kentucky Straight Bourbon Whiskey	99	Bourbon
5	George T. Stagg Kentucky Straight Bourbon Whiskey	96	Bourbon
6	Hudson Baby Bourbon	96	Bourbon
7	Parker's Heritage Collection 10 Year Old Wheated Bourbon	96	Bourbon
8	Rowan's Creek Bourbon	96	Bourbon
9	William Larue Weller Bourbon	96	Bourbon
10	Woodford Reserve Master's Collection Bourbon	96	Bourbon
11	Black Maple Hill 21 Cask 7 Bourbon	94	Bourbon
12	Pappy Van Winkle 20 year old Family Reserve Kentucky Straight Bourbon Whiskey	94	Bourbon
AVG		87.55	

T2

id	movie_title	budget dec	imdb_score >= 8
24	The Help	25000000	8.1
25	Warrior	25000000	8.2
26	Her	23000000	8
27	12 Years a Slave	20000000	8.1
28	Spotlight	20000000	8.1
29	Baahubali: The Beginning	18026148	8.4
30	Kickboxer: Vengeance	17000000	9.1
31	The Artist	15000000	8
32	The King's Speech	15000000	8
33	The Imitation Game	14000000	8.1
34	Black Swan	13000000	8
35	Room	13000000	8.3
36	The Perks of Being a Wallflower	13000000	8
37	My Name Is Khan	12000000	8
38	Mr. Church	8000000	8
39	Incendies	6800000	8.2
40	Dallas Buyers Club	5000000	8
41	Running Forever	5000000	8.6
Median		15000000	

T3

id	movie_title	gross dec	country = UK
1	The Jungle Book	362645141	UK
2	Skyfall	304360277	UK
3	Gravity	274084951	UK
4	Spectre	200074175	UK
5	The King's Speech	138795342	UK
6	Jason Bourne	108521835	UK
7	Non-Stop	91439400	UK
8	The Imitation Game	91121452	UK
9	Percy Jackson & the Olympians: The Lightning Thief	88761720	UK
10	Paddington	76137505	UK
11	Exodus: Gods and Kings	65007045	UK
12	London Has Fallen	62401264	UK
13	Unknown	61094903	UK
14	Me Before You	56154094	UK
15	The Woman in Black	54322273	UK
16	The BFG	52792307	UK
17	Selma	52066000	UK
Total		3085582954	

T4

id	movie_title	budget	title_year = 2016	imdb_score > 8
1	Captain America: Civil War	250000000	2016	8.2
2	Deadpool	58000000	2016	8.1
3	Kickboxer: Vengeance	17000000	2016	9.1
4	Airlift	4400000	2016	8.5
FOCUS SUBSPACE SUM BUDGET		329400000		
TOTAL SUBSPACE SUM BUDGET		4525775000		
ACTUAL %		7.3%		

T5

DATE 3/31/2020 < Date < 3/30/2023	UNRATE
2020-04-01	14.7
2020-05-01	13.2
2020-06-01	11
2020-07-01	10.2
2020-08-01	8.4
2020-09-01	7.9
2020-10-01	6.9
2020-11-01	6.7
2020-12-01	6.7
2021-01-01	6.3
2021-02-01	6.2
2021-03-01	6.1
2021-04-01	6.1
2021-05-01	5.8
2021-06-01	5.9
2021-07-01	5.4
2021-08-01	5.2
2021-09-01	4.8
2021-10-01	4.5

T6

Rank	Player	FG dec	G >= 50	MP >= 30
1	Giannis Antetokounmpo	11.2	63	32.1
2	LeBron James	11.1	55	35.5
3	Joel Embiid	11	66	34.6
4	Luka Dončić	10.9	66	36.2
5	Shai Gilgeous-Alexander	10.4	68	35.5
6	Jaylen Brown	10.1	67	35.9
7	Stephen Curry	10	56	34.7
8	Donovan Mitchell	10	68	35.8
9	Devin Booker	9.9	53	34.6
10	Kyrie Irving	9.9	60	37.4
11	Jayson Tatum	9.8	74	36.9
12	Anthony Davis	9.7	56	34.0
13	Damian Lillard	9.6	58	36.3
14	Nikola Jokić	9.4	69	33.7
15	De'Aaron Fox	9.3	73	33.4
16	Ja Morant	9.3	61	31.9
17	Bradley Beal	8.9	50	33.5
18	DeMar DeRozan	8.9	74	36.2
19	Anthony Edwards	8.9	79	36.0

T7

Rank	Player	3P% dec	G >= 40	MP >= 20	Pos = PG
1	Malcolm Brogdon	0.44	67	26.0	PG
2	T.J. McConnell	0.44	75	20.3	PG
3	Stephen Curry	0.43	56	34.7	PG
4	Jevon Carter	0.42	81	22.3	PG
5	Jalen Brunson	0.42	68	35.0	PG
6	Darius Garland	0.41	69	35.5	[illegible]
7	Tyrese Haliburton	0.4	56	33.6	PG
8	Jamal Murray	0.4	65	32.8	PG
9	D'Angelo Russell	0.4	71	32.5	[illegible]
10	Collin Sexton	0.39	48	23.9	PG
11	Mike Conley	0.39	67	30.3	PG
12	James Harden	0.39	58	36.8	PG
13	Jrue Holiday	0.38	67	32.6	PG
14	Monte Morris	0.38	62	27.3	PG
15	Chris Paul	0.38	59	32.0	PG
16	Tyus Jones	0.37	80	24.3	PG
17	Damian Lillard	0.37	58	36.3	PG
18	Cameron Payne	0.37	48	20.2	PG
19	Kevin Porter Jr.	0.37	59	34.3	PG

T8

id	Whiskey	Rating dec	ABV
1	Dalmore 18 Year Old Single Highland Malt Scotch Whisky	100	43
2	Tyrconnell 10 Year Old Single Malt Madeira Cask Finish Irish Whiskey	100	46
3	Aberlour 18 Year Old Single Malt Scotch	99	43
4	Ardbeg Uigeadail	99	54.2
5	Balvenie 12 Year Old Doublewood Single Malt Scotch	99	40
6	Balvenie 14 Year Old Single Malt Scotch	99	43
7	Bowmore 15 Year Old Single Malt Scotch	99	[illegible]
8	Bunnahabhain 18 Year Old Single Malt Scotch	99	[illegible]
9	Caol Ila Single Malt Scotch Distillers Edition	99	43
10	Cardhu 12 Year Old Single Malt Scotch	99	40
11	Glen Garioch Founders Reserve Scotch	99	48
12	Glenmorangie 10 Year Old Single Malt Scotch Whisky	99	40
13	Glenmorangie Single Malt Scotch	99	46
14	Kingdom 17 Year Old Scotch	99	40
15	Laphroaig 18 Year Old Single Malt Scotch	99	48
16	Noah's Mill Bourbon	99	57.15
17	Pappy Van Winkle 15 Year Old Family Reserve Whiskey	99	53.5
18	Parker's Heritage Bourbon	99	62
19	Powers 12 Year Old Irish Whiskey	99	40

T9

id	Player	TRB
81	Clint Capela	11
531	Christian Wood	7.3
	DIFFERENCE	3.7

T10

id	Name	SAT Average >= 1250	Undergrad Population >= 20000
1	University of California-Berkeley	1350	25951
2	University of California-Los Angeles	1289	28667
3	University of California-San Diego	1276	23792
4	University of Florida	1272	32346
5	University of Illinois at Urbana-Champaign	1303	31663
6	University of Maryland-College Park	1310	26045
7	University of Michigan-Ann Arbor	1352	28077
8	University of Minnesota-Twin Cities	1270	30271
9	New York University	1352	22172
10	Ohio State University-Main Campus	1289	43139
11	The University of Texas at Austin	1262	39323
12	Brigham Young University-Provo	1277	27765
13	University of Wisconsin-Madison	1268	29504
Actual Count: 13		Total n = 120	Total n = 23

T11

Group_ID	Horsepower Range	Count
1	[40, 60]	15
2	[60, 80]	75
3	[80, 100]	87
4	[100, 120]	43
5	[120, 140]	19
6	[140, 160]	33
7	[160, 180]	12
8	[180, 200]	10
9	[200, 220]	5
10	[220, 240]	2

T12

id	Model	MPG dec
58	vw rabbit custom	31.9
59	datsun 210	31.8
60	mazda 626	31.6
61	honda accord cvcc	31.5
62	volkswagen scirocco	31.5
63	toyota corona	31
64	toyota corolla 1200	31
65	datsun b210	31
66	fiat x1.9	31
67	mazda glc custom	31
68	chevy s-10	31
69	pontiac j2000 se hatchback	31
70	dodge omni	30.9
71	volvo diesel	30.7
72	subaru dl	30
73	mercedes-benz 240d	30
74	peugeot 304	30
75	fiat 124b	30
76	buick opel isuzu deluxe	30

T13

id	Model	Weight dec	MPG
285	volkswagen rabbit	1937	29
286	vw rabbit	1937	29
287	vw rabbit custom	1925	31.9
288	dodge colt hatchback custom	1915	35.7
289	plymouth champ	1875	39
290	fiat 128	1867	29
291	honda civic 1500 gl	1850	44.6
292	vokswagen rabbit	1845	29.8
293	volkswagen 1131 deluxe sedan	1835	26
294	volkswagen model 111	1834	27
295	renault 5 gtl	1825	36
296	ford fiesta	1800	36.1
297	honda civic cvcc	1795	33
298	honda civic	1795	33
299	toyota corolla 1200	1773	31
300	honda civic 1300	1760	35.1
301	toyota starlet	1755	39.1
302	toyota corona	1649	31
303	datsun 1200	1613	35

Figure 7: The design of our thirteen data evidence tables. *sorting* and *filtering* widgets are displayed underneath corresponding column names. (T1) a sticky, highlighted row displaying *mean value*; (T2) a sticky, highlighted row displaying *median value* and highlighted row(s) for median rows; (T3) a sticky, highlighted row displaying the *sum (value)*; (T4) three sticky, highlighted summarization rows displaying sum values of the focus set, the reference set, and the computed *proportion*; (T5) chronological table showing the mentioned timeframe; (T6) *extreme* & (T7) *rank* - sorted, indexed table highlighting mentioned entity row; (T8) *association* - two mentioned measures with one measure sorted; (T9) three-row table displaying the two compared entities and a highlighted row displaying the *difference*; (T10) sticky, highlighted row displaying the counts of data points satisfying each category and their *categorization* overlap; (T11) two column table (bins & range) showing the *distribution*; (T12) *univariate outlier* & (T13) *multivariate outlier* - sorted and indexed table displaying individual values and highlighting mentioned entity row.

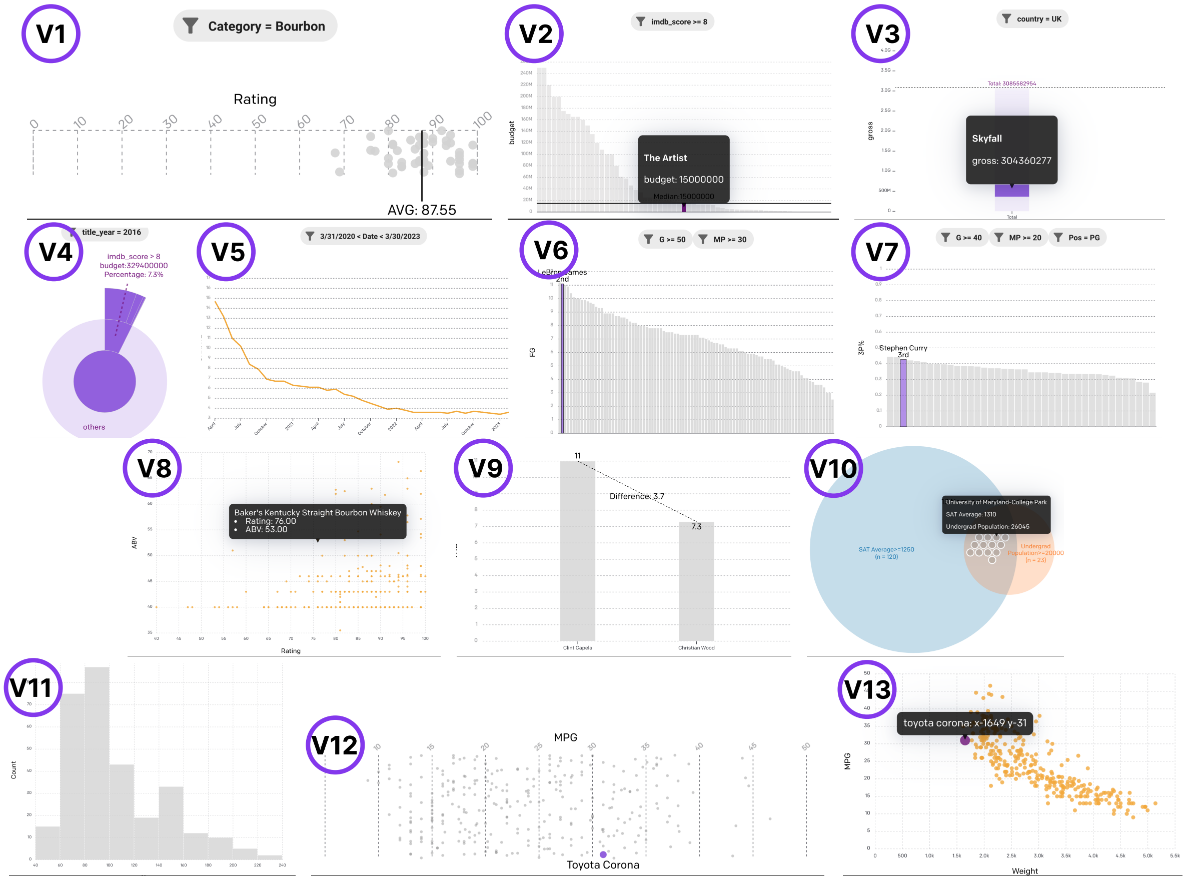
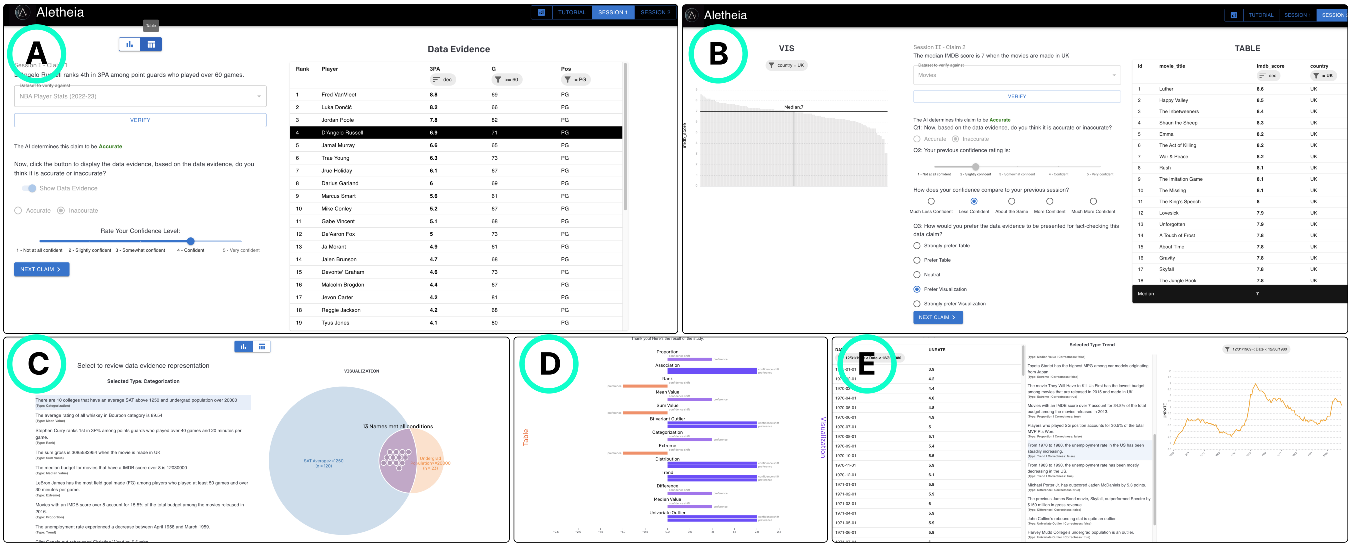


Figure 8: The design of our 13 data evidence visualizations: (V1) a strip plot accompanied by a line that indicates their *mean value*; (V2) a sorted bar chart with the *median values* highlighted and labeled; (V3) a stacked bar chart depicting the *sum (value)*; (V4) a sunburst plot showing the *proportion* of individual data points; (V5) a line graph showing the *trend* for a given timeframe; (V6) *extreme* and (V7) *rank*: a sorted bar chart with the mentioned data point highlighted and labeled; (V8) a scatterplot that shows the *association* between values; (V9) two bars with a comparison line showing the *difference*; (V10) a proportional Venn diagram showing the number of points based on the *categorization* overlap; (V11) a histogram displaying the *distribution*; (V12) *univariate outlier* and (V13) *multivariate outliers* - a strip/scatter plot with the mentioned data point highlighted.



Session I - Claim 1
D'Angelo Russell ranks 4th in 3PA among point guards who played over 60 games.

Dataset to verify against: NBA Player Stats (2022-23)

VERIFY

The AI determines this claim to be Accurate

Now, click the button to display the data evidence, based on the data evidence, do you think it is accurate or inaccurate?

Show Data Evidence

○ Accurate ● Inaccurate

Rate Your Confidence Level:

1 - Not at all confident 2 - Slightly confident 3 - Somewhat confident 4 - Confident 5 - Very confident

NEXT CLAIM >

Data Evidence

Rank	Player	3PA dec	G >= 60	Pos = PG
1	Fred VanVleet	8.8	69	PG
2	Luka Dončić	8.2	66	PG
3	Jordan Poole	7.8	82	PG
4	D'Angelo Russell	6.9	71	PG
5	Jamal Murray	6.6	65	PG
6	Trae Young	6.3	73	PG
7	Jrue Holiday	6.1	67	PG
8	Darius Garland	6	69	PG
9	Marcus Smart	5.6	61	PG
10	Mike Conley	5.2	67	PG
11	Gabe Vincent	5.1	68	PG
12	De'Aaron Fox	5	73	PG
13	Ja Morant	4.9	61	PG
14	Jalen Brunson	4.7	68	PG
15	Devonte' Graham	4.6	73	PG
16	Malcolm Brogdon	4.4	67	PG
17	Jevon Carter	4.2	81	PG
18	Reggie Jackson	4.2	68	PG
19	Tyus Jones	4.1	80	PG

country = UK

Session II - Claim 2
The median IMDB score is 7 when the movies are made in UK

Dataset to verify against: Movies

VERIFY

The AI determines this claim to be Accurate
Q1: Now, based on the data evidence, do you think it is accurate or inaccurate?

○ Accurate ● Inaccurate

Q2: Your previous confidence rating is:

1 - Not at all confident 2 - Slightly confident 3 - Somewhat confident 4 - Confident 5 - Very confident

How does your confidence compare to your previous session?

○ Much Less Confident ● Less Confident ○ About the Same ○ More Confident ○ Much More Confident

Q3: How would you prefer the data evidence to be presented for fact-checking this data claim?

○ Strongly prefer Table
○ Prefer Table
○ Neutral
● Prefer Visualization
○ Strongly prefer Visualization

NEXT CLAIM >

id	movie_title	imdb_score dec	country = UK
1	Luther	8.6	UK
2	Happy Valley	8.5	UK
3	The Inbetweeners	8.4	UK
4	Shaun the Sheep	8.3	UK
5	Emma	8.2	UK
6	The Act of Killing	8.2	UK
7	War & Peace	8.2	UK
8	Rush	8.1	UK
9	The Imitation Game	8.1	UK
10	The Missing	8.1	UK
11	The King's Speech	8	UK
12	Lovesick	7.9	UK
13	Unforgotten	7.9	UK
14	A Touch of Frost	7.8	UK
15	About Time	7.8	UK
16	Gravity	7.8	UK
17	Skyfall	7.8	UK
18	The Jungle Book	7.8	UK
Median		7	

Select to review data evidence representation

Selected Type: Categorization

There are 10 colleges that have an average SAT above 1250 and undergrad population over 20000
(Type: Categorization)

The average rating of all whiskey in Bourbon category is 89.54
(Type: Mean Value)

Stephen Curry ranks 1st in 3P% among points guards who played over 40 games and 20 minutes per game.
(Type: Rank)

The sum gross is 3085582954 when the movie is made in UK
(Type: Sum Value)

The median budget for movies that have a IMDB score over 8 is 12030000
(Type: Median Value)

LeBron James has the most field goal made (FG) among players who played at least 50 games and over 30 minutes per game.
(Type: Extreme)

Movies with an IMDB score over 8 account for 15.5% of the total budget among the movies released in 2016.
(Type: Proportion)

The unemployment rate experienced a decrease between April 1958 and March 1959.
(Type: Trend)

13 Names met all conditions

SAT Average>=1250 (n = 120)

Undergrad Population>=20000 (n = 23)

Thank you! Here's the result of the study.

Proportion, Association, Rank, Mean Value, Sum Value, Bi-variant Outlier, Categorization, Extreme, Distribution, Trend, Difference, Median Value, Univariate Outlier

confidence shift, preference

-2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5

12/31/1969 < Date < 12/30/1980

DATE	UNRATE
1970-01-01	3.9
1970-02-01	4.2
1970-03-01	4.4
1970-04-01	4.6
1970-05-01	4.8
1970-06-01	4.9
1970-07-01	5
1970-08-01	5.1
1970-09-01	5.4
1970-10-01	5.5
1970-11-01	5.9
1970-12-01	6.1
1971-01-01	5.9
1971-02-01	5.9
1971-03-01	6
1971-04-01	5.9
1971-05-01	5.9
1971-06-01	5.9

Selected Type: Trend

(Type: Median Value | Correctness: false)

Toyota Starlet has the highest MPG among car models originating from Japan.
(Type: Extreme | Correctness: false)

The movie They Will Have to Kill Us First has the lowest budget among movies that are released in 2015 and made in UK.
(Type: Extreme | Correctness: true)

Movies with an IMDB score over 7 account for 34.8% of the total budget among the movies released in 2013.
(Type: Proportion | Correctness: true)

Players who played SG position accounts for 30.5% of the total MVP Pts Won.
(Type: Proportion | Correctness: false)

From 1970 to 1980, the unemployment rate in the US has been steadily increasing.
(Type: Trend | Correctness: false)

From 1983 to 1990, the unemployment rate has been mostly decreasing in the US.
(Type: Trend | Correctness: true)

Michael Porter Jr. has outscored Jaden McDaniels by 5.3 points.
(Type: Difference | Correctness: true)

The previous James Bond movie, Skyfall, outperformed Spectre by $150 million in gross revenue.
(Type: Difference | Correctness: false)

John Collins's rebounding stat is quite an outlier.
(Type: Univariate Outlier | Correctness: false)

Harvey Mudd College's undergrad population is an outlier.
(Type: Univariate Outlier | Correctness: true)

Figure 9: Our user study interface. It consists of five components: (A) *Phase I page*: the left side includes the questions and selections while the right side displays the respective data evidence representations; (B) *Phase II page*: its initial state resembles *study I page*. After participants click on *Accurate/Inaccurate* selection, both representations are displayed (left: initial form, right: alternative form). In the middle are the additional questions and scale selections (i.e., confidence shift and preference); (C) *Tutorial page*: it allows participants to click through the demo claims and get familiarized with the encodings for respective representation forms; (D) *Result chart*: it appears after *Phase II* and demonstrates the results of participants’ average confidence shift and preference for each type of data fact; (E) *Exploration page*: supports participants to click through all the test claims and review the corresponding representations during our interview portion.