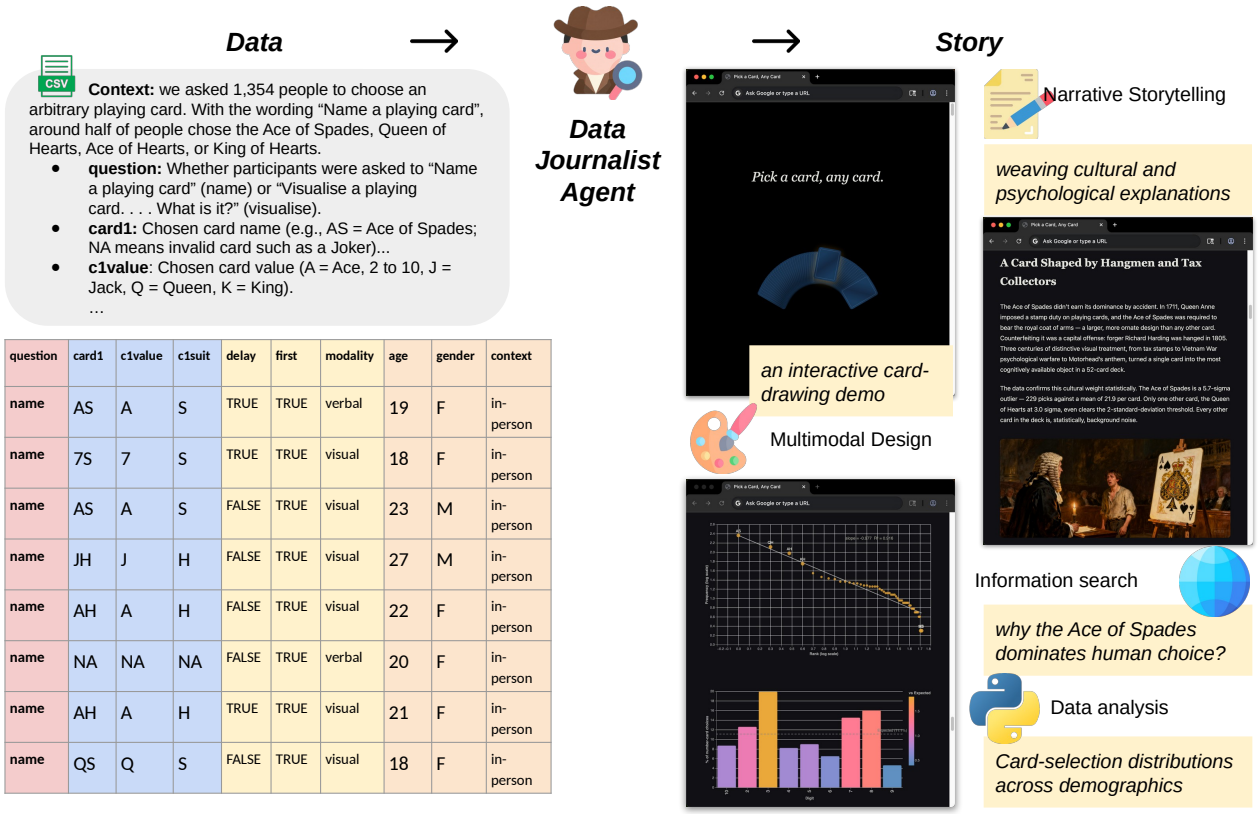
Data Journalist Agent: Transforming Data into Verifiable Multimodal Stories

Kevin Qinghong Lin[♡], **Batu El**[♠], **Yuhong Shi**[♡], **Pan Lu**[♠], **Philip Torr**[✉♡], **James Zou**[✉♠]

[♡]University of Oxford, [♠]Stanford University

Website: https://data2story.github.io
Code: https://github.com/QinghongLin/data2story-skill
Correspondence: philip.torr@eng.ox.ac.uk, jamesz@stanford.edu

Data tells stories that shape society, and the data journalist's job is to turn raw information into a story that non-expert audiences can understand and trust through to the end. A high-quality news feature routinely takes a newsroom team weeks, including hunting for context, running statistics, choosing an angle, and designing visuals. Recent agents are capable at individual steps: automated data-science agents close the analysis loop, while design agents can synthesize beautiful websites. *But can an agent serve as a data journalist end to end?* We introduce **Data Journalist Agent (Data2Story)**, a multi-agent framework that orchestrates specialised roles into a single virtual newsroom. Data2Story highlights two innovations over prior approaches. **(i) Claims are evidence-grounded and verifiable.** We introduce an "Inspector", which links the intermediate results produced by individual roles to their sources so that the numbers, angles, and assets are grounded in data, code, or a reference (*e.g.,* an external URL). **(ii) Articles are multimodally generative.** Rather than defaulting to plain text and static charts, Data2Story reasons about what its readers will want to read visually, then deploys multimodal tools so that the article fits both the data topic and the intended audience (*e.g.,* an interactive map with zoom for a geography piece, or an audio clip for a music piece), making the result readable and engaging. We evaluate Data Journalist Agent on 18 articles from diverse topics and publication sources, each paired with the originally published expert-written piece, along four axes: **(a) Human–agent angle coverage**, measuring the overlap and complementarity of angles between Data2Story and human-authored articles, to characterize what each side covers; **(b) Rubric evaluation with a human study** across 53 human participants, with the rubric covering visual design, narrative pacing, data transparency, claim-data alignment, and insight value; **(c) Computer-use agents as judge**: as an automatic cost-saving proxy for how real-world users navigate and interact with the article, we employ computer-use agents that fully perceive the interface through actions such as clicking and scrolling; and lastly, **(d) Verifiability**, where a coding verifier re-executes every statement against the data and checks that the claims are verifiable or can be grounded in a reference. Our central finding is that Data2Story produces competitive and evidence-traceable multimedia stories, with particularly strong performance on transparency and auditability dimensions. However, human-authored articles retain a clear edge in editorial angle, creative design, and informative presentation. Data2Story is not intended to replace journalists. Rather, it serves as a solution to support story development, enabling reporting that is more evidence-based, transparent, and verifiable.



question	card1	c1value	c1suit	delay	first	modality	age	gender	context
name	AS	A	S	TRUE	TRUE	verbal	19	F	in-person
name	7S	7	S	TRUE	TRUE	visual	18	F	in-person
name	AS	A	S	FALSE	TRUE	visual	23	M	in-person
name	JH	J	H	FALSE	TRUE	visual	27	M	in-person
name	AH	A	H	FALSE	TRUE	visual	22	F	in-person
name	NA	NA	NA	FALSE	TRUE	verbal	20	F	in-person
name	AH	A	H	TRUE	TRUE	visual	21	F	in-person
name	QS	Q	S	FALSE	TRUE	visual	18	F	in-person

Figure 1 | Data2Story turns a raw dataset (*e.g.*, a CSV) into a verifiable, multimodal article (*i.e.*, a website). We use a "Pick a card" dataset as an illustration. This transformation involves information seeking (*e.g.*, "why the Ace of Spades dominates human choice?"), data analysis via programming (*e.g.*, computing card-selection distributions across demographics), narrative storytelling (*e.g.*, weaving cultural and psychological explanations into a cohesive article), and multimodal design (*e.g.*, an interactive card-drawing demo).

1 Introduction

Data journalists turn raw data into stories like "How has the way pop singers use their voice changed across generations?" that everyday readers can follow, helping the public understand what lies behind the data – yet a small newsroom team can spend weeks on a single high-quality article. Recent agents are individually capable at each of these steps: automated data-science agents [1, 2, 3, 4] can profile a dataset, run the right statistics, and return defensible results with reproducible code. Visualization agents [5, 6, 7, 8] generate visual artifacts (such as websites) from a language instruction. *But can agents serve as journalists end to end, taking raw data all the way to a story readers actually want to finish and can trust?*

However, building such an end to end agentic journalist system is non-trivial. Behind each finished article is a long process: gathering background, running careful statistics, choosing an angle, designing assets, building an appealing page, and several rounds of editing. The task is fundamentally **multi-disciplinary**, demanding the simultaneous exercise of multiple skills that rarely co-exist in a single contributor, which is why news is typically the product of a coordinated newsroom team.

Companies such as CitizenPortal and Locunity are already deploying AI agents to produce news articles at scale, signalling that AI-enabled journalism is no longer hypothetical. However, a critical challenge shared by these systems is the lack of **verification and traceability** (as highlighted by the recent discussion [9]): readers and editors have no reliable way to confirm where a number came

from, whether a chart accurately reflects the underlying data, or whether a claim was inferred or hallucinated. This is a particularly demanding requirement for language agents, which are prone to hallucination [10]. Data2Story directly addresses this gap: nearly all statistic, visual asset, and factual claim is grounded in executable code or a verifiable source URL, making the full reasoning chain auditable end to end.

Motivated by this, we introduce **Data Journalist Agent (Data2Story)**, a multi-agent framework that orchestrates seven specialised roles into a virtual newsroom: a Detective for context hunting, an Analyst for running statistics, an Editor for narrative framing, a Designer for visual assets, a Programmer for website creation, an Auditor for reviewing the Programmer's output and offering suggestions for revision, and, most notably, an "Inspector" that traces elements of the final article back to its upstream evidence. As illustrated in Figure 1, Data Journalist Agent takes any data source as input and emits a generative multimedia article. Its key contributions are as follows: **(i) Claims are evidence-grounded.** To ensure the output is grounded in verifiable evidence, we introduce a dedicated agent that links most elements of the published article (*i.e.,* numbers, quotes, and visual assets) back to their provenance (*i.e.,* a specific line of code, a data source, or an external URL). This makes the resulting article *verifiable and auditable*. **(ii) Articles are multimodally generative.** Rather than formatting articles as plain text or static documents, we argue that an article should be multimedia-rich (e.g., interactive charts, images, video, and audio). We let a Designer reason about the topic and what readers will want to see and interact with. For example, as shown in Figure 1, for an article on card-game outcome statistics, we add a playable starter so that readers can interact with this game directly.

To validate the effectiveness of Data Journalist Agent, we first showcase Data2Story on the newest datasets that few humans have yet written up (*e.g.,* the 2026 World Cup schedule), where it discovers original findings of its own, such as an interactive map that ties venue geography to weather and highlights the matches at greatest high-temperature risk. This demonstrates its value for discovery and display via a user-friendly medium. Moreover, we collect 18 data samples from three representative publication sources, each paired with the expert-written piece. For a comprehensive assessment, we design metrics along four complementary axes. **(a) Human–agent angle coverage** extracts the factual claims from articles and reports similarity-matched coverage between human and agent. **(b) Rubric evaluation with human judges** asks 53 participants to score agent-generated or human-written articles blind on five rubric dimensions covering visual design, narrative pacing, data transparency, claim-data alignment, and insight value, and pick the preferred one overall. **(c) Computer-use agent as judge**: we explore a cost-saving automatic proxy for how real-world users navigate and interact with an article, employing computer-use agents that perceive the rendered interface through actions such as clicking and scrolling; **(d) Verifiability** uses a cross-family coding agent to validate claims by verifying statements such as executing code or searching the reference source.

Our experiments show that Data2Story produces multimodal articles that readers find compelling and are independently verifiable, with built-in evidence traceability at the claim level. Human raters judge them favorably across multiple quality dimensions; however, human journalists retain a clear edge in editorial angle, creative design, and informative presentation. Data2Story's greatest advantage instead lies in *auditability*: it makes the evidentiary basis of each claim explicit and measurable — something even carefully crafted human articles rarely provide natively.

We therefore position Data2Story as a collaborator rather than a replacement: humans set the perspective and editorial judgment, while (i) agents handle labor-intensive computation and graphics design and (ii) open the door to specialised, data-rich stories that newsrooms do not have the bandwidth to cover.

Table 1 | Comparison with related works. *Ext. Search*: the system actively browses the web. *Narr. Angle*: the output is organized around a story angle rather than merely presenting data. *Multimodal* (*Image*, *Video*, *Audio*, *Interact.*): whether the system generates the corresponding modality or produces reader-interactive output. *Evidence* (*Source*, *Code*, *Grounded*): whether the output cites sources, ships runnable code, and makes each claim independently verifiable. ✓ present, ✓ partially present or not provided by default, ✗ absent.

System	Inputs	Outputs	Ext. Search	Narr. Angle	Multimodal Generative?				Evidence		
					Image	Video	Audio	Interact.	Source	Code	Grounded
Search Agents											
MindSearch [11]	Query	Report	✓	✗	✗	✗	✗	✗	✓	✗	✓
MMSearch [12]	Query+Image	Text	✓	✗	✗	✗	✗	✗	✓	✗	✓
DR Tulu [13]	Query	Text	✓	✗	✗	✗	✗	✗	✓	✗	✓
Data Visualization Agents											
MatplotAgent [5]	Query+Data	Infographic	✗	✗	✓	✗	✗	✗	✗	✓	✓
LIDA [6]	Query+Data	Infographic	✗	✗	✓	✗	✗	✗	✗	✓	✓
CoDA [7]	Query+Data	Infographic	✗	✗	✓	✗	✗	✓	✗	✓	✓
Data Science Agents											
DSGym [14]	Query+Data	Score	✗	✗	✗	✗	✗	✗	✗	✓	✓
Data Interpreter [15]	Query+Data	Report	✓	✗	✓	✗	✗	✓	✓	✓	✓
AI Scientist [16, 17]	Query	Report	✓	✓	✓	✗	✗	✗	✓	✓	✓
Data Journalist Agents											
LLM writer [18]	Press release	Angle	✗	✓	✗	✗	✗	✗	✓	✗	✓
Human writer [19]	Data	Article	✓	✓	✓	✓	✓	✓	✓	✗	✓
Data2Story(Ours)	Data	**Article**	✓	✓	✓	✓	✓	✓	✓	✓	✓

2 Related Work

In this section, we compare Data2Story against representative works in relevant fields. The comparison is illustrated in Tab.1.

Deep Search Agents take a natural-language query and autonomously browse the web to produce a retrieval-augmented text deliverable [20]. OpenAI's Deep Research [21] is the representative commercial demonstration, which browses the web [22] to collect knowledge, then augments the answer. MindSearch [11] decomposes the query into a graph of atomic sub-questions, each answered by a search-and-summarize role, while DeepResearcher [23] trains the browsing policy end-to-end with reinforcement learning. MMSearch [12] casts the requery, rerank, and summarize loop as a benchmark over short-answer outputs, and OpenResearcher [24] and DR Tulu [13] extend the open-source side of this line with retrieval-augmented scientific question answering and long-form report generation. These systems optimize the retrieval and synthesis of sources in response to a given query, but their deliverable remains a source-centric text document: they surface and summarize evidence rather than construct a narrative angle, and the query, not an editorial judgment about what is worth telling, drives the output.

Data Visualization Agents convert a fixed input into a visual or narrative artifact. LIDA [6] compiles a tabular dataset into executable visualization code, optionally restyled into an infographic. DataNarrative [25] pairs a generator and an evaluator to turn tables and a story intent into a narrative interleaved with chart specifications. MatplotAgent [5] generates plotting code through a collaborative agent system, but fails in metadata analysis. CoDA [7] further coordinates specialized agents to carry a dataset through analysis and into a composed visual report. On the other hand, these systems operate on the data they are given: they assume the input dataset as fixed and do not actively search

for external evidence, and their output is for the most part a static visual artifact rather than an interactive one.

Data Science Agents take a task description with data files and use executed code to produce their deliverable. DSGym [14] scores answer strings or CSV submissions in a sandbox with external tools disabled. DeepAnalyze [26] trains an agentic model end-to-end to interleave analysis, code, and execution into a research report. Data Interpreter [15] plans a task as a hierarchical subtask graph and emits whatever artifact the task requires, from a numeric answer to a playable mini-game. PublicAgent [27] routes an ambiguous question through four agents that discover an open-data table and run validated experiments into a traceable report. AI Scientist [16, 17] chains literature retrieval, experimentation, and writing into a workshop paper that cleared peer review. Across these systems the deliverable is a structured text artifact, and the form stays text-and-charts even when the target reader is a non-expert. In contrast, Data2Story packages the analysis as a multimedia article rather than a static PDF, the form a data-journalism reader actually consumes.

Data Journalists target general-audience data communication, either producing a publishable artifact for a non-expert reader or studying the journalism workflow empirically [19]. Recent work [18, 28, 29, 30, 31] has explored the use of language models in journalistic roles, such as assisting with article planning, recommending angles, and identifying sources. DataDirector [32] fuses Vega-Lite charts, TTS audio, and animation into a passive animated data video. The human data-journalist baseline produces a multimedia article with inline source citations, the gold-standard reader-facing form, but most human articles lack code-line provenance. Data2Story closes this gap: it routes structured multi-source data through seven specialized roles into a multimedia-rich article whose Inspector binds rendered sentences and charts to specific code lines or source URLs.

3 Data Journalist Agent

Given any raw data $\mathcal{D}$, the goal of Data Journalist Agent is to produce an article $\mathcal{U}$ that is narratively compelling, visually appealing, and verifiable in its content.

3.1 The Virtual Newsroom

As illustrated in Figure 2, we define our multi-agent solution as a virtual newsroom composed of specialised agent roles.

Detective. A raw data source is rarely enough on its own: an article almost always depends on context the dataset does not contain. For example, historical events often need to be associated with the time the data were released. The Detective gathers this context before any number is computed, so that downstream roles can frame the data rather than invent claims about it. Concretely, it augments the raw dataset $\mathcal{D}$ via web search into an enriched corpus $\mathcal{D} \cup \widetilde{\mathcal{D}}$, where $\widetilde{\mathcal{D}} \xleftarrow{\text{Web search}} \mathcal{D}$ contains additional context items tagged with category and source URLs, together with a small library of reference media (photographs, maps, short clips) that other agents can later reuse.

Analyst. A news article typically cites dozens of statistics to arrive at its insights. However, given a dataset, it is rarely clear in advance which statistical findings it admits, or which of them will prove most meaningful. The Analyst therefore prioritises completeness: it enumerates every analysis the dataset can support, profiles every column, and runs actual code rather than asking the model to estimate. From the augmented dataset, it derives a set of results $\mathcal{R} = \{r_i\}$ and supporting code

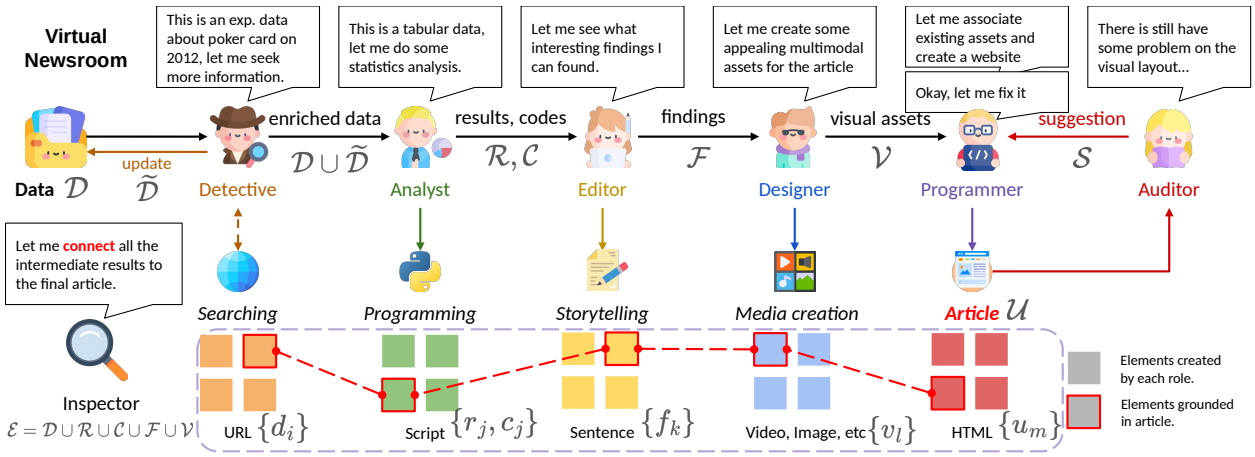


Figure 2 | The Virtual Newsroom for Data2Story. A raw dataset $\mathcal{D}$ flows through a sequence of specialist roles: the *Detective* gathers external context $\widetilde{\mathcal{D}}$ from the web, the *Analyst* writes Python code $\mathcal{C}$ and emits results $\mathcal{R}$ with code-line provenance $\mathcal{R} \xleftarrow{\mathcal{C}} \mathcal{D} \cup \widetilde{\mathcal{D}}$, the *Editor* drafts several findings $\mathcal{F}$ from different angles, the *Designer* produces multimedia assets $\mathcal{V}$ via tool calls, and the *Programmer* renders the final HTML $\mathcal{U}$. The page is then audited by the *Auditor*, which provides suggestions $\mathcal{S}$ for visual and structural defects, and the *Inspector*, which binds every published claim back to its supporting evidence $\mathcal{E} = \mathcal{D} \cup \mathcal{R} \cup \mathcal{C} \cup \mathcal{F} \cup \mathcal{V}$. Each role produces a set of intermediate elements (grey); those that ground the final article are highlighted (red outline), and the Inspector links them into a traceable evidence chain.

$\mathcal{C} = \{c_i\}$ with $r_i \xleftarrow{c_i} \mathcal{D} \cup \widetilde{\mathcal{D}}$, where each finding r_i carries a pointer to the script c_i that generated it, ensuring that every outcome is traceable.

Editor. An interesting analysis is not yet a story. Given a set of findings, the Editor decides what the article actually argues: which findings should lead, which should support, which add colour, and which should be cut. Reasoning over the Analyst's findings, it produces an editorial plan $\mathcal{F} \xleftarrow{\text{LLM}} \mathcal{R}$ that ranks each item by priority, selects the items worth keeping, and drafts a paragraph-level prose outline. Each finding f_i in $\mathcal{F}$ is annotated with the upstream items it draws on, $f_i \sim (r_i, c_i)$.

Designer. An article is not just plain text: multimedia elements can substantially improve readability, such as maps for geography, audio for music, video for events, and interactive widgets for complex findings. For each finding f_i of the editorial plan, the Designer reasons about what a reader would most want to see, then selects the medium that best fits the data, drawing on a suite of external generative tools such as text-to-image and text-to-video. The resulting per-section visual assets $\mathcal{V} \xleftarrow{\text{Tool}} \mathcal{F}$ include the corresponding asset calls needed to realise each medium, where we store every prompt or parameter.

Programmer. Static formats such as PDF cannot natively coordinate multimedia elements; an HTML webpage, by contrast, is the ideal medium for what a reader actually sees. We therefore introduce a Programmer that renders the final page in HTML from the upstream artifacts. The Programmer generates no new facts or numbers; it operates in two modes. (i) In assembly mode, it quotes the upstream artifacts $\{\mathcal{F}, \mathcal{V}\}$ and composes them into a complete interactive article $\mathcal{U} \leftarrow \{\mathcal{F}, \mathcal{V}\}$. (ii) In revision mode, it additionally takes the Auditor's revision suggestions $\mathcal{S}$ and revises the page accordingly, forwarding the audited article $\mathcal{U} \leftarrow \{\mathcal{U}, \mathcal{S}\}$ to the Inspector.

Auditor. The rendered HTML may still harbour visual or structural defects: overlapping elements, broken charts, missing assets, or unresponsive interactions. Such defects can quietly undermine an otherwise well-grounded story. The Auditor therefore reviews the rendered page, $\mathcal{S} \leftarrow \mathcal{U}$, and flags these issues; it returns the page to the Programmer for repair.

3.2 How to ensure claims are verifiable?

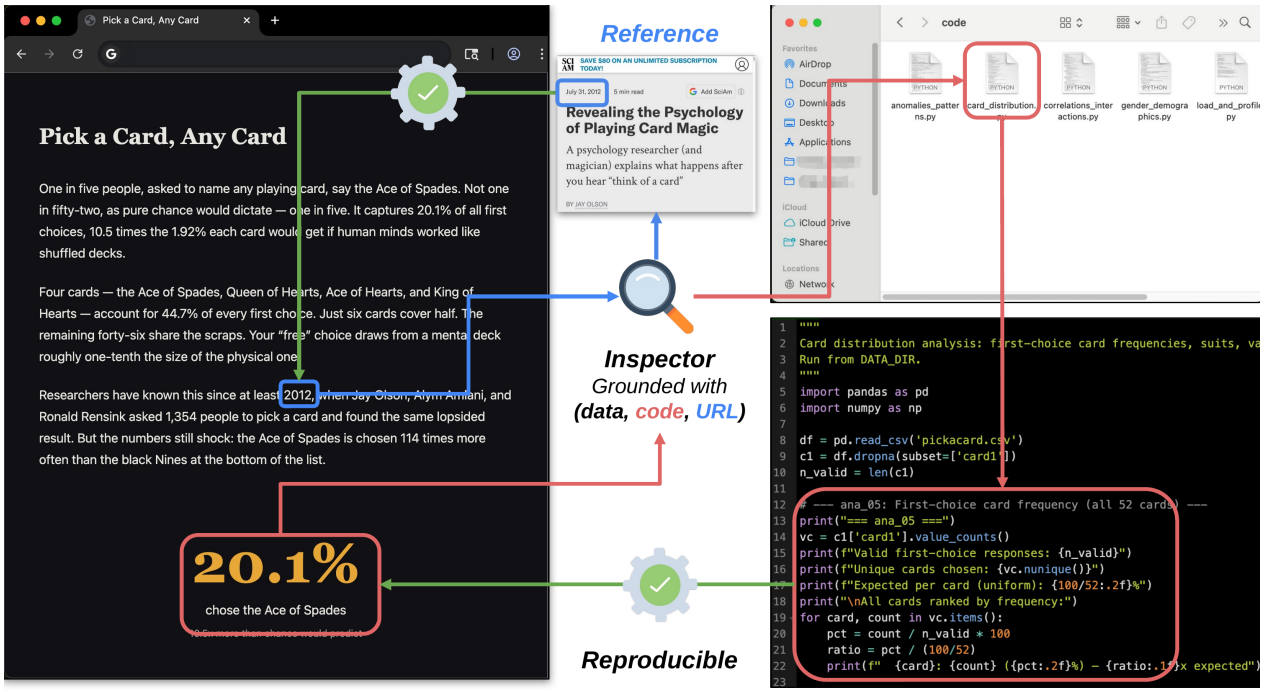


Figure 3 | Illustration of the Inspector. The Inspector binds every output finding back to its supporting evidence, which falls into two types: (i) *code evidence*, the source file and specific line that produced a reported number, and (ii) *reference evidence*, the external article or URL that grounds a contextual claim. The binding establishes auditability / traceability rather than factual correctness.

Inspector 🔍. A central challenge for any multi-agent system that produces an article is that the reader has no reason to trust the page unless every visible element, from the lede sentence to the final tooltip, resolves to something concrete upstream (such as code or reference). We therefore introduce the Inspector, which closes this loop at the level of individual items.

We let all upstream agents each contribute *atomic units of evidence*. The Detective contributes a context $\mathcal{D} = \{d_i\}$, where each d_i is a context item with a source URL. The Analyst contributes findings $\mathcal{R} = \{r_j\}$ paired one-to-one with code $\mathcal{C} = \{c_j\}$, so that every r_j is supported by the script c_j that produced it. The Editor contributes a finding $\mathcal{F} = \{f_k\}$, where each f_k is a paragraph with upstream pointers, and the Designer contributes specifications $\mathcal{V} = \{v_\ell\}$, where each v_ℓ is a per-section specification, and we record the tool call and parameters (such as prompts). Together these form the pool of upstream evidence $\mathcal{E} = \mathcal{D} \cup \mathcal{R} \cup \mathcal{C} \cup \mathcal{F} \cup \mathcal{V}$.

The Inspector decomposes the audited page into a set of partial findings $\mathcal{U} = \{u_m\}$, where each u_m is a self-contained HTML fragment realising a sentence, chart, or interactive element. It then binds every fragment u_m to the entries of the evidence base $\mathcal{E}$ that ground it, *i.e.*, $u_m \sim (d_i, r_j, c_j, f_k, v_l)$, so that each fragment carries an explicit link back to the evidence from which it was derived.

The Inspector recognises two types of evidence link, as illustrated in Figure 3: *code evidence*, where

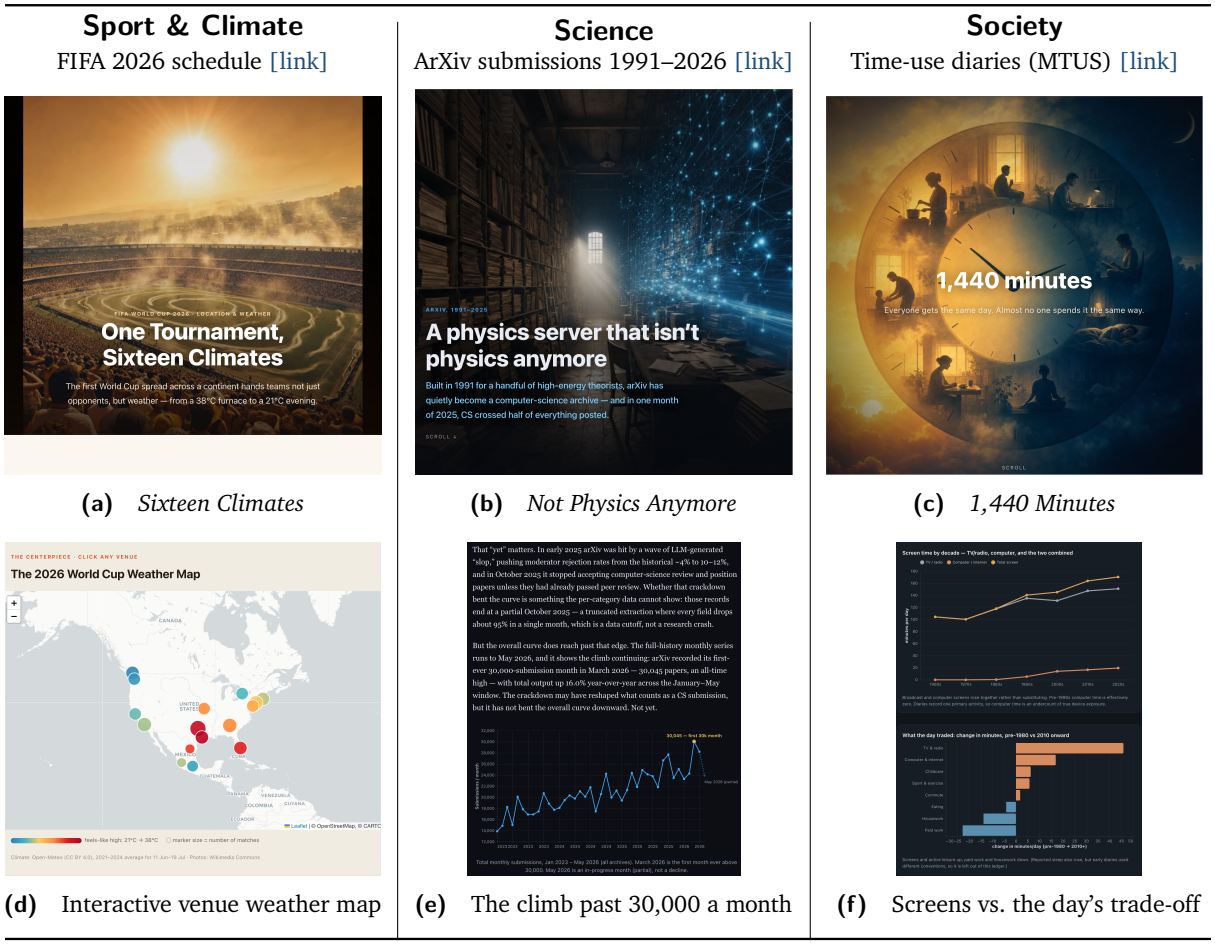


Figure 4 | Data2Story discovering findings on new data with no human reference. Three datasets from 2026 that have no canonical human-written piece, covering sport (a), science (b), and society (c). The **top row** is the opening cover of each piece; the **bottom row** is its signature data view.

a claim traces back to the specific script and line that produced it, and *reference evidence*, where a contextual claim is grounded in an external URL. The result is a page where truthfulness is **evidence-traceable**: every claim can be followed back through the Programmer, the Designer, and the Analyst to the original data file or source reference.

3.3 Data2Story discovers findings on underexplored data

To illustrate Data2Story, we apply it to *new* datasets rarely written by journalists, to show that it can autonomously *discover* an original angle and back it with its own analysis. We chose three datasets that are publicly available in 2026 (Figure 4), spanning society, sport, and the AI industry. For each, we describe Data2Story's writing angle and the core findings the article surfaces.

(a) FIFA 2026 schedule[1]: geography fused with climate. The 2026 World Cup is the first one to spread across a whole continent, so Data2Story fuses each venue's geographic location with its typical climate (Open-Meteo) and FIFPRO's heat-risk flags, interpreting the fixture list as a climate document rather than a sports calendar. The cover (Figure 4a) embodies that tension with a blazing sun over a packed stadium under the title *"One Tournament, Sixteen Climates,"* dramatising a feels-like gulf between a furnace-like Houston and a cool Vancouver, baked into the bracket before a ball is

[1] https://www.fifa.com/

kicked. The core finding is striking: roughly four in ten matches are booked at the venues FIFPRO flags as "extremely high risk," and humidity, not air temperature, drives the worst penalties. The interactive weather map (Figure 4d), the article's centerpiece, lets the reader see this venue by venue. Throughout, the piece keeps the caveat visible: these are typical-climate odds, not a 2026 forecast.

(b) ArXiv submissions[2]: a physics preprint server that has become a computer science platform. Reading three decades of submissions to arXiv, the preprint server that physicist Paul Ginsparg launched in 1991, Data2Story writes from a contrarian angle: the "physics archive" everyone still pictures has quietly become a computer-science one. The cover (Figure 4b), *"A Physics Server That Isn't Physics Anymore,"* embodies this as a sunlit corridor of dusty paper stacks dissolving into a glowing data-network on the right, the founding discipline giving way to the field that overran it. The core finding is stark: computer science is now 42.5% of everything posted, and in May 2025 it crossed half of all submissions in a single month for the first time. The chart (Figure 4e), with data running through 2026, traces the total monthly output still bending upward, reaching arXiv's first-ever 30, 000-submission month in March 2026. The piece ties this surge to a sharper policy turn: facing a wave of LLM-generated "slop" that sent rejection rates climbing, *arXiv stopped treating an institutional email as enough to endorse a first-time submitter in January* 2026, so for the first time the archive is actively deciding who gets to post.

(c) Time-use diaries[3]: the day as a fairness ledger. From the Multinational Time Use Study, harmonised from large-scale national diary surveys across dozens of countries and six decades, Data2Story writes from a single angle: a day is the one resource everyone owns in equal measure, exactly 1,440 minutes, yet who spends them on unpaid work splits sharply by sex, country, and decade. The cover (Figure 4c) embodies that angle with a luminous 24-hour clock face filled with silhouettes of cooking, childcare, and sleep, captioned *"everyone gets the same day, almost no one spends it the same way,"* so an abstract statistic becomes the reader's own morning. The core finding follows from the diaries: women do more than twice men's unpaid work, and once paid and unpaid hours are summed, they work longer days overall. Read by decade (Figure 4f), "**screen time**" (TV, radio, computer, internet) rose while paid work and housework fell, and the gender gap narrowed not because women were freed but because men slowly did more at home. The total work society performs barely moves over the decades; only its division by sex and kind shifts, the "work-time invariance" pattern.

4 Evaluation

In this section, we investigate three research questions: **(i)** How can we fairly evaluate data-journalism articles produced by either humans or agents – what metrics and protocols faithfully capture the quality of such outputs? **(ii)** How do Data2Story-generated articles compare against human-written counterparts, and along which dimensions? **(iii)** To what extent do human and agent judges agree, and how consistent are they across samples?

4.1 Setting

We evaluate Data2Story on various examples drawn from three stylistically distinct sources, deliberately chosen to span the spectrum of contemporary data storytelling. In curating the evaluation set, we sought diversity along the following axes: domain (science, media, sports, politics, health, and

[2] https://arxiv.org/stats/main

[3] https://rdr.ucl.ac.uk/articles/dataset/Multinational_time_use_study_release_version_11/28682660

Table 2 | Evaluation set. Each row pairs a dataset with a published human-written piece. Human articles rarely ship complete code, so ✓in **Code** marks partial code (*e.g.,* data-cleaning only).

#	Year	Domain	Modality	Code	Title of Human Article
Source: The Economist					
1	2018	Science	timeseries		The space race is dominated by new contenders [link]
2	2018	Media	timeseries		TV's golden age is real [link]
3	2019	Sports	panel		Managers in football matter much less than most fans think [link]
4	2019	Politics	geospatial		Israel's growing settlements force stark choices about its future [link]
5	2019	Media	tabular		The Oscars' influence has waned [link]
6	2020	Health	panel		Tourism flows and death rates suggest covid-19 is being under-reported [link]
Source: The Pudding					
7	2018	Culture	text		The Structure of Stand-Up Comedy [link]
8	2019	Music	tabular		Vocal Register in Pop Music [link]
9	2023	Music	tabular		They Won't Play a Lady-O on Country Radio [link]
10	2018	Music	tabular		The Eras of Boy Bands [link]
11	2017	Health	geospatial		How Far Is Too Far for an Abortion Clinic [link]
12	2018	Food	text		Baking the Most Average Chocolate Chip Cookie [link]
Source: TidyTuesday					
13	2019	Tech	timeseries	✓	Technological Progress (Moore's Law) [link]
14	2026	Science	text	✓	How Many Decimals of Pi Do We Really Need? [link]
15	2020	Music	text	✓	Taylor Swift and Beyoncé Lyric Analysis [link]
16	2026	Climate	tabular	✓	Repair Cafés and Consumer Waste [link]
17	2026	Sports	tabular	✓	Milano Cortina 2026 Olympic Schedule [link]
18	2025	Climate	timeseries	✓	Sechseläuten Snowman (Böögg) Forecast [link]

others), temporal coverage (spanning 2018–2026), and data modality (time series and tabular data, among others).

For publication source, we consider: (i) The Economist, featuring concise, analytical economics-style reporting; (ii) The Pudding, known for artistically rich, long-form interactive essays; and (iii) TidyTuesday, a community initiative providing more diverse datasets together with data-processing code and their original source articles. For every example, we pair the underlying data with the human-written reference piece, enabling head-to-head comparison against the Data Journalist Agent outcome. Table 2 lists all 18 original articles.

Potential training-data contamination. We acknowledge that well-known Economist and Pudding articles may sit in model pretraining corpora, which we cannot rule out. But *recalling text alone earns no score*: (i) coverage is bidirectional, rewarding not only matching the human angle but also surfacing claims the human article omits, which memorising that article cannot supply; and (ii) human articles ship no code, so even a memorised angle cannot help pass verifiability, which a cross-family verifier checks by re-running code against the data.

Data Journalist Agent articles are produced using Claude Code with `claude-opus-4.7`. We provide full details information in Appendix B.

4.2 Evaluation Metrics

We evaluate Data Journalist Agent along three orthogonal axes.

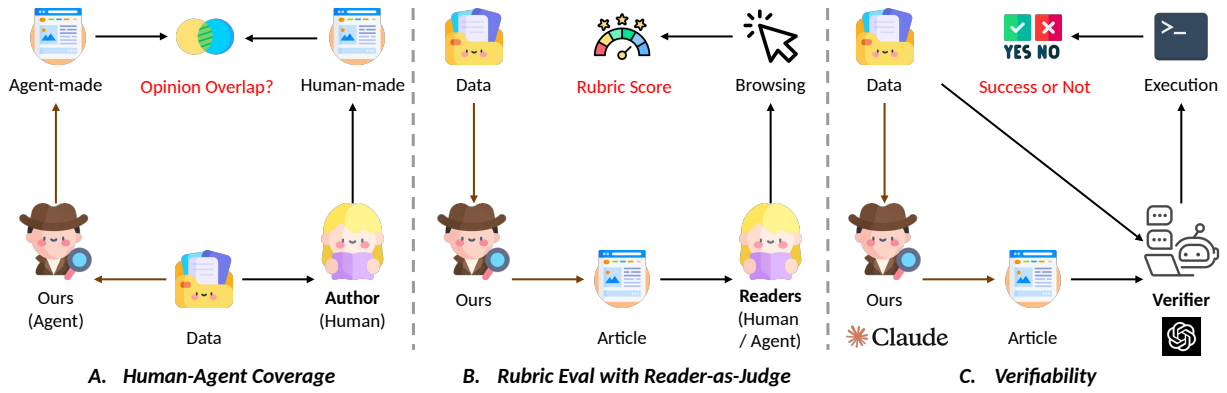


Figure 5 | Three complementary evaluation protocols for Data2Story. (A) Human-agent angle coverage: the agent and a human author independently produce articles from the same dataset; we measure overlap in the claims and insights surfaced by each. **(B) Rubric evaluation with reader as judge:** a human (or a computer-use agent) reader scores the agent's article against the human-written reference along five rubric dimensions, yielding graded quality assessments. **(C) Verifiability:** a verifier agent attempts to reproduce the agent's output from the same inputs, yielding a binary judgment of whether the artefact is faithfully verifiable.

Human-agent angle coverage. For every paired human–agent article, we measure how much overlap exists between the human-written reference and the Data Journalist Agent output. As shown in Figure 5A, we parse the article into various sentences, then apply `gpt-4o-mini` to filter the article content (such as advertisements), resulting in a set of factual claims from the human article Human and from the agent article Agent respectively. We then match claims across the two sides: OpenAI's `text-embedding-3-small` retrieves the top-3 nearest candidates by cosine similarity, and `gpt-4o-mini` decides under a relaxed prompt whether the candidate pair covers the same topic. A claim is covered if at least one of its candidates passes the LLM check. This gives us two directional coverage scores:

- **Human-in-Agent** P(Agent | Human): the fraction of human claims that the agent article also surfaces. *i.e., did the agent catch what a journalist would catch?*
- **Agent-in-Human** P(Human | Agent): the fraction of agent claims that also appear in the human article, indicating how closely the agent's claims track the human-curated angle.

Formally,

$$\mathrm{P}(\mathrm{Agent} \mid \mathrm{Human}) = \frac{|\mathrm{Human} \cap \mathrm{Agent}|}{|\mathrm{Human}|}, \qquad \mathrm{P}(\mathrm{Human} \mid \mathrm{Agent}) = \frac{|\mathrm{Human} \cap \mathrm{Agent}|}{|\mathrm{Agent}|},$$

where $\mathrm{Human} \cap \mathrm{Agent}$ denotes the set of claims matched across both sides. A high P(Agent | Human) indicates that the agent covers more of the human's angle, while a high P(Human | Agent) indicates that the human covers more of the agent's angle; a high value on either side indicates strong coverage of that side's angle, while a gap between the two reflects claims unique to one side, whether from divergence or broader coverage.

Rubric evaluation & Human as judge. An article is ultimately meant to be read, so the *primary* evaluation is to place it in front of readers (illustrated in Figure 5B). Because a data-driven article is not a single output but a composite artefact spanning prose, visuals, and analysis, a one-dimensional score cannot capture its quality [19, 33]; we thus assess it along a rubric. We recruit 53 reviewers via the Prolific platform[4]; each is assigned one Data2Story–human pair (presentation order randomised, blind) and scores both versions along the five rubric dimensions below on a 1–7 scale:

[4]https://www.prolific.com/

1. **Visual Design [34, 35].** Whether palette, typography, layout, and chart-type choice are polished and well matched to the claim each chart supports.
2. **Narrative & Pacing [36, 37].** Whether the hook, ordering, rhythm, and ending make the artefact read as a guided tour rather than a list of facts.
3. **Data & Method Transparency [38, 39].** Whether sources are cited specifically, methodology is described, data is accessible, and limitations are acknowledged with concrete numbers or exclusions.
4. **Claim–Data Alignment [40, 41].** Whether quantitative claims are bounded by what the data can support, confounders are named, and chart encodings are unambiguous.
5. **Insight Value [42, 43].** Whether the reader gains a non-trivial cognitive update; capped at 3 if the takeaway restates common knowledge, capped at 5 if the update is meaningful only to a lay reader.

After viewing both, each reviewer also expresses a binary preference indicating which version they prefer overall.

Computer-use agent as Judge. Beyond human evaluation (which requires costly manual efforts), we also consider a cost-saving protocol that uses a model as judge. This follows the same setup as Figure 5B, with the human reader replaced by an agent. We use an across-family agent, OpenAI's browser-use `gpt-5.5-xhigh`. An article, however, is an interactive website: standard LLM [44, 45] or VLM [46] judges perceive only static screenshots and cannot scroll, hover, or trigger animations, missing precisely the dynamic elements that distinguish a polished interactive piece from a static one. We therefore employ a computer-use agent [47] as judge, which navigates the rendered page like a human reader and scores it along the same rubric dimensions used in our human studies.

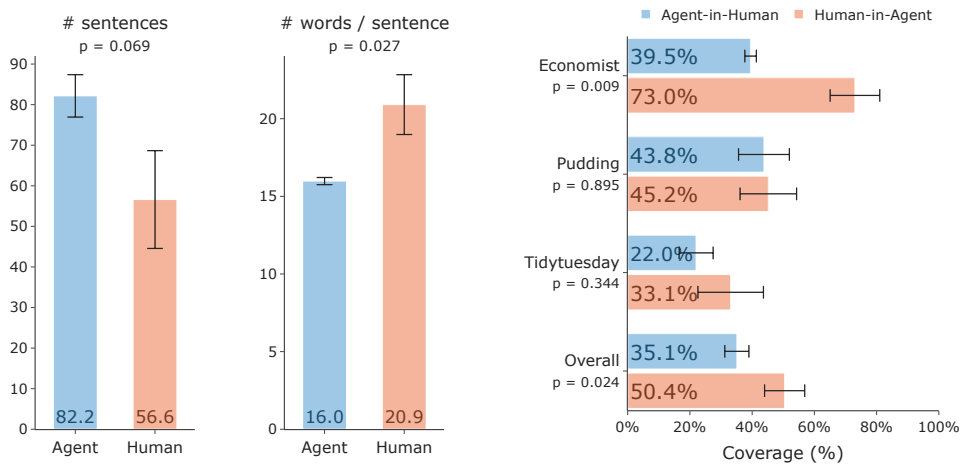
Verifiability. To verify that the published narrative is faithfully grounded in the underlying data (Figure 5 C), we replay every article with an across-family verifier (OpenAI's Coder `codex-GPT-5.4`). From each article, we extract the set of factual statements S, which fall into two categories: (a) *computational claims*, *i.e.,* numbers or findings derived from the data, which the checker verifies by re-executing the supporting `Python` or `R` scripts against the raw dataset; and (b) *reference-supported claims*, *i.e.,* statements backed by an external reference, which the checker verifies by re-fetching the cited source URL and confirming the claim against its content. For each claim, the checker returns a boolean result. We report the average pass rate as the article verifiability rate.

Notably, in verifiability experiments, the verifier has access to the original dataset when evaluating human-written articles, rather than the article text alone. For agent-written articles, the verifier additionally receives the full reasoning trajectory (by our Inspector) — a form of provenance made possible by evidence-grounded design.

4.3 Experiment Results

4.3.1 Distribution of article composition: where do humans and agents differ?

Before examining the article content, a natural first check is whether Data2Story writes at human scale. Across the 18 paired articles (Figure 6a), the total writing volume comes out roughly comparable (1305 for Data2Story and 1557 for humans), while the agent uses 1.45× as many sentences but each is shorter (0.77×); the articles made by Data2Story are broken into shorter, more granular statements.

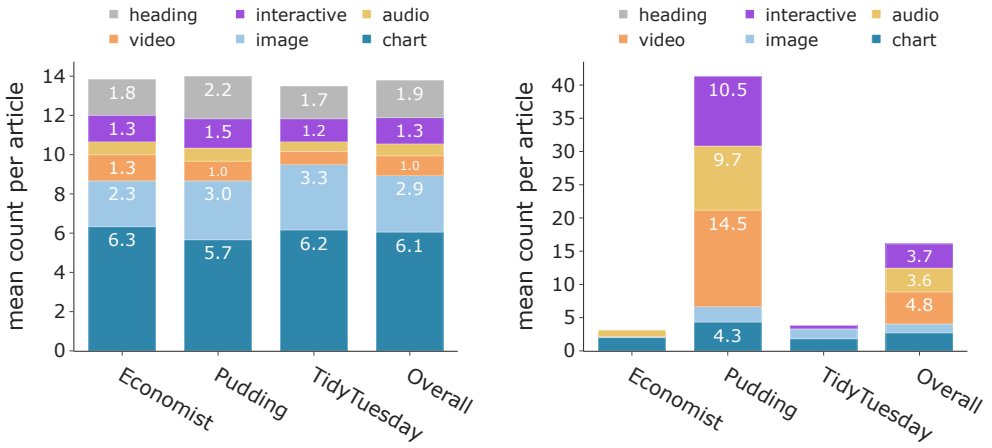


(a) Num. of sentences per article and Avg. words per sentence.

(b) Claim coverage between human-written and agent-generated articles.

Figure 6 | Textual distribution (left) and Content coverage (right) across 18 samples, reported by "mean ±SEM" with p value.

Matching textual statistics is one thing; the angle behind the text is what matters. As shown in Figure 6b, claim-level coverage points clearly one way: about half of the human article angle (50.4%) lands in the agent's article, while only a third (35.1%) of the agent's angle maps back. We find that the pattern is source-shaped, and each gap has a clear cause: it is widest on 'Economist' short briefings ($\Delta = 73.0\% - 39.5\% = 33.5\%$), whose narrow single-topic scope (typically standard statistic or chart) makes them easy for the agent to predict and cover; it stays uniformly lower on 'Pudding' and 'TidyTuesday', whose source articles either carry a single editorial thesis the agent does not fully reproduce (creative long-form storytelling) or span diverse topics as well as external sources ('TidyTuesday'). Data2Story reliably absorbs and rewrites the easy, predictable angles, but *reproducing a human author's narrative arc remains the harder problem.*



(a) Articles made by Data2Story.

(b) Articles made by human.

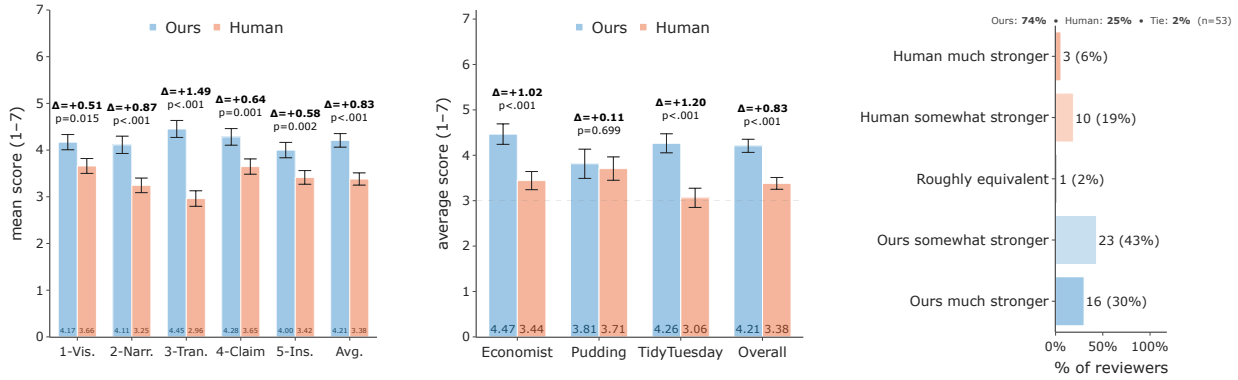
Figure 7 | Multimodal media asset distributions (*e.g.,* video, image, audio, interactive, etc) between Data2Story (left) and human (right).

Beyond text, every article may carry various multimedia assets, leading the article style to diverge sharply. We classify multimedia assets by six categories: heading (big short title), interactive, audio, video, image, and chart. As illustrated in Figure 7a, Data2Story's media distribution is uniform across all three sources: it averages 13–14 assets per article and covers every modality in similar proportions. By contrast, Figure 7b shows that human authors tune the kit to the publication: 'Pudding' carries about 41 assets per article, rich in video, audio, and interactives, while 'The Economist' and 'TidyTuesday' stay near 3–4, almost all charts and images. *Data2Story robustly produces every modality across topics, whereas human designers vary their distribution substantially with editorial style.*

4.3.2 Human studies as primary testbed

Data2Story articles are appreciated by humans across various rubrics. Figure 8a reports per-dimension means across the 53 participants, with the agent ahead on all five axes (with an overall mean of 4.21 for Data2Story and 3.38 for humans). The largest gap is on "Transparency" (+1.49), a margin we attribute to the Inspector per-sentence provenance and we provide an ablation in §4.3.5; the smallest is on "Visual" (+0.51), which we further ablate next.

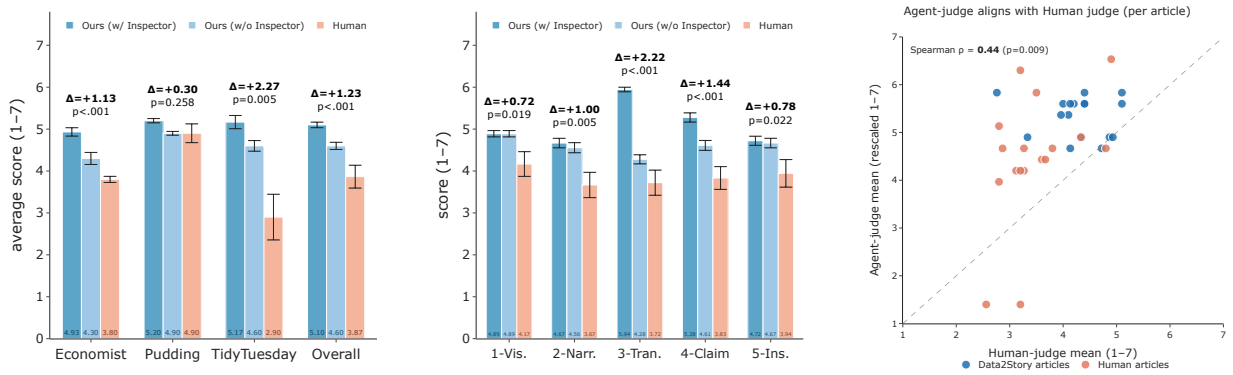
Analytical genres amplify the agent's advantage, while editorial scrollytelling narrows it. Figure 8b shows the breakdown by source: *Economist* (Δ=+1.02, p<.001) and *TidyTuesday* (Δ=+1.20, p<.001) both clearly favour the agent, while *Pudding* is a statistical tie. Pudding's long-form scrollytelling pieces are produced by art designer teams who spend weeks per article on bespoke design and a single committed thesis, an authorial investment the agent does not yet match. *The agent performs best in genres where analytical framing matters more than authorial voice; in the most designer-curated genre, it merely matches human performance.*



(a) By rubric dimension. **(b)** By source category. **(c)** Overall pairwise preference.

Figure 8 | Human evaluation (n=53 reviewers). Scores are grouped by rubric dimension (a) and source category (b). Finally, reviewers were asked to choose the better article through pairwise comparisons (c).

The holistic preference is consistent with the rubric. Beyond the per-dimension scores, each reviewer also gave a single overall preference after seeing both versions. Figure 8c shows the result: of the 53 reviewers, 39 preferred Data2Story, 13 preferred the human version, and 1(2%) calling it a tie. The holistic preference moves in the same direction as the per-dimension rubric, which suggests that the dimensions the rubric isolates (transparency, claim-data alignment, and so on) are also the ones reviewers weigh when forming an overall judgment.



(a) By source category.

(b) By rubric dimension.

(c) Agent judge aligns with human judge.

Figure 9 | Agent-as-judge evaluation. Scores are compared across Data2Story articles with the Inspector, Data2Story articles without the Inspector, and human-written articles. Results are grouped by source category (a) and rubric dimension (b), with score distributions from agent-judge and human-judge compared in (c).

4.3.3 Computer-use agent as a cost-efficient alternative

Figure 9a reports the Agent judge's average score with ablation of the Inspector across the three sources. Notably, we treat the agent judge as a cost-efficient proxy for ranking articles rather than a major quality signal; quality claims rest on the human study alone.

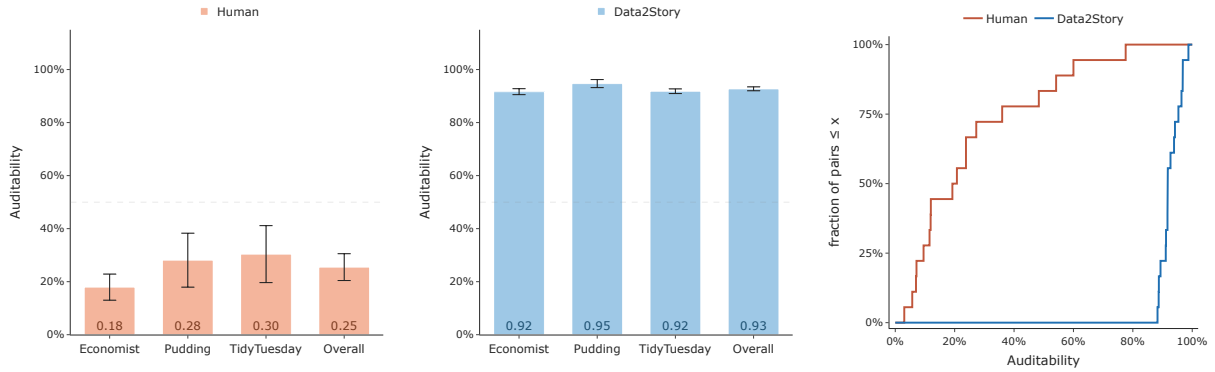
Transparency's Inspector lift is roughly 2.5× *the next-largest dimension and dwarfs the rest.* With the Inspector off, the agent's overall mean is 4.60 (human reference: 3.87), and on *Pudding* the two are identical (4.90 each), consistent with the human study. Opening the Inspector raises the overall mean to 5.10, a further ∼ 0.50. Figure 9b breaks the same three conditions down by rubric dimension: the effect concentrates almost entirely on *3-Transparency* (4.28→5.94, Δ=+1.67), with *4-Claim* a distant second (Δ=+0.67) and the remaining three dimensions barely shifting (Δ≤0.11). The Inspector thus buys a single dominant transparency channel plus a modest claim–data assist, with little spillover onto visual, narrative, or insight.

The agent judge preserves the human ranking at a fraction of the cost. A practical question is whether the cheaper agent judge can stand in for the 53-reviewer study on the same articles. The two judges rank articles together (ρ=0.44, p<.01; Figure 9c), and almost every point (29/34) sits above the y=x line, so the agent keeps the human ordering while scoring both Data2Story and human articles higher in absolute terms. **The agent judge is a usable stand-in for the ranking the human study produces, at a fraction of the cost.**

4.3.4 Verifiability analysis: auditability rather than factuality

Figure 10 (a,b) reports machine-checkable provenance coverage across publication sources. For Data2Story articles, 93% of visible claims resolve to a traceable binding between the rendered text and its upstream evidence. Human reference articles ship no accompanying code, so by construction most claims cannot be checked this way; the Codex verifier has to guess at a plausible reproduction on its own from the raw data and the published text alone. Thus, text-only audit recovers such a binding for 25% of claims. This makes sense as human-written statements are written for general readers and rarely attach a line of code or a traceable source to each claim, whereas our Inspector

question bank probes for exactly that. It is worth noting that **they measure whether a claim carries a verifiable provenance trail, not whether it is factually correct.** The gap therefore reflects the availability of machine-checkable provenance, not the quality of human journalism.

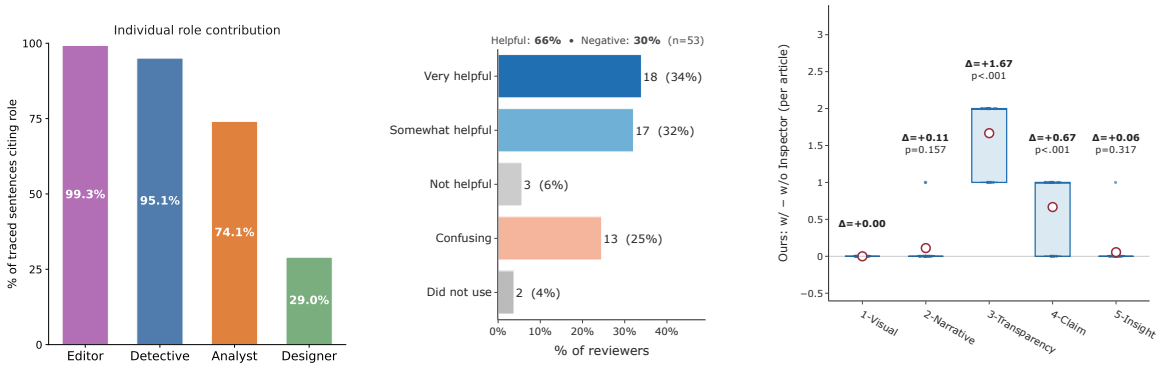


(a) Human, per source. **(b)** Data2Story, per source. **(c)** Empirical CDF over all 18 articles.

Figure 10 | Auditability between Data2Story-generated and human-written articles. Per-source means with SEM error bars for human (a) and Data2Story (b); empirical CDF over all 18 articles (c).

All three sources show a wide and significant gap. The gap is narrowest for Economist, whose briefing-style articles foreground more explicit and standard numerical analysis. This makes the likely human findings easier to anticipate, because many insights can be inferred from visible statistics, comparisons, and trends. In contrast, the gap is widest for Pudding, whose scrollytelling pieces often center on creative editorial ideas and qualitative framing rather than enumerated sub-population statistics. These ideas are less formulaic and therefore harder to guess from the pre-registered questions alone.

Figure 10c shows the empirical distribution of article auditability. Data2Story articles concentrate in a tight band near the top of the auditability axis, while the human distribution is spread more broadly. This suggests that the auditability that Data2Story offers is largely the pipeline itself rather than of any particular reference article; claims are bound to upstream evidence by construction, whereas in human's articles the same kind of binding only appears when the author chose to expose it.



(a) Individual role contribution. **(b)** Human participants' votes on whether the Inspector was useful. **(c)** Within-article Inspector effect on different rubrics.

Figure 11 | Analysis of Inspector effect. Human participants' usefulness ratings of the Inspector (a), and Agent judges inspector-related gains across rubric dimensions (b).

4.3.5 Analysis of different roles

The inspector subagent exposes the per-sentence provenance produced by four different roles: Detective (sourcing), Analyst (computation), Designer (chart authoring), and Editor (storytelling). Figure 11a reports per-role coverage across all articles: *Editor* 99.3%, *Detective* 95.1%, *Analyst* 74.1% and *Designer* 29.0%. These shares reflect each role's working character more than the data itself: Editor and Detective participate in nearly every traced sentence — every claim is storyboarded, and Detective's search-heavy sourcing names at least one external reference; Analyst adds computation to roughly three quarters of sentences (the quantitative subset); Designer is selective, anchoring visual assets in about a third.

Effect of Inspector. Figure 11b reports how the n=53 reviewers experienced the Inspector, which attaches per-claim provenance information including analyst notes, code-line references, and source datasets—to the rendered article. Overall, 66% of participants found the Inspector helpful for forming their evaluations, only 3(6%) rated it not helpful, whereas 25% found it unhelpful or distracting. The main concern was that the provenance traces were sometimes dense and complex, linking each claim to multiple scripts, quoted sources, and data references.

Figure 11c isolates the Inspector behavioral effect: same article, same computer-use agent judge, only difference is whether the Inspector is open. The within-pair lift concentrates on *3-Transparency* (Δ=+1.67, paired, p<.001). This further highlights that *opening the Inspector lifts mainly the transparency rubric*.

4.4 Qualitative assessment: where human did better?

We examine individual examples from each publication source, comparing the articles produced by Data Journalist Agent with those written by human journalists. This qualitative view surfaces values that our numerical experiments miss. Across the paired set, the human edge shows up in three recurring forms: the editorial angle, the creative design, and the informative presentation.

(i) Editorial Angle. The human advantage we could not close is the angle that comes from outside the data. The Repair Cafés reporter (Table 9) frames repair as a matter of accountability, attributing failure to manufacturers that build "phones, cars and tractors" so that mechanics cannot access "diagnostic tools or broken parts" without them. Such a claim is reported rather than computed: it rests on expert testimony and on outside knowledge the dataset never holds. Working only from the table, Data2Story can rank what breaks (knives are saved far more often than printers), but it leaves the cause to the reader. This is the qualitative face of our coverage finding (§4.3). Across the paired set the agent recovers only about half of the human's editorial angle, because the other half lives in reporting it cannot reach.

(ii) Creative Design. On the *Pudding* pieces, human teams invest weeks of bespoke interaction the agent does not attempt. The Stand-Up Comedy article (Table 6) turns the transcript into the interface: "every line" of Ali Wong's special is on the page, and each laugh is marked beside its line as a circle scaled to its length. For the same material, Data2Story links out to a static YouTube thumbnail and summarises the set in standard charts. A similar gap appears in the Internet Boy Band Database (Table 7), which plays as an audio-visual jukebox of all 55 acts, its hand-animated members morphing from band to band as each act's song plays; the agent re-tells the same history in static charts behind click-to-play embeds. The numbers survive, but not the crafted experience built around them.

(iii) Informative Presentation. Even in a single static figure, human designers carry more meaning per frame. The space-race chart (Table 4) sets state against commercial launch providers on one timeline and folds in a second variable for free: each band is shaded lighter where launches failed,

with an annotation explaining why the Soviet count runs so high. Its satellites "lasted only a year and a half on average, compared with nine years for their American counterparts." Data2Story distributes the same material across many single-variable charts, so no one figure carries the story. The football-managers chart (Table 5) overlays managers and star players on one axis, placing Messi and Ronaldo at the high end where the chart's own annotation reads: "Star players' impact can reach ten points per season. Managers rarely add more than two." Our agent plots only the managers, and the comparison the headline promises never appears.

These cases show that the human contribution should not be understated. Data2Story leads on coverage, analysis, and auditable transparency, yet the reported angle, together with the hand-built craft behind a design or a chart, remains a human strength.

5 Discussion

We introduced Data Journalist Agent, a multi-agent framework that orchestrates specialised roles into a single virtual newsroom for end-to-end data journalism. Data2Story contributes two properties absent from prior approaches: an evidence-traceable Inspector that binds each number, quote, and asset to a specific code line or reference, and multimodal generative storytelling in which the agent reasons about audience needs before deploying sub-agents and tools that fit both the data and the reader. Across 18 samples paired with expert references, Data2Story receives favourable ratings from 53 human participants and from computer-use agent judges on both rubric dimensions and side-by-side preference, with the Inspector specifically improving data and method transparency.

We position Data Journalist Agent as a collaborator for human journalists: (i) agent-generated articles can augment the newsroom workflow by contributing creative multimodal assets and an auditability dimension that is rarely formalised. (ii) Beyond augmenting existing coverage, Data2Story opens a complementary path: surfacing specialised or niche datasets that human journalists rarely have the bandwidth to investigate in depth, turning overlooked data into accessible, verifiable stories. We hope this work moves us toward a trustworthy agentic data system. **Limitations.** Data2Story so far runs fully automatically. A more reliable design would let it take human feedback and adjust in the loop – exploring whether an agent can interpret reader feedback and revise as professionally as a journalist. Meanwhile, our multimodal storytelling offers a new perspective on presenting data, yet the depth a human writer brings to the written angle should not be underestimated, and we leave a direct comparison to future work.

References

[1] Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. Dsbench: How far are data science agents from becoming data science experts? *arXiv preprint arXiv:2409.07703*, 2024.

[2] Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024.

[3] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024.

[4] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2023.

[5] Zhiyu Yang, Zihan Zhou, Shuo Wang, Xin Cong, Xu Han, Yukun Yan, Zhenghao Liu, Zhixing Tan, Pengyuan Liu, Dong Yu, et al. Matplotagent: Method and evaluation for llm-based agentic scientific data visualization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11789–11804, 2024.

[6] Victor Dibia. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 113–126, 2023.

[7] Zichen Chen, Jiefeng Chen, Sercan Ö Arik, Misha Sra, Tomas Pfister, and Jinsung Yoon. Coda: Agentic systems for collaborative data visualization. *arXiv preprint arXiv:2510.03194*, 2025.

[8] Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: Benchmarking multimodal code generation for automated front-end engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3956–3974, 2025.

[9] Holly Rusch. Should AI cover your city council meeting? Prevalence of AI-generated articles summarizing public meetings grows in San Mateo County. San Mateo Daily Journal, 2025. Accessed: 2026-06-08.

[10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[11] Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. MindSearch: Mimicking human minds elicits deep AI searcher. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.20183.

[12] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. MMSearch: Benchmarking the potential of large models as multi-modal search engines. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.12959.

[13] Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

[14] Fan Nie, Junlin Wang, Harper Hua, Federico Bianchi, Yongchan Kwon, Zhenting Qi, Owen Queen, Shang Zhu, and James Zou. DSGym: A holistic framework for evaluating and training data science agents. *arXiv preprint arXiv:2601.16344*, 2026.

[15] Sirui Hong, Yizhang Lin, Bang Liu, Bangbang Liu, Binhao Wu, Ceyao Zhang, Chenxing Wei, Danyang Li, Jiaqi Chen, Jiayi Zhang, Jinlin Wang, Li Zhang, Lingyao Zhang, Min Yang, Mingchen Zhuge, Taicheng Guo, Tuo Zhou, Wei Tao, Robert Tang, Xiangtao Lu, Xiawu Zheng, Xinbing Liang, Yaying Fei, Yuheng Cheng, Yongxin Ni, Zhibin Gou, Zongze Xu, Yuyu Luo, and Chenglin Wu. Data Interpreter: An LLM agent for data science. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19796–19821, 2025.

[16] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[17] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

[18] Alexander Spangher, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. Do LLMs plan like human writers? comparing journalist coverage of press releases with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[19] Jonathan Gray, Lucy Chambers, and Liliana Bounegru. *The data journalism handbook: How journalists can use data to improve the news*. " O'Reilly Media, Inc.", 2012.

[20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[21] OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, 2025.

[22] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

[23] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*, 2025.

[24] Zhuofeng Li, Dongfu Jiang, Xueguang Ma, Haoxiang Zhang, Ping Nie, Yuyu Zhang, Kai Zou, Jianwen Xie, Yu Zhang, and Wenhu Chen. Openresearcher: A fully open pipeline for long-horizon deep research trajectory synthesis. *arXiv preprint arXiv:2603.20278*, 2026.

[25] Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. DataNarrative: Automated data-driven storytelling with visualizations and texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 19253–19286, 2024. arXiv:2408.05346.

[26] Shaolei Zhang, Ju Fan, Meihao Fan, Guoliang Li, and Xiaoyong Du. DeepAnalyze: Agentic large language models for autonomous data science. *arXiv preprint arXiv:2510.16872*, 2025.

[27] Sina Montazeri, Yunhe Feng, and Kewei Sha. PublicAgent: Multi-agent design principles from an LLM-based open data analysis framework. *arXiv preprint arXiv:2511.03023*, 2025.

[28] Natalie Grace Brigham, Chongjiu Gao, Tadayoshi Kohno, Franziska Roesner, and Niloofar Mireshghallah. Developing story: Case studies of generative ai's use in journalism. *arXiv preprint arXiv:2406.13706*, 2024.

[29] Sophia Cheng. When journalism meets ai: Risk or opportunity? *Digital Government: Research and Practice*, 6(1):1–12, 2025.

[30] Alexander Spangher, Tenghao Huang, Yiqin Huang, Lucas Spangher, Sewon Min, and Mark Dredze. A novel multi-document retrieval benchmark: Journalist source-selection in newswriting. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 180–204, 2025.

[31] Milad Alshomary, Grace Li, Anubhav Jangra, Yufang Hou, Kathleen McKeown, and Smaranda Muresan. Llms as science journalists: Supporting early-stage researchers in communicating their science to the public. *arXiv preprint arXiv:2601.05821*, 2026.

[32] Leixian Shen, Haotian Li, Yun Wang, and Huamin Qu. From data to story: Towards automatic animated data video creation with LLM-based multi-agent systems. In *IEEE VIS Workshop on Generative AI for Data Storytelling (Gen4DS)*, 2024. arXiv:2408.03876.

[33] Liliana Bounegru and Jonathan Gray. *The Data Journalism Handbook 2: Towards a Critical Data Practice*. Amsterdam University Press, 2021.

[34] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 2nd edition, 2001.

[35] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, San Francisco, CA, 2nd edition, 2004.

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

[37] Cole Nussbaumer Knaflic. *Storytelling with data: A data visualization guide for business professionals*. John Wiley & Sons, 2025.

[38] Sarah Cohen, James T Hamilton, and Fred Turner. Computational journalism. *Communications of the ACM*, 54(10):66–71, 2011.

[39] Nicholas Diakopoulos. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3):398–415, 2015.

[40] Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 2013. Unpublished manuscript.

[41] Alberto Cairo. *The truthful art: Data, charts, and maps for communication*. New Riders, 2016.

[42] H. Paul Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[43] Chris North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006.

[44] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[45] Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*, 2024.

[46] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

[47] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, volume 2024, pages 15585–15606, 2024.

Appendix

A Model Settings

Data Journalist Agent is based on Claude-code `opus-4.7`. We detail the tools employed in the Designer role. We use OpenRouter as the unified provider for all generative models, as summarized in Table 3.

Table 3 | Generative capabilities and the OpenRouter API model backing each tool.

Tool	Modality	API Model (version)
`openrouter-text2image`	Text → Image	`openai/gpt-5.4-image-2`
`openrouter-text2video`	Text → Video	`bytedance/seedance-2.0`
`openrouter-image2video`	Image → Video	`google/veo-3.1-fast`
`openrouter-text2music`	Text → Audio	`google/lyria-3-pro-preview`
`openrouter-embeddings`	Text → Vector	`qwen/qwen3-embedding-8b`

In Human-agent angle coverage, we use OpenAI's `text-embedding-3-small` for retrieval similarity calculation, then use `gpt-4o-mini` to decide matching.

In Computer-use agent as judge experiments, we use the OpenAI's browser-use `gpt-5.5-xhigh`.

B Rubric Evaluation Scoring Standard

In this section, we present the detailed scoring standard used in our rubric evaluation, which applies to both the human study and the agent judges. For each of the five dimensions, we provide detailed instructions for scores ranging from 1 to 7, where a score of 3 serves as our typical default.

Rubric Evaluation scoring guideline

Dimension 1 — Visual Design

What it evaluates: the artifact as a visual product — palette, typography, layout, whitespace, fit between chart type and the claim it supports, overall polish.

Score	Description
7	Indistinguishable from the best work published in the past year by top editorial graphics desks. Every chart is the optimal encoding for its claim. At least one design choice is so well-targeted that you can describe what trade-off it earns.
6	Coherent design system **plus** at least one design move you've rarely seen executed this well — name it specifically. No wrong chart types, no missing labels or legends.
5	Cohesive design with intentional palette, typographic hierarchy, and chart selection. At most one minor named weakness.
4	Polished but with two or three specific weaknesses you can point to (busy chart, label collision, palette drift, missing in-figure legend). No chart is the wrong tool for its claim.
3	Looks like a competent person used a charting library with sensible defaults. Nothing offensive, nothing memorable. **Typical score.**
2	At least one of: mismatched chart type, jarring palette, truncated/missing labels, cluttered layout, visible default-styling residue.
1	Reads as raw library output or a database UI. Aesthetically mishandled to the point of impeding reading.

Dimension 2 — Narrative & Pacing

What it evaluates: the author as a tour guide — hook, ordering, rhythm, ending.

Score	Description
7	A piece you remember the shape of years later. Hook, layered progression, **and** a structural move (reframe, inversion, reveal) that recasts everything before it. The closing line bends back to the opening. The argument has a one-sentence summary that is itself non-obvious.
6	Strong arc with a clear structural move and a closing reframe that wasn't already spoiled by the standfirst.
5	Clear arc, real ending, at most one minor lull.
4	Logical ordering, real ending, two or three minor lulls or one stretched section.
3	Three-act competence: setup → data → conclusion. Readable but unsurprising in shape; no structural move. **Typical score.**
2	Sections organized by topic, not by argument. Reader has to stitch meaning together.
1	No through-line. A list of facts or charts; the artifact does not act as a guide.

***For non-narrative artifacts (datasets, tools)**: score on whatever framing exists (intro, README, landing copy). Bare dataset with title only → 1. README that motivates use without an arc → 2. README with a clear analytic arc → 3. Higher requires explicit narrative scaffolding.*

Dimension 3 — Data & Method Transparency

What it evaluates: procedural credibility. Four things to look for:

1. **Sources cited specifically** — named dataset or URL, not "public data"
2. **Methodology described** — definitions, transformations, exclusions
3. **Data accessible** — link, download, repo, or in-page table
4. **Limitations acknowledged concretely** — specific percentages, specific exclusion rules

Score	Description
7	Fully replicable **and** audit-grade. All four components present; multiple are unusually deep (line-level code references, public repo, processed-data table, dedicated methods note). A motivated outsider could replicate without contacting the author.
6	Replicable in principle. All four components present **and** at least one is unusually deep.
5	All four components present; methodology and caveats are concrete with specific numbers/exclusions; data access at least gestured at (e.g., named CSVs).
4	Three of four components, with sources and methodology mandatory among them. Caveats present but light.
3	Sources cited, methodology gestured at briefly, no data access, caveats minimal. **Industry-typical.**
2	Sources mentioned without specifics ("government data"); no methodology; no caveats.
1	No sources, no methods, claims of unclear origin.

Inspector panel rule for this dimension

- **Panel `not present`**: rate based on the normal reading surface only.
- **Panel `disabled`**: only count sources, methods, and caveats visible in the **normal reading surface** (prose, footnotes, in-page methodology sections, etc.). Do not credit anything that requires the 🔍 button.
- **Panel `enabled`**: you may credit panel contents, **but only if you actually completed the exploration in §1.4**. Reference specific card numbers and tags you saw, not just the card's surface text.

Dimension 4 — Claim–Data Alignment

What it evaluates: substantive credibility — are claims actually supported by the data?

Score	Description
7	Surgical. Every quantitative claim is bounded by what the data can support; every place a reader might overinterpret is explicitly defused; uncertainty is quantified where relevant; chart encodings are unambiguous; correlation is never dressed as causation.
6	Precisely scoped, with at least one explicit sensitivity check or robustness gesture (alternative cut, what-if, named confound addressed quantitatively).
5	Precisely scoped. Quantitative claims bounded; major confounders named in prose; encodings unambiguous; no causal overreach.
4	Solid, with one or two minor stretches where an adjective or annotation slightly exceeds the data. Nothing structurally misleading.
3	Most claims supported; some loose framing; no major distortions. **Typical score.**
2	Several unsupported claims **or** at least one misleading visual element (truncated axis without flag, dual axis implying correlation, cherry-picked window, missing baseline).
1	Charts illustrate unproven claims; clear cherry-picking; encoding is materially deceptive.

Dimension 5 — Insight Value

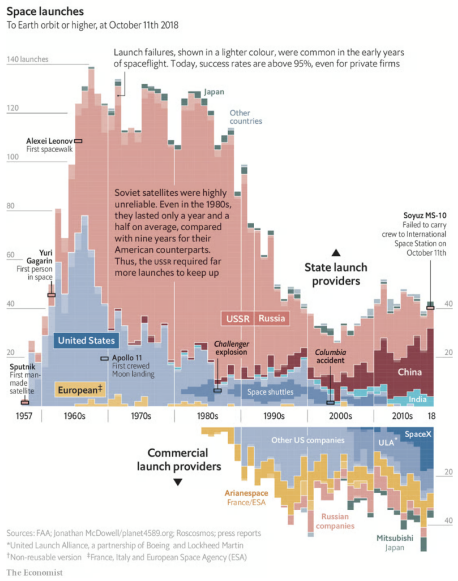
What it evaluates: whether the reader gains a non-trivial cognitive update.

Score	Description
7	Field-shaping for a domain reader. A finding or reframe a specialist would cite. Updates a prior held by people who already know the topic. The reader can articulate the change in one sentence weeks later.
6	Updates a domain prior. Counterintuitive finding, structural pattern not previously legible, or reframe that makes a familiar topic newly readable.
5	Sharpens a domain intuition meaningfully. Quantifies something a specialist suspected but had not seen pinned down at this resolution.
4	Sharpens a lay intuition. Quantifies or sharpens something the lay reader vaguely believed; goes beyond surface summary.
3	Precise common knowledge. Well-supported, but the conclusion is what an informed reader would have predicted before reading. **Typical score.**
2	Awareness, not judgment. Reader leaves knowing the data exists; no new judgment is formed.
1	No thesis. Restates conventional wisdom or has no claim at all.

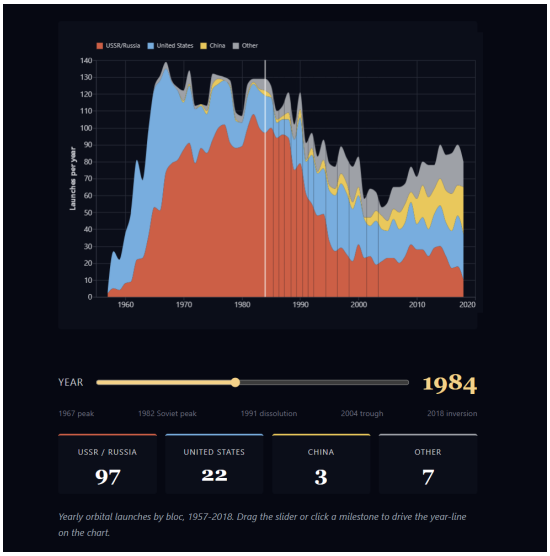
***Quick test**: if the takeaway is just "this dataset exists," cap at 2. If the reader doesn't reach synthesis or evaluation, cap at 3. If the update is meaningful only to a lay reader (not a specialist), cap at 5.*

Table 4 | The Economist: The space race is dominated by new contenders

Data	The Great Launch Inversion — 1957 to 2018
Category	Audio Artifact + Visual Artifact
Human [link]	Title: *The space race is dominated by new contenders*



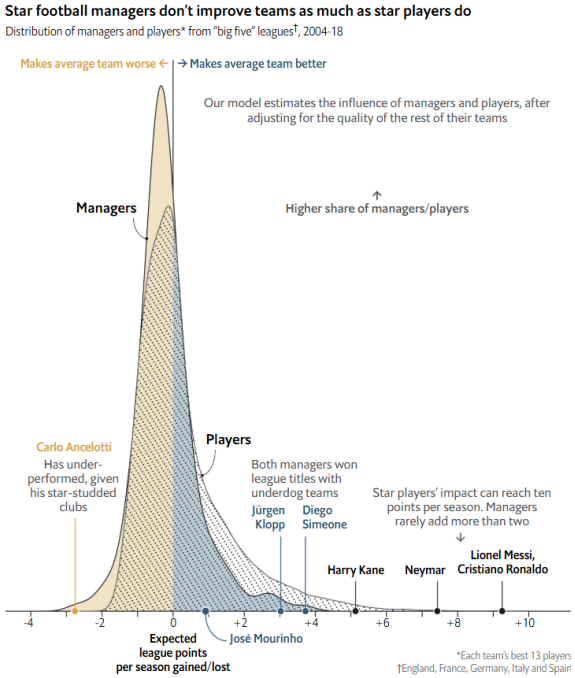
Ours [link]	Title: *The Great Launch Inversion*



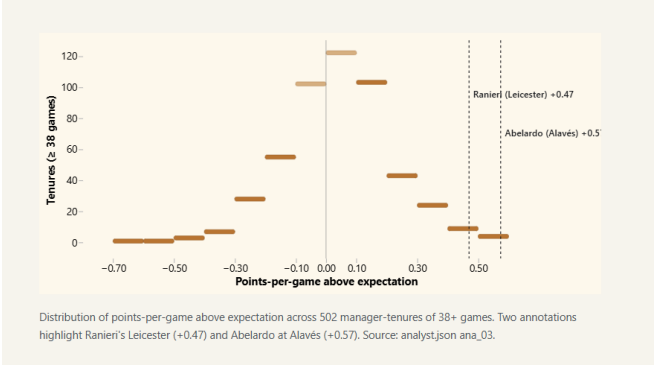
Analyze	In the human-written version, a large amount of information is densely packed into a single image. Key moments are annotated with descriptive text, making the chart richer in content and clearer in explanation. In contrast, the agent's version turns the image into an interactive chart, where users can slide along the year axis to view specific numbers for each year. However, it lacks the descriptive annotations found in the human version, so users can only access the raw figures without the surrounding context.

Table 5 | The Economist: Managers in football matter much less than most fans think

Data	Managers in football matter much less than fans think
Category	Audio Artifact + Visual Artifact
Human [link]	Title: *Managers in football matter much less than most fans think*



Ours [link]	Title: *Managers in football matter much less than fans think*



Analyze	In the human version, the visualization is clear and well-structured: it shows each manager's expected score, with prominent annotations marking both top-tier players and elite managers. This design makes the fine-grained information immediately readable at a glance. In contrast, the agent blog's chart does not present a clean curve; it reads more like a set of labels positioned along the y-axis of scores. The agent annotates only a few managers but omits the standout players, and the overall distribution is not visually salient.

Table 6 | The Pudding: The Structure of Stand-Up Comedy

Data	One Ten-Second Laugh — The Architecture of Ali Wong's Baby Cobra
Category	Video Artifact + Interactive Artifact
Human [link]	Title: *The Structure of Stand-Up Comedy*
Ours [link]	Title: *One ten-second laugh, and what holds it up*
Analyze	In the human-authored version, the video is embedded inline within the dense surrounding text and plays automatically, with a live transcript animating alongside the playback. The effect is polished and attention-holding, reflecting careful design work. The agent-generated version, by contrast, simply embeds a static YouTube iframe that requires the reader to click through and watch the video on YouTube itself.

Table 7 | The Pudding: Internet Boy Band Database

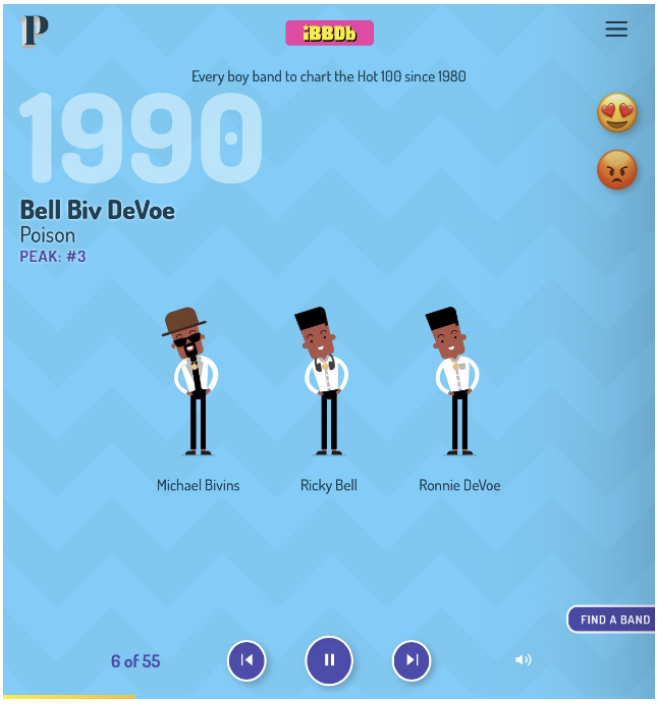
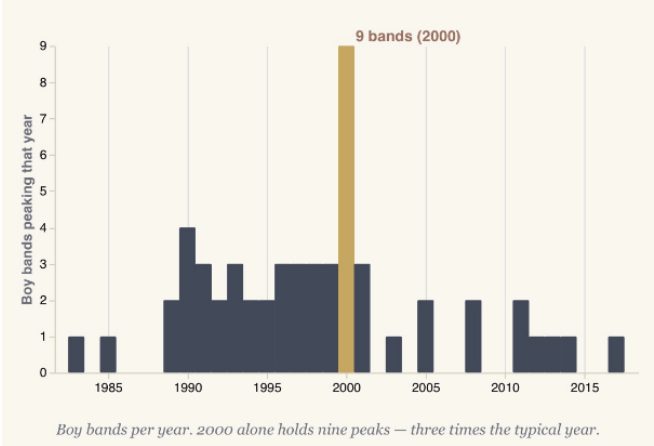
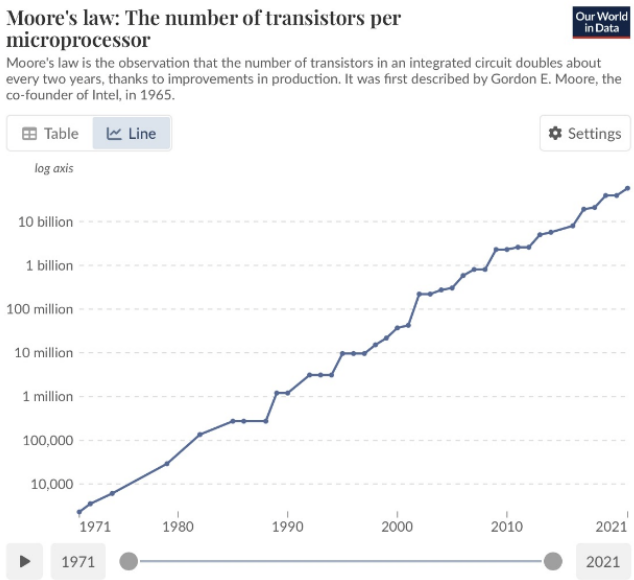
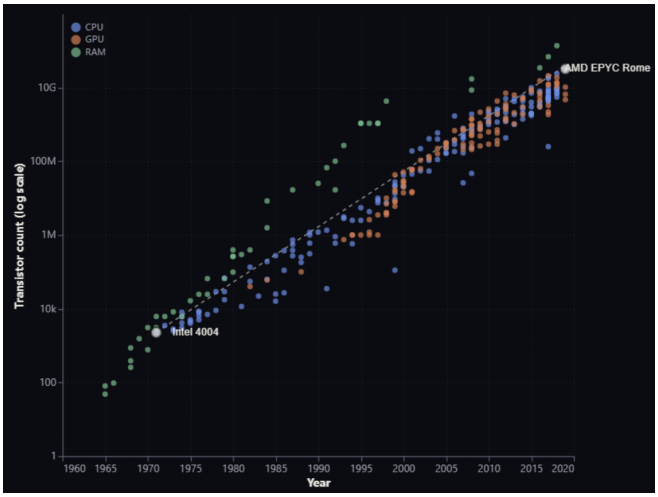
Data	The look you remember is one of four — Boy bands, by the data
Category	Audio Artifact + Interactive Artifact
Human [link]	Title: *Internet Boy Band Database* 
Ours [link]	Title: *The look you remember is one of four* 
Analyze	The human version is a music-driven immersive page: animated avatars for each band move in sync with autoplaying audio, and switching tracks swaps both the song and the illustrated lineup. It produces strong emotional engagement but only one band is visible at a time. The agent version constructs it as a static yearly histogram, exposing the early-2000s peak and the full distribution at a glance.

Table 8 | TidyTuesday: Moore's law: The number of transistors per microprocessor

Data	The forecast that aged — Moore's Law on the data
Category	Interactive Artifact
Human [link]	Title: *Moore's law: The number of transistors per microprocessor*



Ours [link]	Title: *A pencil line, drawn in 1975, that aged.*



Analyze	The human chart distills Moore's Law to a single log-scale line from 1971 to 2021, framed with explicit prose context and a Table/Line/Settings toggle, optimizing for legibility and citation. The agent version expands the same domain into a three-class scatter (CPU, GPU, RAM) on the same log scale, annotating the Intel 4004 and AMD EPYC Rome endpoints to anchor this massive growth. The agent surfaces between-class structure that the human design intentionally hides, at the cost of denser overplotting and higher reader effort.

Table 9 | TidyTuesday: A Growing Number of 'Repair Cafes' Are Popping Up Around the World to Curb Consumer Waste

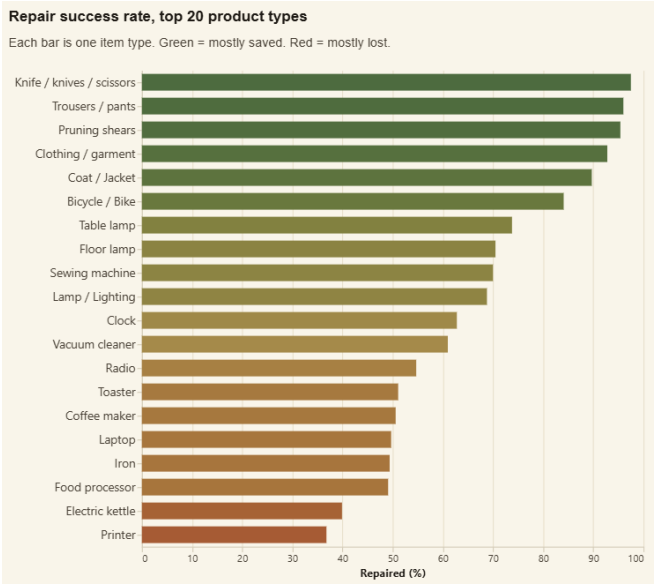
Data	What 178,749 repair attempts say about design
Category	Textual Artifact
Human [link]	Title: *A Growing Number of 'Repair Cafes' Are Popping Up Around the World to Curb Consumer Waste*

Right to Repair: Experts say a few major hurdles are preventing repair culture from dominating modern society.

For many products, such as clothes and appliances, buying a new product can often be cheaper than fixing the one you own. The convenience of free, rapid shipping also means that you are likely to get a new product faster than your old one can be fixed if the repair is more complicated —or there may not even be someone nearby with the expertise to fix it, Fromer said.

On top of this, some products, such as phones, cars and tractors, are produced in a way that prevents mechanics from accessing diagnostic tools or broken parts without taking it to the manufacturer. For example, repair advocates have pointed out that Apple AirPods are extremely difficult to disassemble and repair without damaging the casing, Vox reports. States including

Ours [link]	Title: *What 178,749 Broken Things Tell Us*



Analyze	The human version is qualitative and explanatory: it embeds repair culture inside a causal story where cheap replacement, fast shipping, scarce local expertise, and manufacturer lock-in act as structural barriers to repair. There is no chart; the prose is the artifact. The agent version produces a ranked bar chart of repair success rates across the top twenty product types, drawn from a large multi-country dataset. A diverging green-to-red palette encodes the outcome directly, turning textiles and hand tools into success cases and printers and electric kettles into failures.

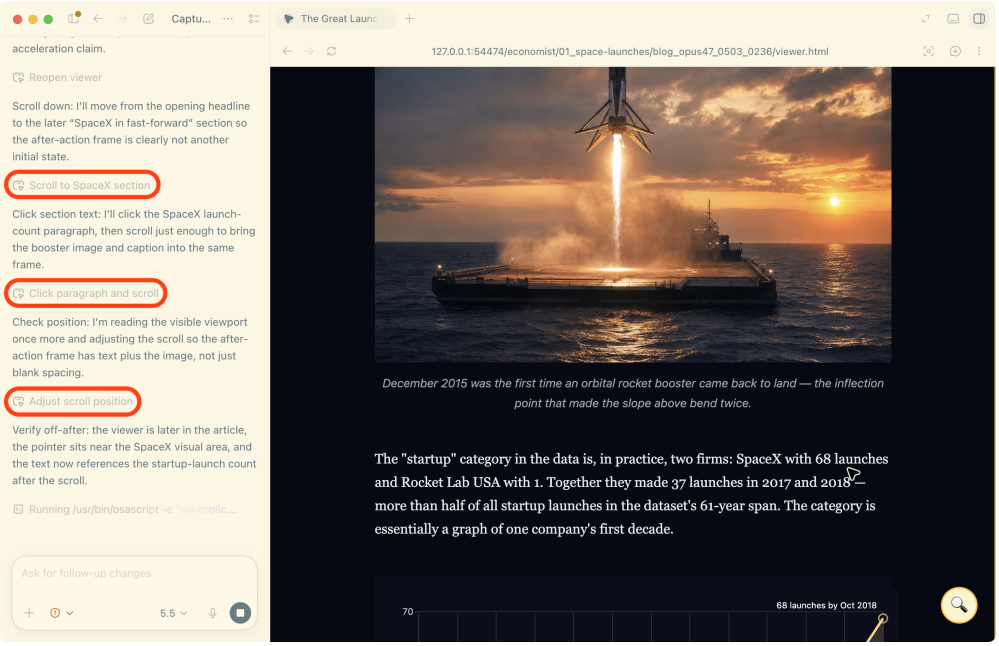
C Agent-as-Judge demonstration

We illustrate how a computer-use agent reads a generated article and prepares its rubric judgements. The actions by computer-use agents are highlighted in red.

Table 10 | Agent-as-judge, *Inspector-off* run on *The Space Launches*.

Initial state

The judge loads the article and observes the introductory animation, mirroring a human reader's first encounter with the page.



Reading the article

The agent then traverses the body via batched scroll-and-screenshot loops, accumulating a visual record of the prose, charts, and stat callouts in the natural reading order.

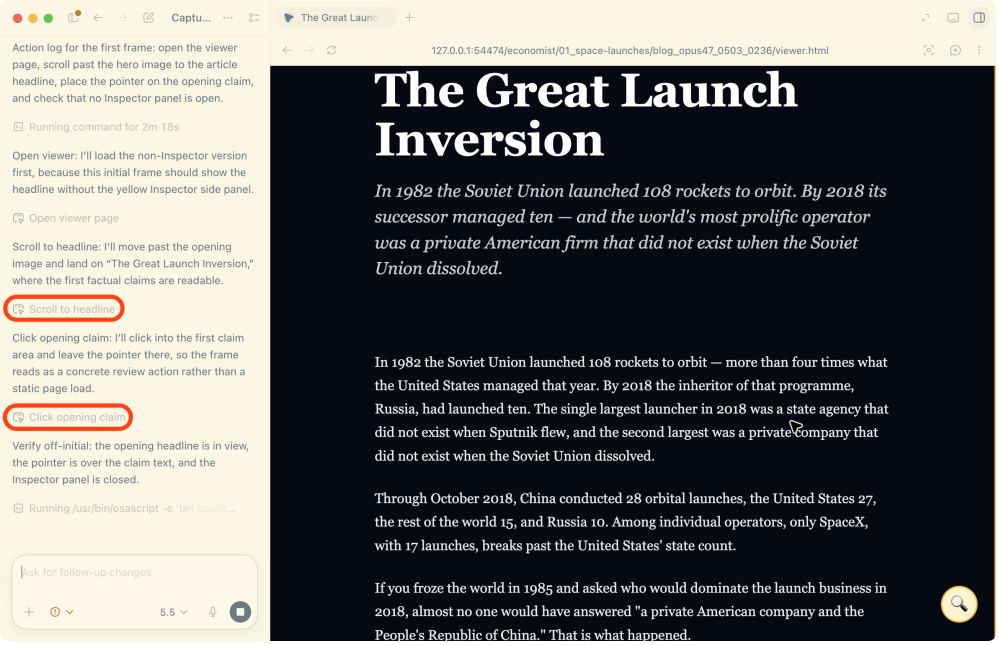
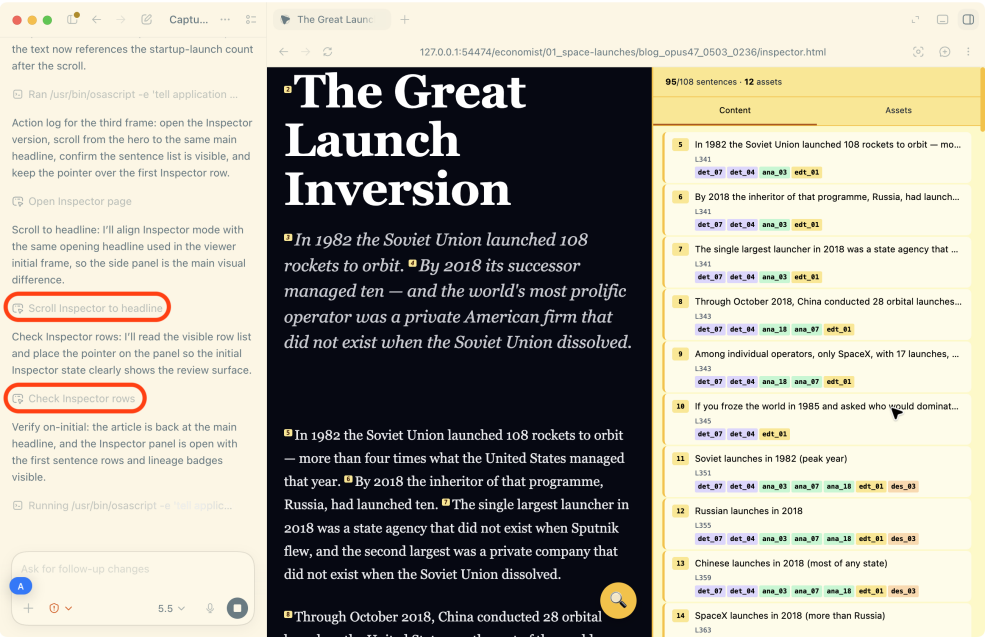
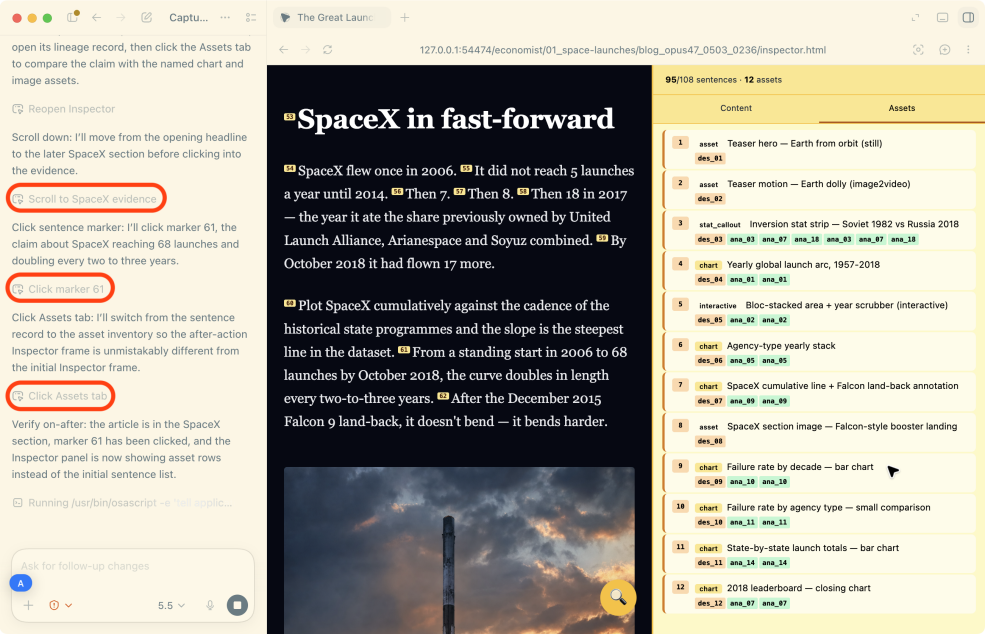


Table 11 | Agent-as-judge, *Inspector-on* run on *The Space Launches*.

Initial state	On arrival, the Inspector panel is already open. It exposes the article as two structured views: a list of every annotated sentence with its lineage badges, and a list of every named asset (chart, callout, interactive element) the article renders.



Reading the article with the Inspector	After reading the body, the agent navigates between the Inspector's two views (the action stream shows it locating and clicking the asset tab, then capturing what it reveals) to verify how each rendered claim and each visual asset traces back to its source (code lines, data tables, or external links) before issuing scores.