DataNarrative: Automated Data-Driven Storytelling with Visualizations and Texts

Mohammed Saidul Islam[♣], **Md Tahmid Rahman Laskar**[♣♡]
Md Rizwan Parvez[♥], **Enamul Hoque**[♣], **Shafiq Joty**[♦♠]
[♣]York University, Canada, [♠]Nanyang Technological University, Singapore,
[♡]Dialpad Inc., Canada, [♥]Qatar Computing Research Institute (QCRI), [♦]Salesforce AI
{saidulis, enamulh, tahmid20}@yorku.ca, mparvez@hbku.edu.qa, sjoty@salesforce.com

Abstract

Data-driven storytelling is a powerful method for conveying insights by combining narrative techniques with visualizations and text. These stories integrate visual aids, such as highlighted bars and lines in charts, along with textual annotations explaining insights. However, creating such stories requires a deep understanding of the data and meticulous narrative planning, often necessitating human intervention, which can be time-consuming and mentally taxing. While Large Language Models (LLMs) excel in various NLP tasks, their ability to generate coherent and comprehensive data stories remains underexplored. In this work, we introduce a novel task for data story generation and a benchmark containing 1,449 stories from diverse sources. To address the challenges of crafting coherent data stories, we propose a multi-agent framework employing two LLM agents designed to replicate the human storytelling process: one for understanding and describing the data (Reflection), generating the outline, and narration and another for verification at each intermediary step. While our agentic framework generally outperforms non-agentic counterparts in both model-based and human evaluations, the results also reveal unique challenges in data story generation.

1 Introduction

Visual data stories have emerged as a powerful medium for communicating data, effectively combining the strengths of visualizations and text to convey contextual information and causal relationships (Hullman and Diakopoulos, 2011). Ranging from data scientists to business analysts to journalists, people frequently write data-driven reports that integrate charts and text to present information to readers in a clear, coherent and visually engaging manner (Otten et al., 2015). The essence of a visual data story involves identifying compelling insights within data ("story pieces"), presenting

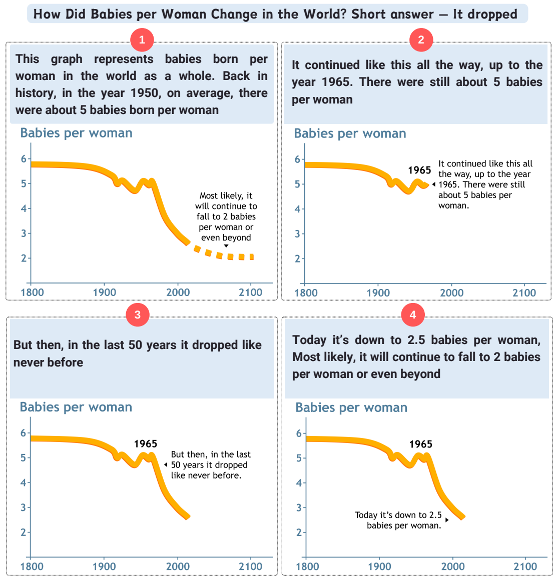


Figure 1: An example data story in our corpus extracted from GapMinder (Rosling, 2023)

them through visualizations and texts, and arranging these representations into a coherent narrative that communicates an overarching message (Lee et al., 2015). Well-crafted visual stories have the potential to significantly enhance data understanding, even for those without specialized technical backgrounds. By combining narrative with data visualization, authors can illustrate trends, highlight correlations, and uncover hidden insights that might be lost in dense tables or reports. For example, Fig. 1 shows a GapMinder data story (Rosling, 2023) in which renowned storyteller Hans Rosling explained how birth rates in the world have changed over time using text and charts.

Data storytelling is widely used across various companies, including Microsoft and Tableau to effectively communicate insights and drive decision-making. In business intelligence, it can help present sales trends and performance metrics, while in healthcare, it can help illustrate patient outcomes and track disease outbreaks. Marketers can lever-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19253–19286
November 12-16, 2024 ©2024 Association for Computational Linguistics

age it to optimize strategies through customer behavior visualization, and financial analysts use it for investment performance and risk assessments. In education, it can help track students' performance highlighting areas where they excel and where they might need additional support, and in public policy, it can communicate the impact of policies on social issues, for instance, a data story could show how a new housing policy affected homelessness rates, providing evidence-based insights to policymakers and the public alike.

Despite the popularity of data-driven stories, crafting them remains challenging and time-consuming, requiring skills in data analysis, visualization, graphic design, and storytelling. To facilitate data-driven storytelling, extensive research has introduced new concepts, theories, and tools. For instance, Segel and Heer (2010) explored different design spaces from a narrative structure point of view, while others (Hullman et al., 2013b; Lan et al., 2022; McKenna et al., 2017; Shi et al., 2021b,c) focused on visual representations for crafting visual stories, tailoring their approaches based on specific tasks and communication objectives. While insightful and coherent, manually created data stories require significant human effort and time. In response, efforts have been made to develop automated methods for generating data stories (Shi et al., 2019, 2021a; Wang et al., 2020b), but these often produce simple facts lacking in quality and engaging narratives.

The rise of LLMs has prompted researchers to explore their effectiveness in tasks like chart summarization (Kantharaj et al., 2022b; Rahman et al., 2023), chart question answering (Masry et al., 2022; Kantharaj et al., 2022a) and natural language story generation (Zhou et al., 2023; Xie and Riedl, 2024). However, the ability of LLMs to generate stories from data tables and to understand their effectiveness remains largely unexplored partly because of the lack of a benchmark dataset.

To address the research gap, we first develop a new task and the corresponding benchmark consisting of 1,449 data stories collected from real-world sources. Motivated by the impressive performance of LLM-based agents in various planning tasks (Ge et al., 2023; Yang et al., 2023a; Wang et al., 2023a; Modarressi et al., 2023; Chen et al., 2024; Wu et al., 2023), we then propose an agentic framework which takes data tables as inputs and employs two LLM agents – a Generator or Actor and an Evaluator or Critic – to mimic the human process of data story generation through writing and revising based on Critic's feedback (Figure 2). The process includes a planning step (reflection and outline generation) and a story generation step (narration), with each step verified and revised by the critic LLM, creating a feedback loop to ensure coherence and factual consistency. Experimental results show that our agentic framework outperforms non-agentic LLM counterparts in terms of generating more insightful and coherent stories with better resemblance to human-written narratives.

Our main contributions include: (*i*) a new automatic data story generation task and a corresponding benchmark dataset, (*ii*) a multi-step LLM-agent framework for Data Story Generation. (*iii*) extensive automatic and human evaluations that demonstrate the state-of-the-art performance of DATANARRATIVE. We make our code and data story corpus publicly available at here.

2 Related Work

2.1 Story Generation Tasks

Automated story generation is an open-ended task focusing on generating a sequence of events based on specific criteria (Li et al., 2013). Generated stories can be textual (Kumar et al., 2006), visual (Li et al., 2019; Cohn, 2020), or multimodal (Bensaid et al., 2021). Visual stories, often found in comics and storyboards, present image sequences centered around main characters (Cohn, 2020). Early visual story generation models primarily utilized either global image features (Yu et al., 2017; Wang et al., 2018; Huang et al., 2019) or local features, which focus on specific parts of an image, such as objects (Wang et al., 2020a; Hong et al., 2020; Braude et al., 2022), to create visually grounded stories.

Data-driven stories differ from visual stories as they produce multimodal outputs in which charts communicate patterns, trends, and outliers in data while text explains such visualizations (Riche et al., 2018a; Kwon et al., 2014; Segel and Heer, 2010; Hullman et al., 2013a). Early work focused on extracting and ranking key insights from data tables using statistical measures (Ding et al., 2019; Tang et al., 2017). Tools like DataShot (Wang et al., 2020b) and Calliope (Shi et al., 2021a) present data facts with visualizations and captions, while Erato (Sun et al., 2023) and Socrates (Wu et al., 2024) incorporate user input to guide the story generation process. In addition, there has been a recent survey (He et al., 2024) that explores the utilization of large models in narrative visualization. However, the methods used in existing works often use simple rule-based approaches that may miss critical insights and lack effective narrative structure.

2.2 LLMs for Story Generation

Recent LLMs such as Gemini (Team et al., 2023), ChatGPT (OpenAI, 2023), and GPT-4 (OpenAI, 2023a) excel at generating fluent stories by repeatedly providing contextual information from both the plan and the current state of the story to an LLM prompt (Yang et al., 2022; Wang et al., 2023b). Several studies confirm the effectiveness of LLMs in generating short (Eldan and Li, 2023), coherent and fluent stories (Peng et al., 2022). However, data story generation using LLMs is rare; one exception is DataTales (Sultanum and Srinivasan, 2023), which uses LLMs for narrative generation from chart images but is limited to only producing textual narratives without charts.

Recent studies also explore LLM agents in decision-making (Yang et al., 2023a), task planning in video games (Wang et al., 2023a), memory function configuration (Modarressi et al., 2023), multi-agent conversations (Wu et al., 2023), and code generation (Ridnik et al., 2024; Islam et al., 2024a). Despite the suitability of this approach for open-ended tasks requiring planning, LLM agents for data story generation remain unexplored.

2.3 Chart-related Downstream Tasks

Several downstream tasks associated with charts have been proposed recently. Masry et al. (2022); Methani et al. (2020) focus on answering factual questions about charts that require arithmetic and visual reasoning, while Kantharaj et al. (2022a) address open-ended question-answering that generates explanatory texts. Chart summarization task involves generating informative summaries from a chart (Kantharaj et al., 2022b; Tang et al., 2023; Rahman et al., 2023), while Chart-to-Table (Choi et al., 2019; Masry et al., 2023, 2024) extracts the underlying data tables from a chart image. Others focus on verifying claims about charts (Akhtar et al., 2023, 2024). Unlike the above tasks which produce only text, data-driven stories are multimodal as they combine visualizations with texts and there are no existing benchmarks for this task.

	Pew		**Tableau**		**GapMinder**	
# of Samples	Train	Test	Train	Test	Train	Test
# of Stories	1,068	321	42	13	-	5
# of Tables	4,729	1,590	340	64	-	42
# of Charts	4,729	1,590	297	64	-	42

Table 1: Distribution of stories, charts, and tables across the train and test split of three datasets

3 Benchmark Construction

Given the lack of a benchmark for automated data storytelling, we started by exhaustively searching across diverse online sources such as news sites, visualization repositories, and data blog sites. At the end, we chose three suitable sources that contain data stories covering a series of visualizations and texts as we described below.

3.1 Data Collection

- **Pew** Pew Research (Pew, 2024) publishes data reports related to social issues, public opinion, and demographic trends. Often, such reports include charts and accompanying texts to communicate a coherent data story. To assemble the Pew corpus, we crawled articles from the Pew Research website until March 14, 2024, resulting in 4,532 articles across 18 topics and 22,760 figures (i.e., charts and other images). For each article, we extracted the title, paragraphs, and chart images and their metadata (e.g., captions and alt-texts).
- **Tableau** Tableau Public Story (Tableau, 2024) allows users to create interactive stories through data visualizations on various topics and make these stories publicly accessible. Collecting data from Tableau with web crawlers proved difficult due to the complicated nature of the story representation, leading us to manually curate stories from the website. Specifically, we looked for stories that presented a paginated view, each page containing text and an associated chart. We searched by terms like 'story', 'data story', and 'narrative-visualization' on the Tableau public, which led us to find over 1,200 dashboards with potential data stories. From these, we filtered out dashboards that did not have paginated views with a series of pages containing both text and charts. This filtering process led us to select 100 candidate stories for our corpus. For each story page, we downloaded the chart image, data table, title, and text.
- **GapMinder** GapMinder (Rosling, 2023) offers interactive data visualization tools and educa-

	Pew		Tableau		GapMinder	
Type	Train	Test	Train	Test	Train	Test
Bar	3949	1159	155	46	-	-
Line	433	360	69	8	-	31
Pie	191	53	9	2	-	-
Scatter	42	10	36	6	-	-
Bubble	-	-	16	1	-	11
Other	114	8	12	1	-	-
Total	4729	1590	297	64	-	42

Table 2: Chart type distribution

tional resources on global trends in health, wealth, and development indicators. Similar to Tableau stories, GapMinder stories were challenging to crawl due to the tool's interactive nature. Additionally, only a small subset of data articles featured both a paginated view and a combination of text and charts, resulting in 11 data stories. For each page in these stories, we downloaded the chart image and other associated data.

3.2 Data Processing & Annotation

Data processing and annotations follow three steps: *(i)* story filtering, *(ii)* chart data extraction, *(iii)* chart-text pairs identification.

- **Story Filtering** To ensure the quality of our corpus, we applied the following exclusion criteria (**EC**) for filtering data stories from the initial collection: (*i*) stories with texts shorter than 500 tokens for Pew and 140 tokens for Tableau and GapMinder samples, (*ii*) Stories with fewer than 3 or more than 10 charts. By applying these criteria, we carefully selected the stories from Pew, Tableau, and GapMinder, resulting in a total of 1,449 stories. Also, some Tableau stories included complex and unconventional visualizations, such as infographics and treemaps, so we filtered these stories to retain the ones with common visualizations.
- **Chart data extraction** Chart data tables are essential for the story-generation process as we use them as inputs to the proposed framework. Also, to identify the text associated with each chart, we first need to extract the underlying data table of the chart image. We managed to download some gold data tables either from the story page (for Tableau) or from external sources (OWID (2024) for Gapminder). However, for Pew, we needed to automatically extract data from chart images as the original data tables were not available. Specifically, we utilized the multi-modal large language model Gemini-1.0-pro-vision (Team et al., 2023) to extract data from chart images, which has been found to be effective for this task (Islam et al., 2024b). On 100 chart images from the ChartQA (Masry et al., 2022) corpus, where gold tables were already available, we manually evaluated and found that the model correctly generated the tables in 77% of the cases (more details in Appendix A.3).
- **Identification of chart-text pairs** Since data stories usually come with descriptive texts for charts, it was essential to identify the texts related to each chart. Given the relatively small sizes of the Tableau and GapMinder corpus, we manually extracted the paragraphs associated with each chart image. For Pew, the chart-text pairs were already identified in the Chart-to-Text corpus (Kantharaj et al., 2022b) for 321 articles. However, for the remaining 1068 articles, we did not have the chart-text pairs. Due to the large sample size, collecting chart-text manually would be labor-intensive and time-consuming. Therefore, we utilized the state-of-the-art GPT-4-turbo model (OpenAI, 2023b) to collect relevant paragraphs corresponding to each of the charts in the training set. On a small subset of human-annotated Chart-to-Text corpus, the model accurately linked paragraphs to data tables 70% of the time (more details in Appendix A.4).

Data Splits After conducting the filtering process using the **ECs**, we selected 1,389 articles from the Pew Research corpus, 55 stories from Tableau story dashboards, and 5 stories from GapMinder, and split them into training and test sets as shown in Table 1. To create the test set from the Pew corpus, we selected the articles that also appear in the Chart-to-Text (Kantharaj et al., 2022b) corpus, as their chart-summary pairs were identified by human annotators to ensure the quality of the test set. For the Pew training set, we used GPT-4 model-generated annotations as explained earlier.

3.3 Features of DATANARRATIVE

We analyze our corpus statistics to highlight the key features of DATANARRATIVE. More details of the corpus analysis are included in Appendix A.5.

Diversity: Our benchmark contains stories covering a wide range of topics, from 'Politics & Policy' to 'International Affairs,' 'Education,' and 'Economy' (Fig. 4, and Fig. 7). Topics in GapMinder and Tableau are more evenly distributed while Pew is dominated by 'Politics & Policy' (57.24%). The corpus also includes a diverse range of chart types

	Pew		Tableau		GapMinder	
Statistics	Train	Test	Train	Test	Train	Test
Avg. length of Stories	1804	2865	837	1009	-	707
Avg. # of Tokens	353	561	159	194	-	146
Avg. # of Paragraphs	4	5	5	4	-	8
Avg. V. : T. ratio (↑)	0.51	0.46	0.64	0.63	-	0.63
Avg. # of unique V. (↑)	14	23	5	11	-	5
Avg. % of diverse V. (↑)	44	47	25	30	-	39
% of Intra 3-gram rep. (↓)	18.38	17.94	12.79	14.24	-	11.30
% of Inter 3-gram rep. (↓)	14.84	11.28	0.64	0.45	-	2.45

Table 3: DataNarrative dataset statistics. Here, 'V.' denotes 'Verb', 'T.' denotes 'Token', and 'rep.' denotes 'repetition'.

such as bars, lines, pies, and scatter plots (Table 2), with bar charts being the most common (78.98%), followed by line charts (13.40%).

Long, multimodal outputs: Unlike existing chart domain benchmarks that produce short summaries (Kantharaj et al., 2022b) or answers (Masry et al., 2022) related to charts, DATANARRATIVE have stories with multiple text paragraphs (Table 3), suggesting the open-ended nature of the task. Among them, Pew stories tend to be longer with an average story length of 2334.5 characters and 457 average tokens. Each story contains 4.5 charts and corresponding paragraphs on average, demonstrating the need for planning a narrative structure that has a multimodal output covering several visualizations and related texts.

Semantically rich stories: To assess semantic richness, we analyzed Vocab: Token Ratio, unique verbs, diverse verbs per story, and intra/inter-story trigram repetitions, common metrics for measuring content originality and diversity in story corpus (Goldfarb-Tarrant et al., 2020). As shown in Table 3, the Tableau corpus has the highest verb-to-token ratio (0.63), while the Pew has the most unique verbs (18.5) and the highest percentage of diverse verbs (45.5%), indicating high semantic richness. Trigram repetition is also higher in Pew, likely due to the greater length of Pew stories.

4 Methodology

4.1 Overall Framework

Task Formulation: Given one or more data table(s) and associated titles D, a user intent I representing the main theme of the story, and additional guidelines G as inputs, the expected output is a coherent data story S consisting of multiple textual paragraphs and corresponding visualization specifications (e.g., chart type, x-axis/y-axis values, x-axis/y-axis labels, etc.). These visualization

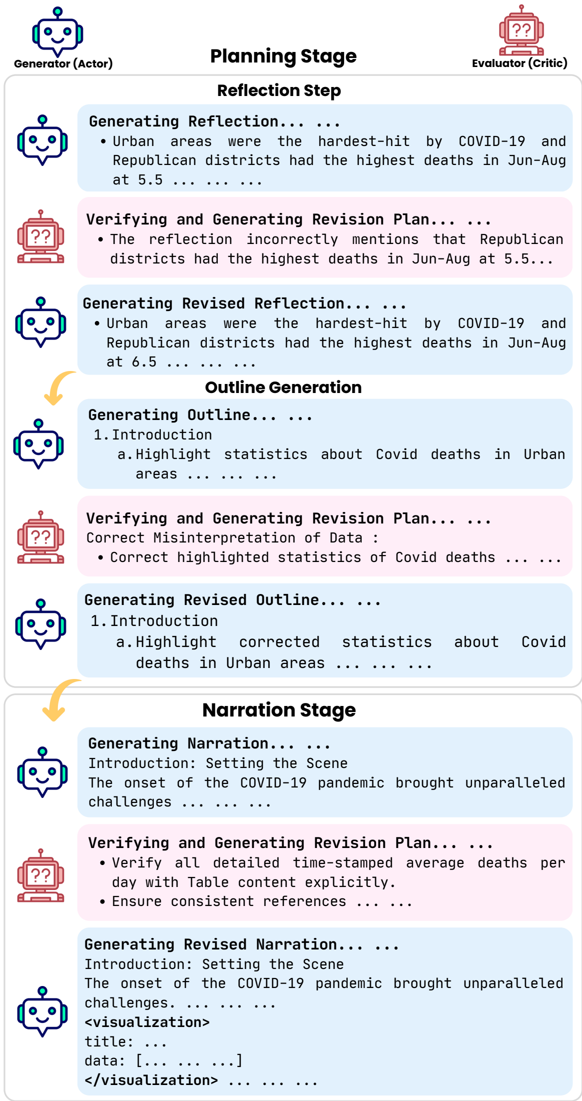


Figure 2: An overview of the proposed LLM-Agent framework for data story generation.

specifications are later utilized to generate visualizations based on the relevant data tables. Here, the user intent I refers to the main idea or message that the author aims to convey, enabling them to achieve their communicative goal. In our corpus, we select report/story titles as user intents.

To this end, our goal is to develop a novel multi-agent-based approach to effectively generate the narration of a data story. To achieve this, we propose a system that uses two LLM agents – a Generator (Actor) and an Evaluator (Critic) – to mimic the human process of data story generation. This process includes a planning step that involves understanding the data (reflection), creating an outline (outline generation), and the story generation step that involves narrating the story (narration), with each step being verified and revised. We introduce a pipeline approach where the response from one LLM agent serves as the context for the next agent in the sequence. In each of the stages, the generator LLM first produces an initial version of the content, which is then assessed by the critic agent based on some fixed criteria; the generator then makes a revision based on the assessment feedback (fig. 2).

4.2 Planning Stage

Planning is crucial for all types of storytelling, particularly when it comes to data storytelling. The planning stage is divided into two intermediary steps: *(i)* Reflection, and *(ii)* Outline Generation.

- **Reflection** The goal of this stage is to understand and create a comprehensive description of the data presented in the data tables. First, the Generator Agent identifies and presents the most impactful insights, focusing on critical trends, notable patterns, and outliers that influence the overall narrative. The agent assesses the relevance, implications, and significance of the data points to determine their importance and explains the interconnections between different attributes of the data. After generating an initial reflection, the Evaluator Agent is called to verify the generation based on the data tables and asked to prepare a revision plan if necessary. At the time of verification, the Evaluator Agent cross-matches the data description with the data tables and identifies any inconsistencies and factual inaccuracies in the data description. If it determines a revision is needed, then the Generator Agent is called again to revise the initial reflection based on the revision plan. We present the prompts used at this stage in Fig. 18 - 20 in the Appendix. The whole process can be summarized as follows:

> **Input:** Data tables with titles (D), and Additional Guidelines (G).
> **Process:**
> (a) The Generator Agent generates initial reflections (R_{init}) in bullet points.
> (b) Verification: The Evaluator Agent reviews the reflection, producing a revision plan (R_{rvp}) if necessary.
> (c) Revision: The reflection is revised by the Generator Agent based on (R_{rvp}), resulting in final reflection (R_{f}).

- **Outline Generation** Once the 'reflection' is generated, the next step in the Planning stage is outlining the data story. In this step, the Generator Agent constructs an outline following a linear narrative structure (Riche et al., 2018b; Segel and Heer, 2010), consisting of a beginning, middle, and end, to ensure a coherent flow of the story. It also breaks down each major point into smaller sub-points, highlighting specific aspects of the data such as key figures, patterns, notable exceptions, and comparisons over time and including simple visualization specifications to enhance the narrative. Additionally, the user provides an 'intention' that depicts the overarching theme of the data story, and the agent is instructed to ensure that the theme is consistently emphasized throughout the outline. After generating an initial outline, the Evaluator Agent is deployed to verify the generation based on the data tables and the reflection and asked to prepare a revision plan if necessary. The agent evaluates the initial outline in two aspects, *(a)* whether the insights, trends, or outliers included in the initial outline are consistent with the data presented in the tables or not, and *(b)* whether the outline is coherent with the 'intention' or not. If it determines a revision is needed, then the Generator Agent is called again to revise the initially generated outline accordingly. We present the prompts used at this stage in Fig. 21 - 23. The whole process is summarized as follows:

> **Input:** Final reflection (R_{f}) from the previous step, data tables with titles (D), and user intention (I).
> **Process:**
> (a) The Generator Agent generates an initial outline (O_{init}) following the narrative structure.
> (b) Verification: The Evaluator Agent reviews the outline, producing a revision plan (O_{rvp}) if necessary.
> (c) Revision: The outline is revised based on (O_{rvp}), resulting in the final outline (O_{f}).

4.3 Narration Stage

The final stage of the framework is the Narration stage. The aim of this step is to generate the actual narrative text and associated visualizations. The goal is to generate a coherent data story that adheres to the narrative structure and user intention. The agent is also instructed to emphasize key statistics essential to understanding the theme, presenting them in a way that balances technical precision with accessibility thereby ensuring the story is approachable for both non-specialists and experts. Additionally, the agent is instructed to outline detailed specifications for visualizations, including chart titles, types (e.g., line, bar, pie, scatter plot), and axis data, where required by the outline. After the initial narration is generated, the Evaluator Agent assesses it to confirm its alignment with the input outline. The agent also verifies that the insights, trends, and patterns discussed are substantiated by the data tables and that the visualization specifications are factually correct. Finally, if revisions are necessary, the agent produces a revision plan. The Generator Agent then uses this plan to further refine the narration. We present the prompts used at this stage in Fig. 24 - 26. In summary:

Input: Final outline (O_f), data tables with titles (D), and user intention (I).
Process:
(a) The Generator Agent generates the initial narration (N_{init}), incorporating relevant story texts and vis-specs.
(b) Verification: The Evaluator Agent reviews the narration for factual accuracy and consistency, producing a revision plan (N_{rvp}) if necessary.
(c) Revision: Finally, the narration is revised based on (N_{rvp}), resulting in the final narration (N_f).

In each step of the framework, the LLMs are employed three times: twice for generation and once for critique. With three steps, this totals nine LLM calls. We summarize the overall working principle of the proposed agentic framework in the algorithm provided in the Appendix B.

5 Evaluation

5.1 Evaluation Methods

We employed GPT-4o (OpenAI, 2024), LLaMA-3-8b-instruct, and LLaMA-3-70b-instruct (Meta, 2024) models as the Generator and Evaluator Agents for story generation. GPT-4o was chosen for its exceptional performance across various NLP downstream tasks (OpenAI, 2024). Additionally, we utilized the leading open-source model LLaMA-3-70b-instruct and the smaller-scale option LLaMA-3-8b-instruct (Chiang et al., 2024). To generate the stories, we used the data tables from our test set which has 339 stories. To assess the efficacy of the agentic framework for story generation, we used two rigorous evaluation methods: *(i)* automatic evaluation using Gemini-1.5-pro (Team et al., 2024) as an LLM-judge and *(ii)* human evaluation.

5.2 Automatic Evaluation

Method Previous studies have found that reference-based evaluation metrics like the BLEU score often do not align with the attributes of text quality as perceived by humans (Smith et al., 2016; Liu et al., 2023). In addition, given the inherently objective nature of the story generation task, especially in data story generation, we established comprehensive methods for both automatic and human evaluations. Following the work of Zheng et al.

Model	Agentic Win (%)	Direct Win (%)	Tie (%)
GPT-4o	**78.17**	20.05	1.78
LLaMA-3-70b-instruct	**58.70**	39.82	1.48
LLaMA-3-8b-instruct	41.59	**54.57**	3.84

Table 4: An overview of the results from automatic evaluation with pairwise comparison.

(2023) and Yuan et al. (2024), we implemented an automatic evaluation method i.e., pairwise comparison of the stories generated by the agentic framework versus direct prompting. The evaluation criteria included 'Informativeness', 'Clarity and Coherence', 'Visualization Quality', 'Narrative Quality', and 'Factual Correctness'.

Results As illustrated in Table 4, the agentic framework significantly outperformed the direct approach, as demonstrated by GPT-4o, which attained an average win rate of 75.93% across three test sets, compared to the direct approach's 23.47%, highlighting a substantial difference of 52.46%. Similarly, LLaMA-3-70b-instruct using the agentic approach attained an average win rate of 58.7%, while the direct approach only achieved 39.82%. These results indicate a clear preference by the LLM judge (Gemini-1.5-pro-001 in our case) for stories generated with the agentic approach over direct prompting. However, the LLaMA-3-8b-instruct model demonstrated balanced performance with our agentic approach outperforming its counterpart in only 40.59% of cases. This outcome may be attributed to its relatively smaller size, and its limited 8k context length. These factors indicate that there is still potential for improvement through task-specific fine-tuning. Overall, these findings underscore the superior efficacy of the LLM-agent framework in producing coherent data stories.

5.3 Human Evaluation

Method For human evaluation, in line with similar research in story generation (Wang et al., 2023b; Yang et al., 2023b), we assess the stories produced by the LLMs using various subjective metrics. These metrics include 'Informativeness', 'Clarity and Coherence', 'Visualization Quality', 'Narrative Quality', and 'Factual Correctness'. We conducted a human evaluation on 100 story samples generated by the top-performing model (GPT-4o). For each sample, two annotators performed a pairwise comparison between the two versions, one gener-

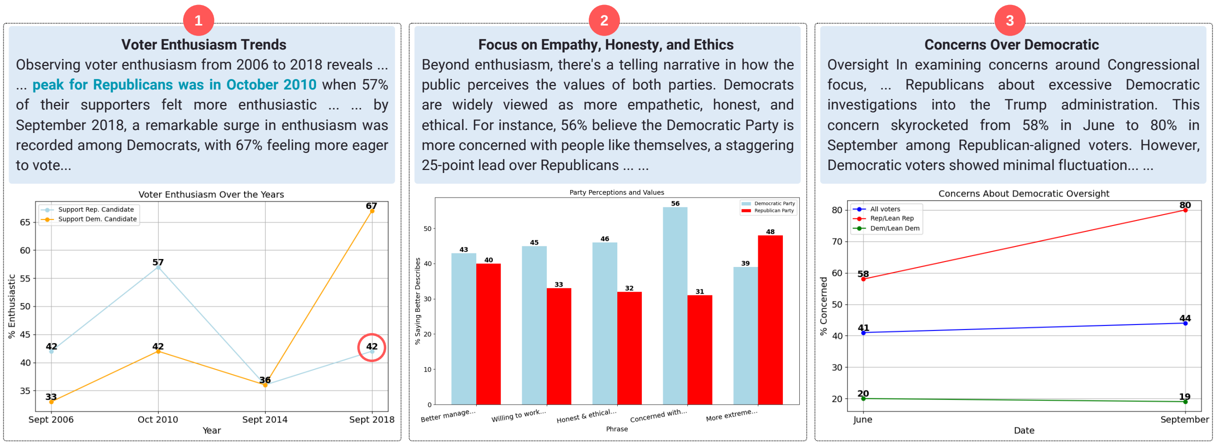


Figure 3: An example of a GPT-4o-generated story using the agentic framework: The text in **Blue** color denotes hallucinated fact, while the **red circled** value is factually incorrect according to '*Table_0*' of Fig. 13.

Metrics	GPT-4o (Agentic vs. Direct) Agentic Win (%)	Direct Win (%)	Tie (%)	p-value (sign test)
Informativeness	**74**	11	15	1.29e−12
Clarity and Coherence	**73**	11	16	2.25e−12
Visualization Quality	**59**	15	26	2.55e−07
Narrative Quality	**75**	12	13	2.71e−12
Factual Correctness	**75**	11	14	7.37e−13

Table 5: Human evaluation results of the story generation setup: GPT-4o (Agentic) vs. GPT-4o (Direct)

Planning Stage				Narration Stage	
Refl.	Refl. ver.	Out. Gen.	Out. ver.	Narr.	Narr. ver.
✓	✓	✓	✓	✓	✓
✗	✗	✓	✓	✓	✓
✓	✓	✗	✗	✓	✓
✗	✗	✗	✗	✓	✓
✓	✗	✓	✗	✓	✗

Table 6: Ablation Strategy. Here, 'Refl', 'Out.', 'Narr.', and 'Ver' denotes 'Reflection', 'Outline', 'Narration', and 'Verification' respectively

Strategy	Loss (%)	Win (%)	Tie (%)
w/o 'Reflection'	64%	35%	1%
w/o 'Outline'	64%	32%	4%
w/o 'Reflection' and 'Outline'	79%	18%	3%
w/o 'Verification'	73%	22%	5%

Table 7: The results from our ablation experiment in four different setups. We report the 'Loss', 'Win', and 'Tie' of different setups against the Agentic framework.

ated by the agentic framework and the other one by the direct prompting method, and the agreement between them for these comparisons was 85.0%.

Results The results from Table 5 indicate that the stories generated by the agentic approach are of significantly higher quality compared to those produced by the non-agentic version. This is demonstrated by an impressive average win rate of 71.2% across all five evaluation criteria. Furthermore, we compared the human-evaluated stories with our automatic evaluation and found that our human annotators agreed with the LLM judge in 67.0% of the cases, suggesting that human annotators' scores are roughly consistent with the LLM judge.

5.4 Ablation Studies

To assess the efficacy of the agentic approach, we perform ablation experiments on a randomly selected subset of 100 stories and evaluate them automatically by the LLM judge (Gemini-1.5-pro-001). These experiments focused on excluding different steps (see Table 6) and comparing the generated stories with those produced by the agentic approach.

From Table 7, we observe that The most significant decline occurred when all steps, especially when the Planning stage (Reflection and Outline Generation), were skipped (79% loss). Skipping either the Reflection or Outline Generation step also led to a decline in performance, though less severe, with a 64% loss in both cases. This demonstrates that the agentic framework's performance is roughly twice as effective as other approaches, underscoring its importance and value. Finally, omitting the verification step resulted in a 73% loss, compared to a 22% case of win, emphasizing the crucial role of the 'Critic' agent in the framework.

5.5 Error Analysis and Challenges

We manually analyzed 100 sample data stories generated by the agentic framework to understand the key challenges in addressing our new task.

Factual errors: Despite the verification steps at each stage, factual errors sometimes occur during the narration phase. For instance, the red circle in slide (1) of Fig. 3 highlights a factual error where the actual value is 59% instead of 42%, as per '*Table_0*' of Fig. 13.

Hallucination errors Although hallucinating facts is a rare occurrence in the GPT4o-generated stories using the agentic approach, some cases appear where the model is prone to hallucinating facts. For example in Fig. 3, the model mentions that 'the peak of Republican enthusiasm was in 'October 2010', whereas according to '*Table_0*' of Fig. 13 it was 'September 2018' at 59%.

Ambiguous visualization specifications In some cases, the model generates ambiguous chart specifications such as 'side-by-side bar chart,' 'multi-dimensional infographic,' 'summary chart,' or 'combined' as chart types. Such ambiguous specifications make it difficult to render charts correctly, illustrating the limitations of existing models in generating multimodal outputs with charts.

Lack of coherence and verbosity issue A key challenge faced by the open-source LLaMA-3 models is maintaining a coherent narrative structure, particularly when using the agentic approach which tends to produce more verbose text. On average, the length of stories generated by the LLaMA-3-8b-instruct model is approximately 610 tokens, while those generated using the non-agentic approach contain about 500 tokens. Fig. 14 shows that despite the story's theme being the '*EU's response to COVID-19*,' the third slide features unrelated statistics, and the fourth slide repeats text from the third. This highlights the limitations of relatively smaller open-source LLMs (8B) in producing long, multimodal stories with complex narratives.

6 Conclusion and Future Work

We present DATANARRATIVE, a new benchmark for multimodal data story generation that combines text generation, data analysis, and information visualization. Our benchmark includes 1,449 diverse data stories with open-ended multimodal outputs, each featuring various charts and related texts. We then propose an LLM-agent-based story generation framework that mimics the human process of creating data stories by using a generator and an evaluator agent. Our experiments show that this framework generally outperforms the direct method in both automatic and human evaluations.

The study also highlights unique challenges in multimodal long-form data story generation, such as the difficulty of building open-source models that generate long, coherent stories with rich narratives. To address this, we release a training corpus for the community to explore fine-tuning open-source models for this task. Additionally, our agentic framework can serve as a foundation for human-in-the-loop co-authoring of data stories with LLMs, where humans act as critics, collaborating and co-editing with the LLM to create coherent and informative stories. We hope our research inspires further work in multimodal data storytelling.

Acknowledgement

The authors would like to thank the anonymous reviewers for their helpful comments. The authors would also like to thank Mizanur Rahman for his valuable contributions to the human evaluation process. This research was supported by the Natural Sciences and Engineering Research Council (NSERC), Canada, Canada Foundation for Innovation, and the CIRC grant on Inclusive and Accessible Data Visualizations and Analytics.

Limitations

Despite the fact that the proposed agentic framework is capable of producing coherent and informative data stories, there are instances where the model may generate factually inaccurate statements within the text. Furthermore, in certain rare cases, the visualization specifications might be sufficient to create a chart image but may still lack critical information. Furthermore, because of the expense associated with API access, we were unable to assess other state-of-the-art proprietary LLMs similar to GPT-4o, such as Claude-3 (Anthropic, 2024). Due to resource constraints, we were unable to fine-tune an open-source model within the limited time available. However, we plan to release a fine-tuned model as part of our future research. Additionally, we will make the training corpus available to the community to facilitate further exploration of fine-tuning open-source models for this task.

Ethics Statement

At the time of the dataset collection process, we carefully considered various ethical aspects. The three sources of our data story corpus (Pew Research Center (Pew, 2024), Tableau Public (Tableau, 2024), and GapMinder (Rosling, 2023)) approve publication rights for academic utilization of their content. We plan to make the whole corpus and all the collected metadata publicly available.

To ensure our chart images are free of harmful content, we utilized Google search, benefiting from its rigorous content policies[1]. Moreover, during the data extraction process, the chart images were analyzed using the Gemini API, which is specifically designed to filter out unsafe content[2], thereby ensuring an additional degree of certainty concerning the appropriateness of the content included in our dataset.

The human evaluation was conducted by the authors and their collaborators associated with this research. Since the primary aim was to assess the models' capabilities, effectiveness, and limitations in generating stories across various experimental conditions, the evaluation by the authors does not introduce any ethical concerns or unwanted biases. The instructions given to the human evaluators are provided in Fig. 10. There were no paid participants in the human evaluation study. For the human evaluation study, we selected two human evaluators. The first evaluator has more than three years of industry experience (also has a graduate degree in computer science) in data science and information visualization. The second evaluator comes from an academic background (and has an undergraduate degree in computer science) and has one year of experience in information visualization. Additionally, since the evaluators were volunteers, there were no paid participants in the human evaluation study. Lastly, the evaluation did not involve any information that could be used to identify individuals.

References

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia. Association for Computational Linguistics.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. Chartcheck: Explainable fact-checking over real-world chart images.

Anthropic. 2024. Introducing the next generation of claude.

[1]https://blog.google/products/search/when-and-why-we-remove-content-google-search-results/

[2]https://ai.google.dev/docs/safety_setting_gemini

Eden Bensaid, Mauro Martino, Benjamin Hoover, and Hendrik Strobelt. 2021. Fairytailor: A multimodal generative framework for storytelling. *arXiv preprint arXiv:2108.04324*.

Tom Braude, Idan Schwartz, Alex Schwing, and Ariel Shamir. 2022. Ordered attention for coherent visual storytelling. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3310–3318, New York, NY, USA. Association for Computing Machinery.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. 2024. Autoagents: A framework for automatic agent generation.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Jinho Choi, Sanghun Jung, Deok Gun Park, Jaegul Choo, and Niklas Elmqvist. 2019. Visualizing for the non-visual: Enabling the visually impaired to use visualization. In *Computer Graphics Forum*, volume 38, pages 249–260. Wiley Online Library.

Neil Cohn. 2020. Visual narrative comprehension: Universal or not? *Psychonomic Bulletin & Review*, 27(2):266–285.

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*, SIGMOD '19, page 317–332, New York, NY, USA. Association for Computing Machinery.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english?

Yingqiang Ge, Wenyue Hua, Kai Mei, jianchao ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023. Openagi: When llm meets domain experts. In *Advances in Neural Information Processing Systems*, volume 36, pages 5539–5568. Curran Associates, Inc.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Yi He, Shixiong Cao, Yang Shi, Qing Chen, Ke Xu, and Nan Cao. 2024. Leveraging large models for crafting narrative visualization: A survey.

Xudong Hong, Rakshith Shetty, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2020. Diverse and relevant visual storytelling with scene graph embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 420–430, Online. Association for Computational Linguistics.

Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8465–8472.

Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. 2013a. Contextifier: Automatic generation of annotated stock visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2707–2716, New York, NY, USA. Association for Computing Machinery.

Jessica Hullman and Nick Diakopoulos. 2011. Visualization rhetoric: Framing effects in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2231–2240.

Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. 2013b. A deeper understanding of sequence in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2406–2415.

Md. Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024a. Mapcoder: Multi-agent code generation for competitive problem solving.

Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024b. Are large vision language models up to the challenge of chart comprehension and reasoning? an extensive investigation into the capabilities and limitations of lvlms. *arXiv preprint arXiv:2406.00257*.

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022a. OpenCQA: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11817–11837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022b. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. In *Computing Krippendorff's Alpha-Reliability*.

Deept Kumar, Naren Ramakrishnan, Richard F Helm, and Malcolm Potts. 2006. Algorithms for storytelling. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 604–610.

Bum Chul Kwon, Florian Stoffel, Dominik Jäckle, Bongshin Lee, and Daniel Keim. 2014. Visjockey: Enriching data stories through orchestrated interactive visualization. In *Poster compendium of the computation+ journalism symposium*, volume 3, page 3.

Xingyu Lan, Yang Shi, Yanqiu Wu, Xiaohan Jiao, and Nan Cao. 2022. Kineticharts: Augmenting affective expressiveness of charts in data stories with animation design. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):933–943.

Bongshin Lee, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. 2015. More than telling a story: Transforming data into visually shared stories. *IEEE Computer Graphics and Applications*, 35(5):84–90.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 598–604.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (to appear)*. Association for Computational Linguistics.

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. Chartinstruct: Instruction tuning for chart comprehension and reasoning.

S. McKenna, N. Henry Riche, B. Lee, J. Boy, and M. Meyer. 2017. Visual narrative flow: Exploring factors shaping data visualization story reading experiences. *Computer Graphics Forum*, 36(3):377–387.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. 2023. Ret-llm: Towards a general read-write memory for large language models.

OpenAI. 2023. ChatGPT. `https://chat.openai.com/`.

OpenAI. 2023a. Gpt-4 technical report. `https://openai.com/research/gpt-4`. Accessed: 2023.

OpenAI. 2023b. Gpt-4-turbo.

OpenAI. 2024. Hello gpt-4o | openai.

Jennifer J. Otten, Karen Cheng, and Adam Drewnowski. 2015. Infographics and public policy: Using data visualization to convey complex information. *Health Affairs*, 34(11):1901–1907.

OWID. 2024. Our world in data.

Xiangyu Peng, Siyan Li, Sarah Wiegreffe, and Mark Riedl. 2022. Inferring the reader: Guiding automated story generation with commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7008–7029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pew. 2024. Pew research center.

Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *arXiv preprint arXiv:2304.13620*.

Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos, and Sheelagh Carpendale. 2018a. *Data-driven storytelling*. CRC Press.

Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos, and Sheelagh Carpendale, editors. 2018b. *Data-Driven Storytelling*, 1 edition. A K Peters/CRC Press.

Tal Ridnik, Dedy Kredo, and Itamar Friedman. 2024. Code generation with alphacodium: From prompt engineering to flow engineering.

Hans Rosling. 2023. Gapminder.

Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148.

D. Shi, Y. Shi, X. Xu, N. Chen, S. Fu, H. Wu, and N. Cao. 2019. Task-oriented optimal sequencing of visualization charts. In *2019 IEEE Visualization in Data Science (VDS)*, pages 58–66, Los Alamitos, CA, USA. IEEE Computer Society.

Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. 2021a. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):453–463.

Yang Shi, Xingyu Lan, Jingwen Li, Zhaorui Li, and Nan Cao. 2021b. Communicating with motion: A design space for animated visual narratives in data videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Yang Shi, Zhaorui Li, Lingfei Xu, and Nan Cao. 2021c. Understanding the design space for animated narratives applied to illustrations. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Aaron Smith, Christian Hardmeier, and Joerg Tiedemann. 2016. Climbing mont BLEU: The strange world of reachable high-BLEU translations. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 269–281.

Nicole Sultanum and Arjun Srinivasan. 2023. Datatales: Investigating the use of large language models for authoring data-driven articles.

Mengdi Sun, Ligan Cai, Weiwei Cui, Yanqiu Wu, Yang Shi, and Nan Cao. 2023. Erato: Cooperative data story editing via fact interpolation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):983–993.

Tableau. 2024. Tableau public.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298.

Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, and Dongmei Zhang. 2017. Extracting top-k insights from multi-dimensional data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 1509–1524, New York, NY, USA. Association for Computing Machinery.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, and Jean-Baptiste Alayrac et al. 2023. Gemini: A family of highly capable multimodal models.

Gemini Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillicrap, and Jean baptiste Alayrac et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models.

Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. 2020a. Storytelling from an image stream using scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9185–9192.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 899–909, Melbourne, Australia. Association for Computational Linguistics.

Yichen Wang, Kevin Yang, Xiaoming Liu, and Dan Klein. 2023b. Improving pacing in long-form story planning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10788–10845, Singapore. Association for Computational Linguistics.

Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. 2020b. Datashot: Automatic generation of fact sheets from tabular data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):895–905.

G. Wu, S. Guo, J. Hoffswell, G. Chan, R. A. Rossi, and E. Koh. 2024. Socrates: Data story generation via adaptive machine-guided elicitation of user feedback. *IEEE Transactions on Visualization & Computer Graphics*, 30(01):131–141.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation.

Kaige Xie and Mark Riedl. 2024. Creating suspenseful stories: Iterative planning with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2391–2407, St. Julian's, Malta. Association for Computational Linguistics.

Hui Yang, Sifu Yue, and Yunzhong He. 2023a. Auto-gpt for online decision making: Benchmarks and additional opinions.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023b. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Licheng Yu, Mohit Bansal, and Tamara Berg. 2017. Hierarchically-attentive RNN for album summarization and storytelling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 966–971, Copenhagen, Denmark. Association for Computational Linguistics.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. 2023. Recurrentgpt: Interactive generation of (arbitrarily) long text.

Appendices

A Dataset Construction Process

In this section, we provide further detail on our dataset curation process.

A.1 Data Sources

The corpus for DATANARRATIVE consists of stories collected from three different platforms: Pew Research Pew (2024), Tableau Public Data Story Tableau (2024), and Gapminder (Rosling, 2023). Pew Research releases articles based on data that focus on social issues, public opinion, and demographic trends. These articles frequently include various charts and are complemented by high-quality descriptions from professional editors. Gapminder is a Swedish foundation dedicated to fighting misconceptions about global development by promoting a fact-based worldview. They provide interactive data visualization tools and publish educational resources, such as data stories, and interactive visualizations that emphasize global trends in health, wealth, and other development indicators. On the other hand, Tableau Public Story, a feature of Tableau Public, is a platform that enables users to create interactive presentations through a series of data visualizations. It makes data stories publicly accessible, covering a wide range of topics including economy, social issues, and international affairs. Therefore, the corpus benefits from this diversity by providing stories with varying topics, styles, and themes.

A.2 Raw Data Collection

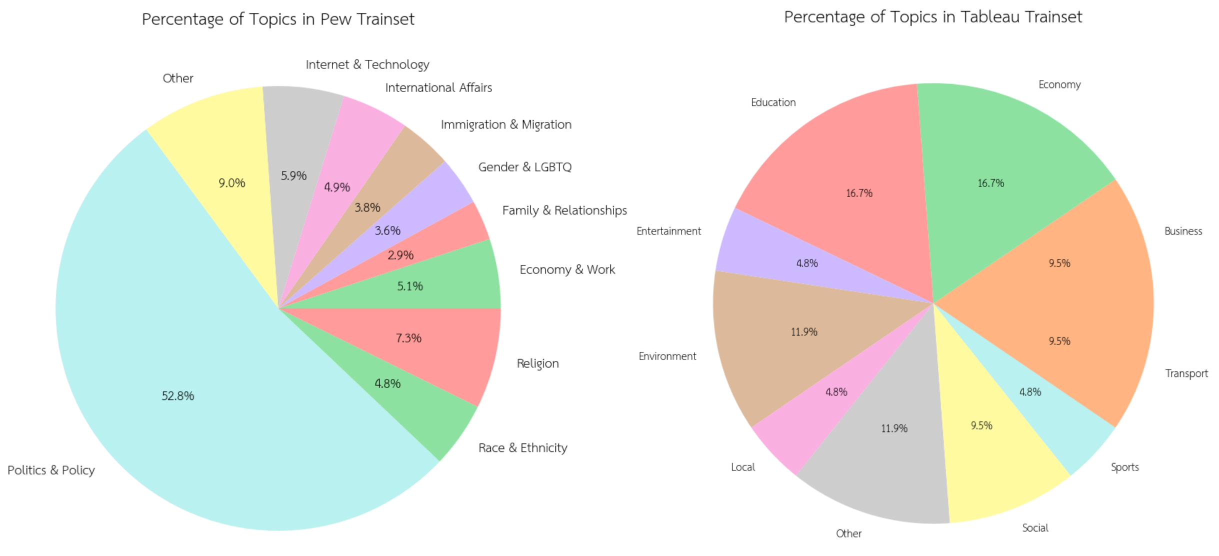
To assemble the Pew corpus, we created a web crawling script that initially stores research topics and their corresponding URLs. This script systematically processes the HTML elements from these URLs to collect all links, categorizing them under general topics while excluding irrelevant ones like "Methodological Research" and "Full topic list" that do not link to any meaningful article webpage. Subsequently, another script is employed to visit all the article pages for each topic, extracting and parsing HTML content to gather various data such as article texts, titles, and image links. These image links are then filtered by specific criteria (e.g., 'jpg', 'jpeg', 'SVG', or 'png' formats) to ensure data integrity, eliminating duplicates. A secondary script downloads these images in 'PNG' format. We gathered articles from the Pew Research website until March 14, 2024, resulting in 4532 articles across 18 topics. Additionally, we collected metadata related to the images, including captions and alt-texts.

A.3 Chart Data Extraction

We utilize the multi-modal large language model (MLLM) Gemini-1.0-pro-vision (Team et al., 2023) to extract data from chart images. In order to verify the factual correctness of the generated data tables, we conducted a small experiment using 100 chart images from the ChartQA (Masry et al., 2022) corpus, where gold tables were already available, allowing for direct comparison between the gold tables and the generated tables. We performed a human evaluation of the generated data tables and found that the model correctly generated the tables in 77% of the cases. Most errors occurred when the model either produced incomplete tables (missing one or two values or an entire row) or failed to generate any output at all. Fig. 5 presents an overview of the chart data extraction process.

A.4 Chart-text pair Collection

As the Pew corpus is larger than the other corpora, collecting paragraphs associated with the data tables manually is labor-intensive and time-consuming. Therefore, for the Pew training set, we adopted an automatic approach using the GPT-4-turbo model (OpenAI, 2023b). The model selected relevant paragraphs from articles based on data tables for the chart images that we extracted automatically. In addition to collecting the original paragraphs, we also generated the paraphrased version of the paragraphs using the GPT-4-turbo model as well. To evaluate the effectiveness and accuracy of this approach, we compared human-curated paragraphs from Pew articles with those selected by GPT-4-turbo. By examining 50 randomly selected samples from the Chart-to-Text corpus, we found that GPT-4-turbo accurately linked paragraphs to data tables 70% of the time. As a result, we decided to use GPT-4-turbo-generated paragraphs for the Pew training set. To create the test set from the Pew corpus, we selected the articles and the paragraph-table pairs from each of the articles that appear in the Chart-to-Text (Kantharaj et al., 2022b) Pew corpus. Fig. 6 illustrates an overview of the chart-text collection process.



(a) Pew

(b) Tableau

Figure 4: The figure demonstrates the distribution of Story Topics in the Train set.

A.5 Detailed Corpus Analysis

In this section, we present a more fine-grained analysis of the proposed dataset for DATANARRATIVE.

- **Pew** The Pew training corpus includes 1,068 stories, encompassing a total of 4,729 tables and 4,729 charts. On average, the length of these stories is 1,804 characters, consisting of an average of 353 tokens and organized into on average 4 paragraphs per story. The vocabulary-to-token ratio averages 0.51, with each story typically featuring 14 unique verbs, and 44% of these verbs are diverse. Trigram repetition within stories stands at 18.37%, while between stories it is 14.83%. From Table 2 we observe that in the Pew train set, a significant majority of the charts are bar charts (both simple as well as stacked and group bar charts) (83.51%), followed by line charts (9.16%), and pie charts (4.04%), etc. Regarding topic variety, 51.84% of the stories focus on 'Politics & Policy', 7.17% on 'Religion', and 5.79% on 'Internet & Technology', among other categories.

The Pew test corpus comprises a total number of 321 stories, with a total of 1590 tables and 1590 charts. The average length of stories in the train set is 2865 characters, the average token count is 561 and there are 5 paragraphs in each sample story on average. Additionally, the average vocabulary-to-token ratio is 0.46, with an average of 23 unique verbs per story, and 47% of the verbs used are diverse. The intra-story trigram repetition rate is 17.94%, while inter-story trigram repetition is 11.28%. Similarly, Table 2 indicates that in the Pew test set, the majority of the charts are bar charts (simple, stacked, and group) at 77.79%, followed by line charts at 17.45%, and pie charts at 3.56%. Regarding topic diversity, about 71.96% of the stories are related to 'Politics & Policy', 8.09% to 'International Affairs', and 5.29% to 'Internet & Technology'.

- **Tableau** The training corpus for Tableau consists of 42 stories with a total of 340 tables and 297 charts. Each story in the training set averages 837 characters, 159 tokens, and 5 paragraphs. The vocabulary-to-token ratio averages 0.64, and each story typically includes 5 unique verbs, with 25% of them being diverse. The percentage of intra-story trigram repetition is 12.79% and inter-story trigram repetition is 0.64%. The Tableau test corpus consists of 13 stories, with 64 tables and 64 charts. From Table 2 we can see that bar charts are the most common chart type in the Tableau train set, accounting for 52.19% of all charts. They are followed by line charts (23.23%) and scatter plots (12.12%). In terms of topic diversity, approximately 16.67% of the stories are about the 'Economy', followed by 'Education' (16.67%) and the 'Environment' (11.9%), among others.

In the test set, the average story length is 1009 characters, the average token count is 194, and each story contains an average of 4 paragraphs. Additionally, the vocab: token ratio is 0.63, the average number of unique verbs per story is 11, and 30% of the verbs in a story are diverse. The percentage of intra-story trigram repetition is 14.24%,

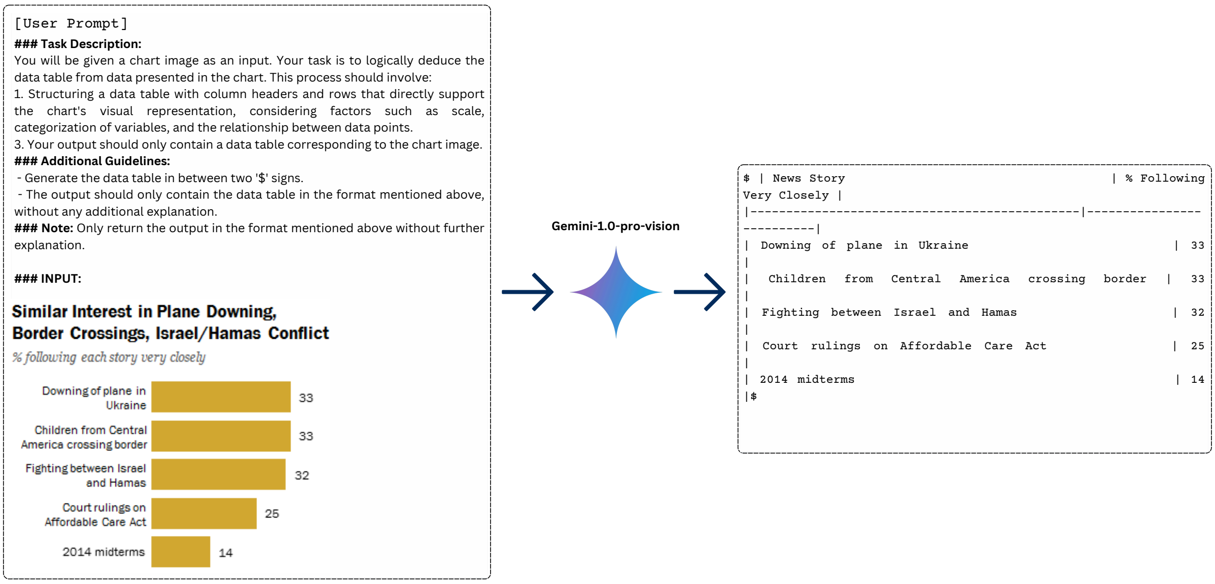


Figure 5: The figure presents an overview of the Chart data extraction process using the Gemini-1.0-pro-vision (Team et al., 2023) model.

and the percentage of inter-story trigram repetition is 44.67%. Similarly, regarding the charts in the Tableau test set, Table 2 shows that bar charts (simple, stacked, and grouped) comprise the majority (71.88%), followed by line charts (12.5%) and scatter plots (9.37%). In terms of topic diversity, approximately 30.77% of the stories are about the 'Economy', followed by 'Education' (15.38%) and the 'Environment' (7.69%), among others.

• **Gapminder** The GapMinder test corpus consists of five stories, with a total of 42 tables and 42 charts. The average length of stories in the train set is 707 characters, and there are 8 paragraphs in each sample story on average. The average token count is 146. Additionally, the average vocab: token ratio is 0.63, the average number of unique verbs per story is 5, and there are 39% of diverse verbs present in a story. Furthermore, the percentage of intra-story trigram repetition is 11.3% and inter-story trigram repetition is 2.45%. From Table 2 we observe that the Gapminder dataset mainly focuses on topics such as 'World Population', 'World Economy', and 'Population Birthrate'. The dataset only consists of line charts (73.81%) and bubble charts (26.19%).

In addition, Fig. 4 and Fig. 7 detail the overall topic distribution in the train and test set respectively. Furthermore, Fig. 8 and Fig. 9 show the distributions of Charts / Tables per Story in the Pew train and test set respectively.

B LLM Agent Framework

We summarize the whole working process of the proposed agentic framework in the Alg. 1:

C Additional Results and Evaluation Details

In this section, we detail our human evaluation approach and present a detailed result analysis (see Fig. 10)

Human Evaluation Our human evaluation metrics include 'Informativeness', 'Clarity and Coherence', 'Visualization Quality', 'Narrative Quality', and 'Factual Correctness'. Below we present the description of the metrics:
(a) Informativeness: The extent to which the data story provides substantial and useful information.
(b) Clarity and Coherence: The logical organization, ease of understanding, and connectivity between different parts of the data story.
(c) Visualization Quality: The effectiveness of visualization, i.e., charts in enhancing understanding of the data.
(d) Narrative Quality: The ability of the narrative to engage the reader and provide deep insights.
(e) Factual Correctness: The accuracy of the data and information presented.

We assessed each story using two human annotators for each evaluation criterion. For every story, we presented two versions—one generated using the Agentic framework and the other using the Direct prompting method—without disclosing which

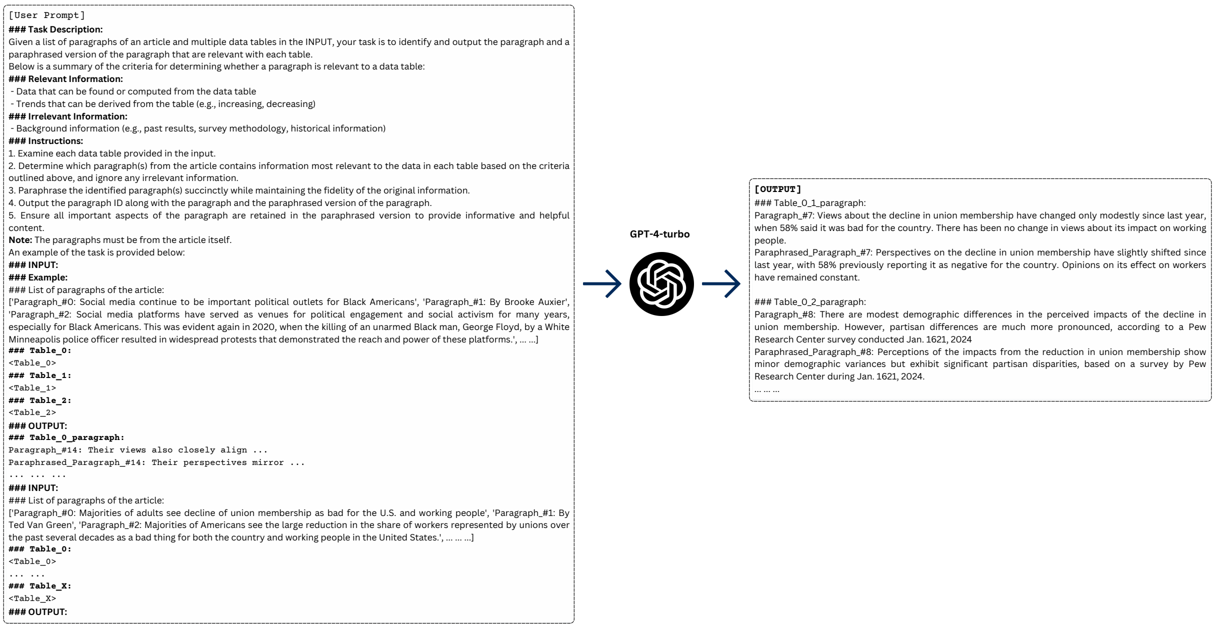


Figure 6: The figure presents an overview of the Paragraph table pair generation using the GPT-4-turbo (OpenAI, 2023b) model.

Input: Data tables with titles D, Additional Guidelines G, Intention I
Output: Final narration N_f
$R_0 \leftarrow \text{Generate}(D, G)$; `// Generate initial reflection`
$V_R \leftarrow \text{Verify}(D, R_0)$; `// Verify reflection`
$R_f \leftarrow \text{Revise}(R_0, V_R)$; `// Revise reflection`
$O_0 \leftarrow \text{Generate}(R_f, D, I)$; `// Generate initial outline with intention`
$V_O \leftarrow \text{Verify}(D, R_f, O_0)$; `// Verify outline`
$O_f \leftarrow \text{Revise}(O_0, V_O)$; `// Revise outline`
$N_0 \leftarrow \text{Generate}(O_f, D, I)$; `// Generate initial narration with intention`
$V_N \leftarrow \text{Verify}(D, O_f, N_0)$; `// Verify narration`
$N_f \leftarrow \text{Revise}(N_0, O_f, V_N, I)$; `// Revise the narration (if necessary) and generate the final version`

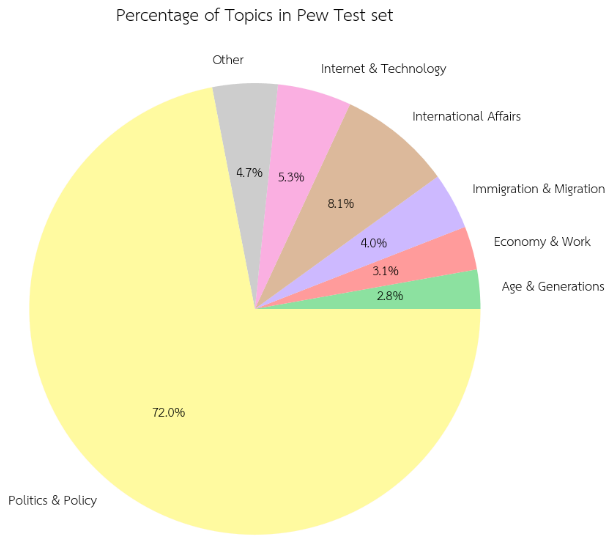
Algorithm 1: Data Story Generation Framework

version was which. The annotators were then asked to determine which version was superior based on each criterion. In cases where the annotators disagreed, we considered the result as a tie. We measured Krippendorff's alpha (Krippendorff, 2011) to determine inter-annotator agreement and found a moderate level of agreement (0.505%) between the annotators.

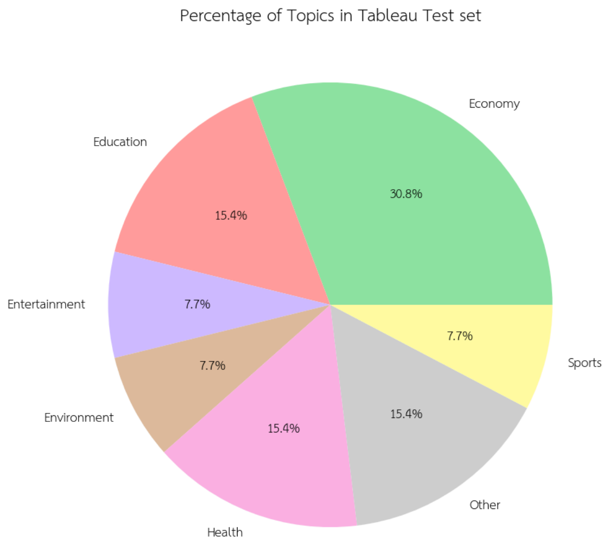
Results In this section, we present a detailed breakdown of the performance of the agentic framework against the direct prompting strategy across the different test sets. Table 8 presents the detailed results from the experiments. We also present our ablation study strategy in Table 6.

D Additional Error Analysis

In this section, we present examples of errors that occurred in the generated stories. For instance, Fig. 12 illustrates a story generated by the LLaMA-3-8b-instruct model where factual errors are in '**Section 2**' where it mentions '*average approval rating for presidents in the third year is 55%*' according to the '**Table_#0**' in the figure, however, it is actually less than 55% (the average is 53.8%). Furthermore, we found that most factual error occurs in the '*Visualization Specifications*' as exemplified by Fig. 15. Additionally, hallucinating data values is another concern at the time of narration generation, even though verification steps are included at each stage of the agentic framework. One such case is illustrated in Fig. 12, where the LLaMA-3-8b-instruct model hallucinated facts



(a) Pew



(b) Tableau

Figure 7: The figure demonstrates the distribution of Story Topics in the Test set.

Model	Pew				Tableau				Gapminder			
	Samples	Agentic Win (%)	Direct Win (%)	Tie (%)	Samples	Agentic Win (%)	Direct Win (%)	Tie (%)	Samples	Agentic Win (%)	Direct Win (%)	Tie (%)
GPT-4o	321	78.50 252	19.63 63	1.87 6	13	69.23 9	30.77 4	0 0	5	80.00 4	20.00 1	0 0
LLaMA-3-8b-I	321	40.81 131	55.45 178	3.74 12	13	53.85 7	38.46 5	7.69 1	5	60 3	40 2	0 0
LLaMA-3-70b-I	321	58.25 187	40.19 129	1.56 5	13	69.23 9	30.77 4	0 0	5	60 3	40 2	0 0

Table 8: Automatic Evaluation results of generated stories (Agentic vs. Non-agentic) with pairwise additive prompting. Here, 'I' in 'LLaMA-3-Xb-I' stands for Instruction tuned versions, and 'Agentic' and 'Direct' stands for Agentic framework and Direct prompting strategy respectively. We calculate the % of wins for these two different strategies and report them in this table. The Gray text indices the number of samples for each case.

such as '*Trump's presidency has been marked by low approval ratings throughout his term*', whereas the data in the table only gives a picture of first three years. Similar to the factual errors, most of the hallucinations are prevalent in the '*Visualization Specifications*' like Fig. 15.

E Examples

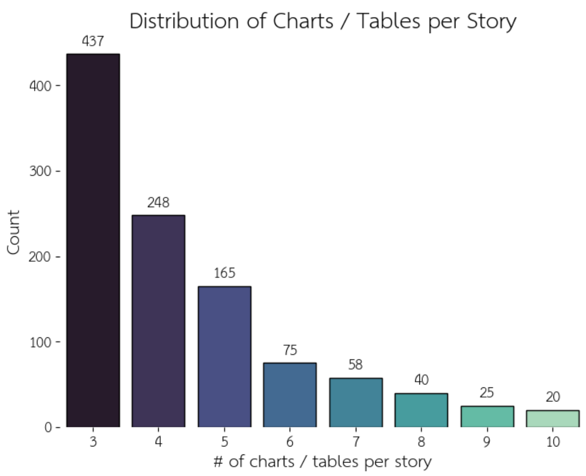


Figure 8: Distribution of # of charts / tables per story (Pew Train).

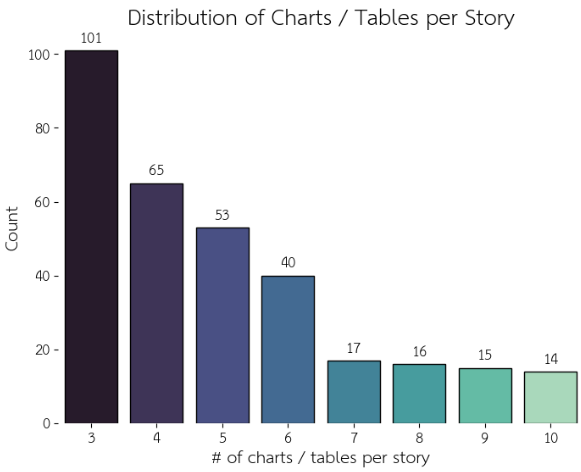


Figure 9: Distribution of # of charts / tables per story (Pew Test).

Human Evaluation Instruction:
Review the provided two versions of a data story based on the evaluation criteria mentioned below:
Evaluation Criteria:
1. **Informativeness:** The extent to which the data story provides substantial and useful information.
2. **Clarity and Coherence:** The logical organization, ease of understanding, and connectivity between different parts of the data story.
3. **Visualization Quality:** The effectiveness of visualization, i.e., charts in enhancing understanding of the data.
4. **Narrative Quality:** The ability of the narrative to engage the reader and provide deep insights.
5. **Factual Correctness:** The accuracy of the data and information presented.

For each of the abovementioned criteria, rate the data story on a scale of 1 to 5, where 1 is the worst quality and 5 is the best quality. Here, user `intention' refers to the title of the story
User Intention: <Input intention → The article title of sample the gold test set>
After reviewing both data stories (Story A and Story B), evaluate which version of each story excels in the specific criteria. Conclude by providing a final verdict on which story is overall superior.
Informativeness: [story version]
Clarity and Coherence: [story version]
Visualization Quality: [story version]
Narrative Quality: [story version]
Factual Correctness: [story version]
Final Verdict: [story version]

Figure 10: Instruction for our Human Evaluation settings.

```
Automatic Evaluation Prompt:
### Task Description:
You will receive:
- A plausible gold data story as a reference
- A user intention representing the overarching theme of the story
- Data tables used to generate the data story
- Two model-generated stories
Ignore any extra white spaces and newlines in the stories. Your task is to evaluate the quality
of the LLM-generated stories based on the criteria listed below:

### Evaluation Criteria:
1. **Relevance and Informativeness:** The extent to which the data story addresses the given
user `intention` and provides substantial and useful information.
2. **Structure and Coherence:** The logical organization such as a linear narrative structure
(a beginning, a middle and a conclusion), ease of understanding, and connectivity between
different parts of the data story.
3. **Visualization Specification Quality:** The visualization specifications defined within
`<visualization>` tags are well-suited for creating visualizations that enhance the
understanding of the data.
4. **Narrative Quality and Insightfulness:** The ability of the narrative to engage the reader,
provide important insights, and follow the `intention` provided by the user.
5. **Factual Correctness:** The accuracy of the data and information presented considering the
input data tables.
### Point Allocation Criteria:
1. For each evaluation criterion, give 1 point to 'Story A' if it is better than 'Story B', or
vice versa.
2. If both stories perform equally well in a criterion, give 1 point to both.
3. Evaluate the stories based on their total points.
### Additional Guidelines:
 - Systematically attribute points to `Story A` and `Story B` based on the `Point Allocation
Criteria`.
 - Make sure total accumulated points for each story is within a range of 1 to 5.
 - Briefly justify your total score, up to 100 words.
 - Avoid any position biases and ensure that the order in which the stories were presented does
not influence your decision.
 - Do not allow the length of the stories to influence your evaluation.
 - Be as objective as possible.
 - Remember to assess the data story from the perspective of relevance, clarity, coherence,
informativeness, and factual correctness, taking the plausible gold story as a reference.
 - After providing your explanation, output your final verdict based on the total points each
story received by strictly following this format: '[[A]]' if the story A is better, '[[B]]' if
the story B is better, and '[[C]]' for a tie.
### INPUT:
<intention>
### Gold Story:
<gold_story>
### Story A:
<story_a>
### Story B:
<story_b>
```

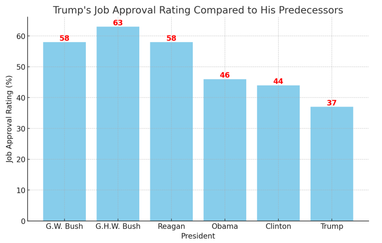
Figure 11: Pairwise automatic evaluation prompt.

```
### Table_0:
At start of Trump's third year in office, his job approval lags most of his recent predecessors
Presidential job approval at beginning of third year in office (%)
| President        | Approval Rating (%) |
|------------------|---------------------|
| G.W. Bush ('01-Jan. '03)  | 58                 |
| G.H.W. Bush ('89-Jan. '91)  | 63                 |
| Reagan ('81-Jan. '83)       | 58                 |
| Obama ('09-Jan. '11)        | 46                 |
| Clinton ('93-Feb. '95)      | 44                 |
| Trump ('17-Jan. '19)        | 37                 |
Source: Survey data from Pew Research Center (Trump through Clinton) and the Gallup organization (G.H.W. Bush and Reagan). Current data on Trump approval from survey of
U.S adults conducted Jan. 9-14, 2019.
PEW RESEARCH CENTER
```

1

Trump Begins Third Year With Low Job Approval and Doubts About His Honesty

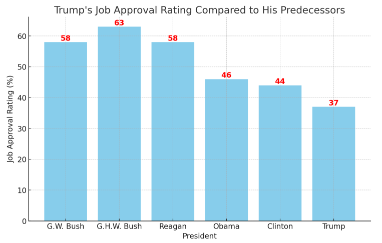
Trump's approval rating at the start of his third year in office stands at 37%, lower than most of his recent predecessors (Table_0). This is not a new phenomenon, as Trump's presidency has been marked by low approval ratings throughout his term.



2

Trump's Job Approval Rating Lags Behind Most Recent Predecessors

It is important to consider the historical average of job approval ratings for presidents in their third year in office. According to Gallup, the average approval rating for presidents in their third year is 55%. Trump's approval rating of 37% ... ...



•••

6

Conclusion

In conclusion, Trump's third year in office has been marked by low job approval ratings and doubts about his honesty. ... ... Trump faces significant challenges in his presidency, and that he will need to find ways to overcome these challenges in order to achieve his policy goals.

Figure 12: Examples of Factual and Hallucination errors in LLaMA-3-8b-instruct generated story using the Agentic framework. Here, **Blue** color denotes hallucinated text, and **Red** color denotes text containing factual errors.

```
### Table_0: Voter Enthusiasm Rises – Especially Among Democrats
% of registered voters who say they are more enthusiastic than usual about voting:
| Year       | Support Rep. Candidate | Support Dem. Candidate |
|-----------|------------------------|-----------------------|
| Sept 2006 | 42                     | 33                    |
| Oct 2010  | 57                     | 42                    |
| Sept 2014 | 36                     | 36                    |
| Sept 2018 | 59                     | 67                    |
### Table_1: Democrats lead the GOP on empathy, honesty and ethics; neither party has edge on managing govt.
% saying the phrase ___ better describes the...
| Phrase                                              | Democratic Party | Republican Party | Difference |
|-----------------------------------------------------|------------------|------------------|------------|
| Better manage government                            | 43               | 40               | D+3        |
| More willing to work with political leaders from the other party | 45              | 33               | D+12       |
| Governs in more honest & ethical way                | 46               | 32               | D+14       |
| Concerned with people like me                       | 56               | 31               | D+25       |
| More extreme in its positions                       | 39               | 48               | R+9        |
### Table_2: % of registered voters who say they are concerned that...
| Concern / Group                                  | Very | Somewhat | Not too | Not at all |
|--------------------------------------------------|------|----------|---------|------------|
| **If Democrats take control of Congress,**                                            |
|     they   will   focus   too   much   on   investigating   the   Trump   administration
|
| All voters (June)                               | 41   | 34       | 15      | 9          |
| All voters (Sept)                               | 44   | 31       | 15      | 9          |
| Rep/Lean Rep (June)                             | 58   | 22       | 12      | 7          |
| Rep/Lean Rep (Sept)                             | 80   | 9        | 7       | 4          |
| Dem/Lean Dem (June)                             | 20   | 27       | 35      | 18         |
| Dem/Lean Dem (Sept)                             | 19   | 21       | 34      | 25         |
| **If Republicans keep control of Congress,**                                        |
|     they   will   not   focus   enough   on   oversight   of   the   Trump   administration
|
| All voters (June)                               | 40   | 31       | 12      | 16         |
| All voters (Sept)                               | 40   | 34       | 11      | 15         |
| Rep/Lean Rep (June)                             | 22   | 20       | 27      | 29         |
| Rep/Lean Rep (Sept)                             | 20   | 26       | 24      | 29         |
| Dem/Lean Dem (June)                             | 59   | 20       | 9       | 9          |
| Dem/Lean Dem (Sept)                             | 66   | 13       | 10      | 10         |
```

Figure 13: The tables corresponding to Fig. 3.

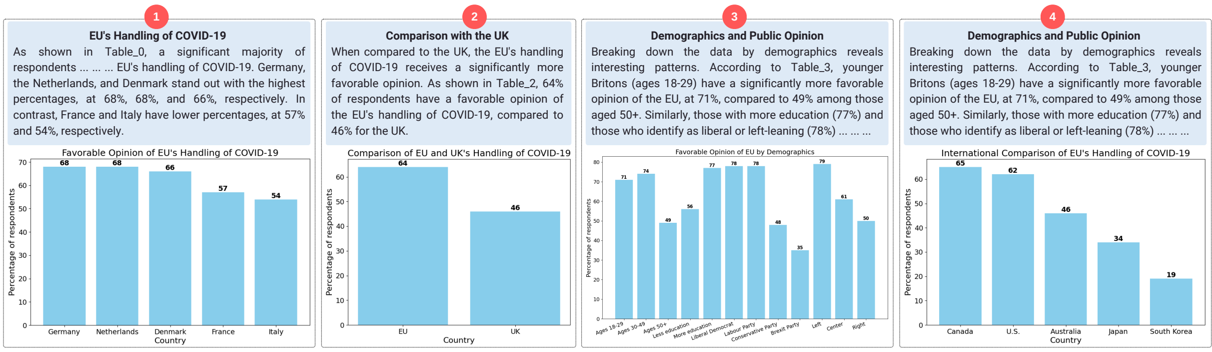


Figure 14: A figure demonstrating the 'Coherence' issue of the LLaMA-3-8b model.

[Data Tables]

Table_#1:

Public widely views cyberattacks as a major threat to the United States: % who say that each is a __ to the well-being of the United States | Threat | Major threat | Minor threat | Not a threat | |--|--------------|--------------|--------------| | Cyberattacks from other countries | 74 | 21 | 4 | | Iran's nuclear program | 57 | 31 | 8 | | Global climate change | 57 | 23 | 18 | | China's power and influence | 54 | 32 | 11 | | The Islamic militant group, known as ISIS | 53 | 35 | 8 | | North Korea's nuclear program | 53 | 36 | 9 | | Russia's power and influence | 50 | 38 | 10 |

...

...

Table_#4:

Far more Republicans than Democrats say Russia is 'not much of a problem' % who think of Russia as... Total | Adversary | Serious problem | Not a problem July 2019 | 26 | 44 | 26 May 2018 | 28 | 42 | 24 Jan 2017 | 29 | 44 | 24 Oct 2016 | 23 | 46 | 23 April 2016 | 23 | 44 | 30 Among Rep/Lean Rep | Adversary | Serious problem | Not a problem July 2019 | 20 | 41 | 38 May 2018 | 25 | 39 | 33 Jan 2017 | 20 | 45 | 32 Oct 2016 | 25 | 42 | 28 April 2016 | 27 | 41 | 29 Among Dem/Lean Dem | Adversary | Serious problem | Not a problem July 2019 | 34 | 49 | 14 May 2018 | 30 | 47 | 17 Jan 2017 | 38 | 42 | 17 Oct 2016 | 25 | 51 | 18 April 2016 | 20 | 47 | 29 Note: Don't know responses not shown. Source: Survey of U.S adults conducted July 10-15, 2019 PEW RESEARCH CENTER

[LLM Generated Story]

Climate Change and Russia Are Partisan Flashpoints in Public's Views of Global Threats

The Pew Research Center conducted a survey of U.S. adults from July 10-15, 2019, ...

Visualization 1: Cyberattacks as a Major Threat to the United States

* Chart title: Cyberattacks as a Major Threat to the United States
* Chart type: Line chart
* X-axis: Years (2009-2019)
* Y-axis: Percentage of Americans viewing cyberattacks as a major threat
* **Data values: 2009 (34%), 2011 (34%), 2013 (35%), 2015 (35%), 2017 (35%), and 2019 (74%)**

<visualization>

{"chart_title": "Cyberattacks as a Major Threat to the United States",
"chart_type": "line",
"x_axis": "Years (2009-2019)",
"y_axis": "Percentage of Americans viewing cyberattacks as a major threat",
"**data_values": [34, 34, 35, 35, 35, 74]}**

</visualization>

Partisan Divide

The data in Table_2 shows a significant partisan divide in ... ...

<visualization>

Visualization 2: Partisan Divide on Russia

* Chart title: Partisan Divide on Russia
* Chart type: Bar chart
* X-axis: Political party (Dem/Lean Dem, Rep/Lean Rep)
* Y-axis: Percentage of Americans viewing Russia as a serious problem or adversary
* **Data values: Dem/Lean Dem (49%, 47%), Rep/Lean Rep (41%, 39%)**

<visualization>

{"chart_title": "Partisan Divide on Russia",
"chart_type": "bar",
"x_axis": "Political party (Dem/Lean Dem, Rep/Lean Rep)",
"y_axis": "Percentage of Americans viewing Russia as a serious problem or adversary",
"**data_values": [[49, 47], [41, 39]]}**

</visualization>

Figure 15: Examples of Factual and Hallucination errors in LLaMA-3-8b-instruct generated story using the Agentic framework. Here, **Blue** color denotes hallucinated text, and **Red** color denotes text containing factual errors.

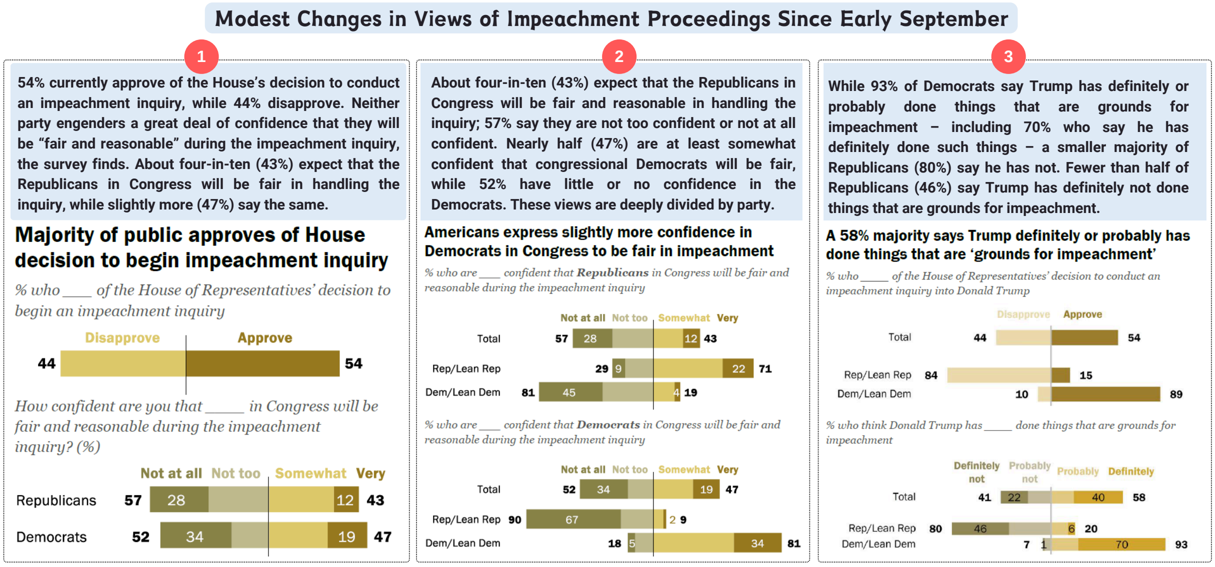


Figure 16: An example data story in our corpus collected from Pew (Pew, 2024).

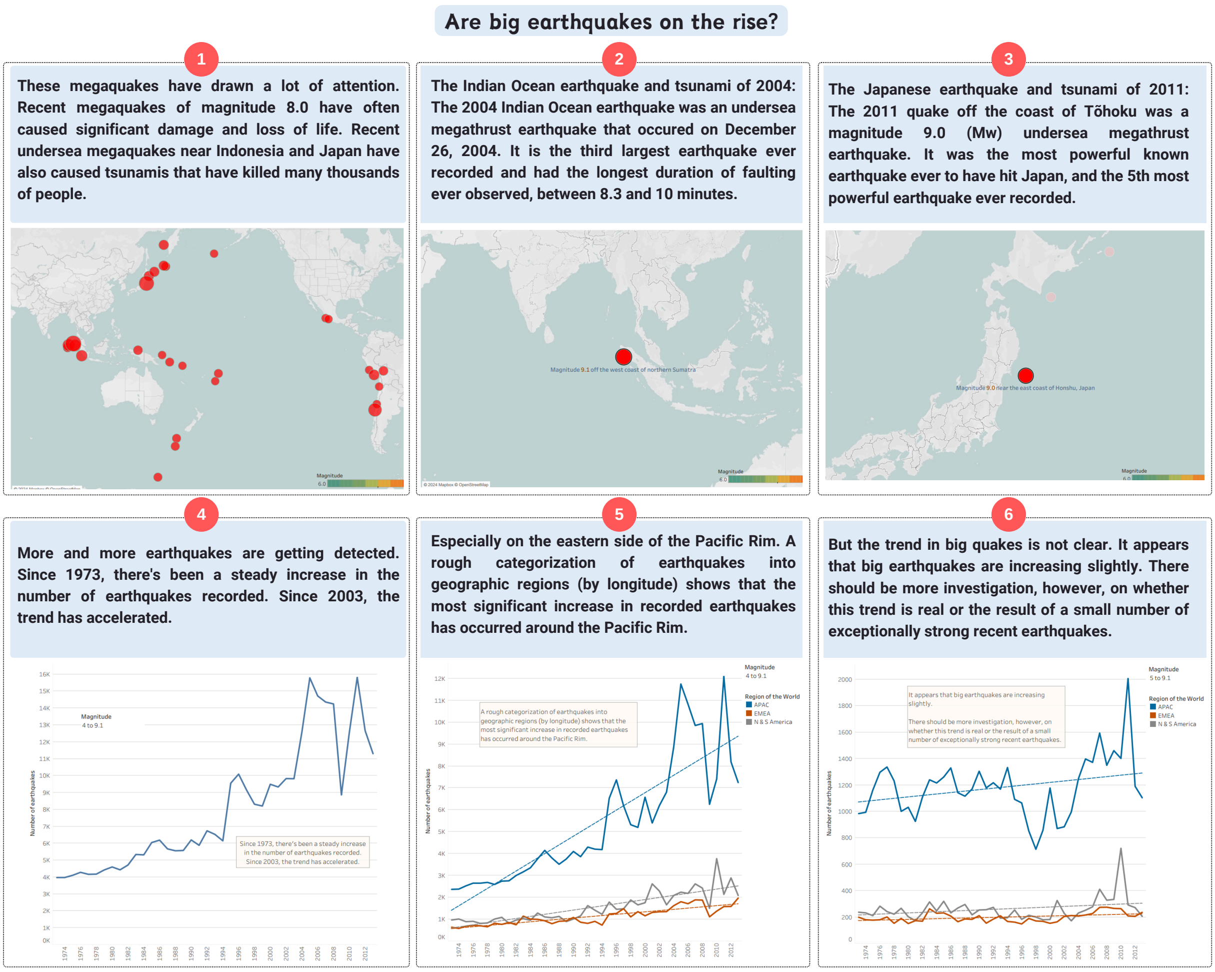


Figure 17: An example data story in our corpus collected from Tableau (Tableau, 2024).

[System Prompt]
As an intelligent data analyst and insight extraction specialist, your role is to generate a 'reflection' from data tables that must cover every important detail that can be observed in the data tables. Pay attention to small details and nuances as well as any trends or outliers in the given tables.

[User Prompt]
Task Description:
Given the data tables corresponding to a data story in the input, your task is the following:
1. Generate a coherent 'reflection' on the data tables given in the input, in bullet points. Here, 'reflection' is defined as the systematic examination and interpretation of data tables to narrate a coherent story, involving a comprehensive understanding of the data structure, identification of key variables, analysis of data distribution and trends, and understanding of the data's broader context.
2. Identify and discuss the most impactful insights from the data tables. Focus on elements that significantly influence the narrative or findings, such as critical trends, notable patterns, and significant outliers.
3. Factual accuracy in the data description is of utmost importance, so review the data tables carefully and thoroughly.
4. Determine the importance of details based on their relevance to the overall story, potential implications, and their statistical significance.
5. Explain how different attributes of the data tables are interconnected. Highlight any causal relationships, correlations, or patterns that emerge from the data.
6. Discuss any observed trends or outliers, explaining their potential implications or causes.

Additional Guidelines:
- The output must be in plain text and structured in bullet points.
- Think step by step and generate the response 'reflection' in between two <reflection> tags.

INPUT:
Tables:
<Tables>

Figure 18: The figure presents the prompt used to generate the initial 'Reflection'.

[System Prompt]
As an analytical critic, your role is to meticulously examine the alignment between data presented in tables and the narrative provided in a reflection. Focus on identifying any discrepancies and factual inaccuracies in the details. Consider not just the numbers but also the context and implications of the data.

[User Prompt]
Task Description:
Given the data tables and a reflection corresponding to a data story in the input, your task is the following:
1. Carefully analyze the data tables and the reflection. Identify any discrepancies or inconsistencies, focusing on numerical data, contextual interpretations, and the reflection's fidelity to the data. Discrepancies might include but are not limited to incorrect data interpretation, or overlooked details.
2. Factual correctness of the data is of utmost importance, so review the data tables and the given `reflection` carefully and thoroughly, and include instructions for necessary corrections.
3. Based on your analysis, draft a revision plan to refine the reflection if needed, and output the revision plan. Otherwise just output: 'No revision needed'.
4. The revision plan if needed must coherently and logically relate to the attributes of the data.
5. Be as specific as possible.

Additional Guidelines:
- The output must be in plain text and structured in bullet points.
- Think step by step and generate the response 'reflection' in between two <reflection> tags.

INPUT:
Tables:
<Tables>
Reflection:
<reflection>

Figure 19: The figure presents the prompt used to generate the 'Reflection' revision plan.

[System Prompt]
As an intelligent data analyst and insight extraction specialist, your role is to generate a 'reflection' from data tables that must cover every important detail that can be observed in the data tables. Pay attention to small details and nuances as well as any trends or outliers in the given tables.

[User Prompt]
Task Description:
Given the data tables corresponding to a data story and a revision plan for reflection in the input, your task is the following:
1. Revise the reflection according to the revision plan. Pay attention to small details and nuances and any trends or outliers in the given tables.
2. Factual accuracy in the data description is of utmost importance, so review the data tables carefully and thoroughly.
3. The generated reflection must coherently and logically relate to the attributes of the data.
4. Be as specific as possible.

Additional Guidelines:
- The output must be in plain text and structured in bullet points.
- Think step by step and generate the response 'reflection' in between two <reflection> tags.

INPUT:
Tables:
<Tables>
Previous Reflection:
<reflection>
Revision Plan:
<reflection_revision_plan>

Figure 20: The figure presents the prompt used to generate the revised 'Reflection'.

[System Prompt]
You are an expert at generating outlines for data stories. The generated outline should cover every important detail that can be observed in the data tables. Pay attention to small details and nuances as well as any trends or outliers in the given tables.

[User Prompt]
Task Description:
Given a reflection and the data tables corresponding to a data story in the input, you have the following tasks:
1. Generate an outline of the story following a linear narrative structure considering the reflection and the data presented in the tables. A linear narrative structure is defined as the narrative structure that contain a start (introduction), a middle, and an end (conclusion). Think of it as setting the scene, unveiling the adventure, and wrapping up with a satisfying conclusion.
2. The data story's overarching theme should focus on ***<intention>***. Make sure this theme is consistent throughout the outline.
3. Each of the points in the outline, break it down into sub-points that spotlight specific aspects of the data. This could include: significant figures or patterns, noteworthy exceptions or deviations, comparisons or changes over time. Add instructions for visualizations, i.e., charts, where necessary.
4. Remember, the essence of a compelling data story is not just in the numbers but in how you tell the tale, so inclusion of visualization instruction is of utmost importance.
5. The generated outline must coherently and logically relate to the attributes of the data and rigourously follow the theme. Be as specific as possible.

Additional Guidelines:
- The output must be in plain text and structured in bullet points.
- Think step by step and generate the response outline in between two <outline> tags.

INPUT:
Tables:
<Tables>
Reflection:
<final_reflection>

Figure 21: The figure presents the prompt used to generate the initial 'Outline'.

[System Prompt]
You are an intelligent critic, whose job is to identify inconsistencies between data presented in data tables, and a reflection and an outline. Pay attention to small details and nuances as well as any trends or outliers in the given tables.

[User Prompt]
Task Description:
Given the data tables, a reflection and an outline corresponding to a data story in the input, your task is the following:
1. Identify inconsistencies and factual inaccuracies in the outline considering the data in the tables, and the reflection. The information in the outline must be factually correct.
2. Adjust the narrative flow if needed, to keep this theme central to the story, ensuring that each section contributes meaningfully to the theme.
3. Based on your analysis, draft a revision plan to refine the the outline if needed, and output the revision plan. Otherwise just output: 'No revision needed'.
4. Make sure the revision plan is consistent with the intention or the main theme of the story: ***<intention>***, and is completely aligned with the theme.
5. The revision plan must coherently and logically relate to the attributes of the data. Be as specific as possible.

Additional Guidelines:
- The output must be in plain text and structured in bullet points.
- Think step by step and generate the response outline in between two <outline> tags.

INPUT:
Tables:
<Tables>
Reflection:
<final_reflection>
Outline:
<outline>

Figure 22: The figure presents the prompt used to generate the 'Outline' revision plan.

[System Prompt]
You are an expert at generating outlines for data stories. The generated outline should cover every important detail that can be observed in the data tables. Pay attention to small details and nuances as well as any trends or outliers in the given tables.

[User Prompt]
Task Description:
Given the data tables, the revision plan and the outline corresponding to a data story in the input, your task is the following:
1. Apply the changes suggested in the revision plan to the existing outline.
2. Ensure Theme Consistency: The data story's overarching theme, defined as ***<intention>***, should be clearly reflected throughout the revised outline.
3. The revised outline should be detailed in plain text, with each bullet point clearly articulating the specific aspect of the data story it addresses.
4. Be specific, be clear, and most importantly, be engaging. The generated outline must coherently and logically relate to the attributes of the data and rigourously follow the theme. Be as specific as possible.

Additional Guidelines:
- The output must be in plain text and structured in bullet points.
- Think step by step and generate the response outline in between two <outline> tags.

INPUT:
Tables:
<Tables>
Previous Outline:
<outline>
Revision Plan:
<outline_revision_plan>

Figure 23: The figure presents the prompt used to generate the revised 'Outline'.

[System Prompt]

You are an expert at generating engaging data stories. The generated data story should cover every important detail that can be observed in the data tables. Pay attention to small details and nuances as well as any trends or outliers in the given tables.

[User Prompt]

Task Description:

Given a outline and the data tables corresponding to a data story in the input, you have the following tasks:
1. Follow the outline rigorously to generate a "data story" that is highly informative and engaging to the audience.
2. The overarching theme, denoted as ***<intention>***, should be the narrative's backbone. Ensure that this theme resonates throughout the story, tying together different data points and insights into a coherent whole.
3. Highlight key statistics that are critical to understanding the theme. Explain these elements in a way that balances technical accuracy with accessibility, ensuring that your narrative is approachable for a non-specialist audience while still offering depth for those more familiar with the subject matter. Think about the narrative flow and how each piece of data contributes to the overall story arc.
4. In the outline, if it is mentioned to include a visualization, then include a 'visualization' placeholder. Each visualization placeholder should also suggest a narrative element that the visualization supports or explains.
5. Ensure that each paragraph in the story is in between two `<paragraph>` tags.
6. Ensure that each of the paragraph headers is in between two `<head>` tags.
7. The visualization placeholder must contain detailed information about the visualization, such as:
 - chart title
 - chart type (such as, `line`, `bar`, `pie`, `scatter plot`, etc.). Keep the chart types simple and appropriate to present the data. Do not include any complicated visualizations or infographics.
 - x-axis and y-axis
 - x-axis data values and y-axis data values, etc.
8. The visualization specifications must be sufficient to generate informative visualizations. Make sure the specifications are in `json` format and put in between two <visualization> tags.
9. Make sure that the story is highly informative and engaging to the audience.
10. Ensure coherence and clarity, connect information with proper synthesis and make connection to the overall narrative.

Additional Guidelines:

- The output must be in plain text.
- Generate the response narration in between two <narration> tags.

INPUT:

Tables:

<Tables>

Outline:

<final_outline>

Figure 24: The figure presents the prompt used to generate the initial 'Narration'.

[System Prompt]
You are an intelligent critic, whose job is to identify inconsistencies between data presented in data tables, and an outline and a data story. Pay attention to small details and nuances as well as any trends or outliers in the given tables.

[User Prompt]
Task Description:
Given the outline, the data tables and a data story in the input, you have the following tasks:
1. Examine the data presented in the tables, the story's outline, and the narrative itself. Look for discrepancies, factual inaccuracies, or any details that do not align.
2. Provide a step-by-step analysis, highlighting specific data points and narrative elements that contribute to these inconsistencies.
3. Make sure the story fully aligns with the intention or the main theme: ***<intention>***. Ensure that this theme resonates throughout the story, tying together different data points and insights into a coherent whole.
4. Based on your analysis, draft a revision plan to refine the data story. Your plan should address identified inconsistencies and enhance theme alignment. Otherwise output: 'No revision needed'.
5. The output must be coherent, logically structured, and detailed, aiming for constructive feedback that enhances the data story's impact.

Additional Guidelines:
- The output must be in plain text and in bullet points.
- Generate the response narration in between two <narration> tags.

INPUT:
Tables:
<Tables>
Outline:
<final_outline>
Data Story:
<narration>

Figure 25: The figure presents the prompt used to generate the 'Narration' revision plan.

[System Prompt]
You are an expert at generating engaging data stories. The generated data story will cover every important detail that can be observed in the data tables. Pay attention to small details and nuances as well as any trends or outliers in the given tables.

[User Prompt]
Task Description:
Given the data tables, the outline, the revision plan, and the data story in the input, your task is the following:
1. Revise the data story according to the revision plan. Use the provided outline as your guide, adjusting the narrative according to the revision plan.
2. The overarching theme, denoted as ***<intention>***, should be the narrative's backbone.
3. Ensure that this theme resonates throughout the story, tying together different data points and insights into a coherent whole.
4. In the outline, if it is mentioned to include a visualization, then include a 'visualization' placeholder. Each visualization placeholder should also suggest a narrative element that the visualization supports or explains.
5. Ensure that each paragraph in the story is in between two `<paragraph>` tags.
6. Ensure that each of the paragraph headers is in between two `<head>` tags.
7. The visualization placeholder must contain detailed information about the visualization, such as:
- chart title
- chart type (such as, `line`, `bar`, `pie`, `scatter plot`, etc.). Keep the chart types simple and appropriate to present the data. Do not include any complicated visualizations or infographics.
- x-axis and y-axis
- x-axis data values and y-axis data values, etc.
8. The visualization specifications must be sufficient to generate informative visualizations. Make sure the specifications are in `json` format and put in between two <visualization> tags.
9. Make sure that the story is highly informative and engaging to the audience.
10. Ensure coherence and clarity, connect information with proper synthesis and make connection to the overall narrative.

Additional Guidelines:
- The output must be in plain text and in bullet points.
- Generate the response narration in between two <narration> tags.

INPUT:
Tables:
<Tables>
Outline:
<final_outline>
Previous Data Story:
<narration>
Revision plan:
<narration_revision_plan>

Figure 26: The figure presents the prompt used to generate the revised 'Narration'.

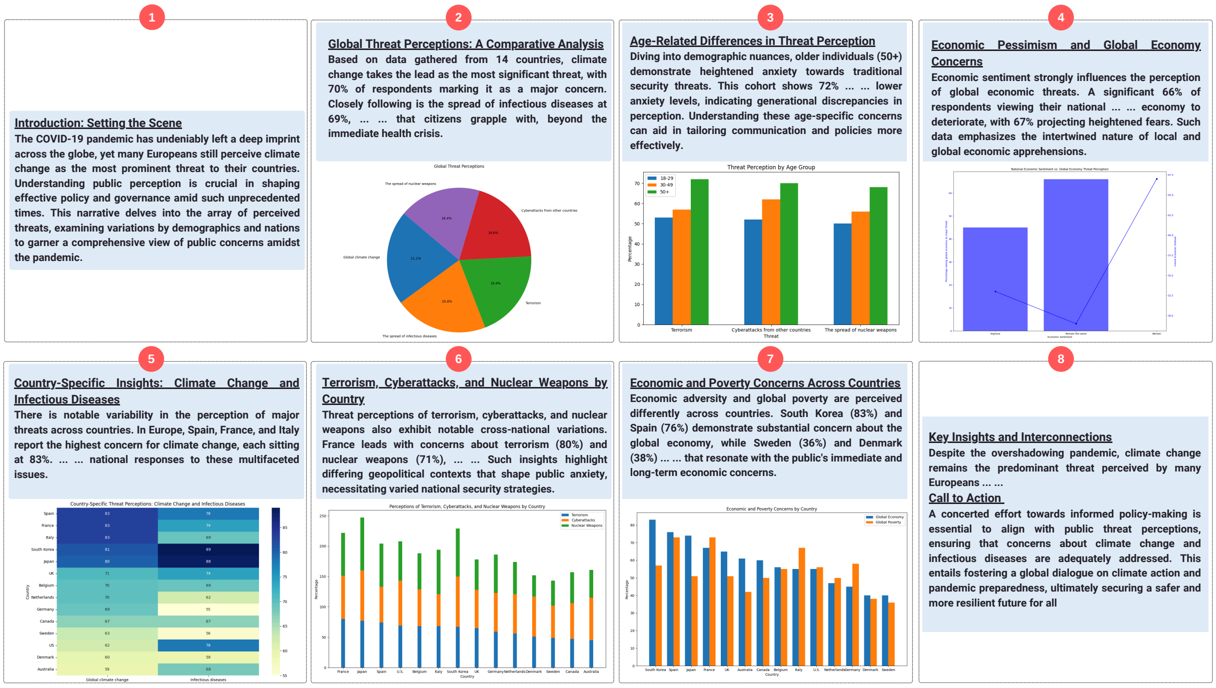


Figure 27: The figure demonstrates an example data story generated by GPT-4o using the agentic framework. Here, '...' indicates abbreviated text for brevity.

Example Data Story Generated by GPT-4o w/ Agentic Framework

Introduction: Setting the Scene

The COVID-19 pandemic has undeniably left a deep imprint across the globe, yet many Europeans still perceive climate change as the most prominent threat to their countries. Understanding public perception is crucial in shaping effective policy and governance amid such unprecedented times. This narrative delves into the array of perceived threats, examining variations by demographics and nations to garner a comprehensive view of public concerns amidst the pandemic.

Global Threat Perceptions: A Comparative Analysis

Based on data gathered from 14 countries, climate change takes the lead as the most significant threat, with 70% of respondents marking it as a major concern. Closely following is the spread of infectious diseases at 69%, showcasing an escalating focus on health amid the ongoing pandemic. Detailed threat rankings further reveal terrorism (66%), cyberattacks from other countries (65%), and the spread of nuclear weapons (61%) rounding out the top five threats. These insights underscore the intricate tapestry of fears that citizens grapple with, beyond the immediate health crisis.

```
<visualization>
{
 "chart_title": "Global Threat Perceptions",
 "chart_type": "pie",
 "data": {
  "labels": ["Global climate change", "The spread of infectious diseases", "Terrorism", "Cyberattacks from other countries", "The spread of nuclear weapons"],
  "values": [70, 69, 66, 65, 61]
 }
}
</visualization>
```

Age-Related Differences in Threat Perception

Diving into demographic nuances, older individuals (50+) demonstrate heightened anxiety towards traditional security threats. This cohort shows 72% concern about terrorism, 70% regarding cyberattacks, and 68% over nuclear weapons spread. Conversely, younger populations (18-29) display relatively lower anxiety levels, indicating generational discrepancies in perception. Understanding these age-specific concerns can aid in tailoring communication and policies more effectively.

```
<visualization>
{
 "chart_title": "Threat Perception by Age Group",
 "chart_type": "bar",
 "x_axis": "Age Group",
 "y_axis": "Percentage",
 "data": {
  "labels": ["Terrorism", "Cyberattacks from other countries", "The spread of nuclear weapons"],
  "18-29": [53, 52, 50],
  "30-49": [57, 62, 56],
  "50+": [72, 70, 68]
 }
}
</visualization>
```

...

...

Call to Action

A concerted effort towards informed policy-making is essential to align with public threat perceptions, ensuring that concerns about climate change and infectious diseases are adequately addressed. This entails fostering a global dialogue on climate action and pandemic preparedness, ultimately securing a safer and more resilient future for all.

Figure 28: The figure demonstrates an example data story generated by GPT-4o in natural language text. Here, '...' indicates abbreviated text for brevity.

Example Data Story Generated by LLaMA-3-8b-Instruct w/ Agentic Framework

Despite the ongoing COVID-19 pandemic, many Europeans still consider climate change as the greatest threat to their countries. According to a recent survey, 70% of respondents across 14 countries polled consider climate change a major threat to their country. This data highlights the importance of considering the global context in understanding local concerns and threats, and the interconnectedness of global economies and the spread of infectious diseases can have significant implications for national security and economic stability.

```
<visualization>
**Chart Title:** Global Threats to Countries
**Chart Type:** Bar Chart
**X-axis Label:** Countries
**Y-axis Label:** Percentage of Respondents
**X-axis Data Values:** 14 countries (US, Canada, Belgium, Denmark, France, Germany, Italy, Netherlands, Spain, Sweden, UK,
Australia, Japan, and South Korea)
**Y-axis Data Values:** 70% (climate change), 28% (infectious diseases), 66% (global economy), 53% (terrorism), 30%
(cyberattacks), 61% (nuclear weapons), 58% (poverty), and 40% (long-standing conflict)
</visualization>
```

The survey also found that the spread of infectious diseases is a significant concern, with 28% of respondents considering it a major threat to their country. This data highlights the importance of global cooperation and collective action to address the spread of infectious diseases, which can have devastating consequences for individuals and communities. Older people (ages 50+) are more concerned about traditional security threats, such as terrorism, cyberattacks, and the spread of nuclear weapons, with 72%, 70%, and 68% of respondents considering these threats major concerns, respectively. This data suggests that traditional security threats are still a significant concern for many Europeans, particularly older generations.

Pessimism about national economies is tied to concerns about the global economy, with 66% of respondents who say the economic situation in their country is bad also considering the condition of the global economy a major threat. This data highlights the interconnectedness of global economies and the impact of global economic trends on individual countries and communities. Majorities across all countries surveyed see climate change and infectious disease transmission as major threats, with some countries showing higher levels of concern. This data suggests that climate change and infectious disease transmission are global issues that require collective action and cooperation to address.

In conclusion, despite the pandemic, many Europeans still see climate change as the greatest threat to their countries, highlighting the importance of considering the global context in understanding local concerns and threats. The data suggests a need for collective action and cooperation to address these global challenges and ensure a more secure and prosperous future for all.

```
<visualization>
**Chart Title:** Global Threats to Countries
**Chart Type:** Bar Chart
**X-axis Label:** Countries
**Y-axis Label:** Percentage of Respondents
**X-axis Data Values:** 14 countries (US, Canada, Belgium, Denmark, France, Germany, Italy, Netherlands, Spain, Sweden, UK,
Australia, Japan, and South Korea)
**Y-axis Data Values:** 70% (climate change), 28% (infectious diseases), 66% (global economy), 53% (terrorism), 30%
(cyberattacks), 61% (nuclear weapons), 58% (poverty), and 40% (long-standing conflict)
</visualization>
```

The survey found that 70% of respondents across 14 countries polled consider climate change a major threat to their country. This is a significant finding that highlights the importance of climate change as a global threat. Older people (ages 50+) are more concerned about traditional security threats, such as terrorism, cyberattacks, and the spread of nuclear weapons. This is an important finding that highlights the importance of considering the perspectives of different age groups when addressing global challenges.

Pessimism about national economies is tied to concerns about the global economy. This is a critical finding that highlights the interconnectedness of global economies and the impact of global economic trends on individual countries and communities. The data suggests that climate change and infectious disease transmission are global issues that require collective action and cooperation to address.

Figure 29: The figure demonstrates an example data story generated by the LLaMA-3-8b-instruct model in natural language text.