Data Journalist Agent: Transforming Data into Reproducible Multimedia Story

Kevin Qinghong Lin
University of Oxford

Batu EI
Stanford University

Yuhong Shi
University of Oxford

Pan Lu
Stanford University

Philip Torr
University of Oxford

James Zou
Stanford University

Abstract

Data tells stories that shape society, and the data journalist's job is to turn raw information into a piece that a non-expert reader will actually finish. A single high-quality new feature routinely takes a newsroom team weeks including hunting for context, running statistics, choosing an angle, designing visuals. Recent agents are individually capable at each step: automated data-science agents close the analysis loop, and scientific-writing agents synthesise long-form drafts. *But can an agent serve as a data journalist end to end?* We introduce **Data Journalist Agent (Agent-J)**, an agentic harness that orchestrates specialised roles into a single virtual newsroom. Agent-J offers two distinguishing properties over prior approaches. **(i) Evidence-traceable inspector.** The numbers, opinions and assets are grounded in a specific code line or external reference (*e.g.,* a URL), so verifiability is built into the fact itself: every claim is auditable. **(ii) Multimodal generative article.** Agent-J reasons about what its readers will want to read and interact with, then deploys multimodal tools so the article fits both the data and the audience (such as an interactive map with zoom for a geography piece or an audio clip for a music piece) rather than collapsing every dataset onto plain charts. We evaluate Data Journalist Agent on 18 article, each paired with the originally published expert-written piece from diverse topics and publication source, along three axes: **(a) Reproducibility**, where a Codex verifier re-executes every statement based on the data and checks that the rendered numbers are reproducible; **(b) Human–agent alignment**, measuring statement-matched recall and precision against the paired human article; and **(c) Rubric evaluation with human and agent judges** across 53 human participators and computer-use agents as judges. The results show that Agent-J achieves score comparable to human-authored articles on both rubric-based evaluation and side-by-side preference. Moreover, Agent-J's Inspector substantially improves the transparency of data and methods, making auditability explicit and measurable. We hope this work moves data journalism toward a reproducible, auditable, data intelligent system. Our **live demo** is available at data-journalist-agent.github.io.

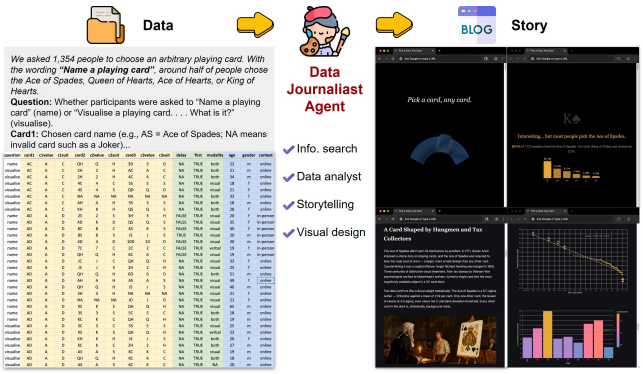


Figure 1: Data Journalist Agent turns a raw dataset into a verifiable, multimedia article by integrating information seeking, data analysis, storytelling, and visual design.

ACM Reference Format:
Kevin Qinghong Lin, Batu EI, Yuhong Shi, Pan Lu, Philip Torr, and James Zou. 2026. **Data Journalist Agent: Transforming Data into Reproducible Multimedia Story**. In *ACM Conference on AI and Agentic Systems, May 26, 2026, San Jose, California.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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. 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. A small newsroom team can spend weeks on a single high-quality article. Recent agents are individually capable at each of these steps such as automated data-science agents [1–4] close the analysis loop end to end: they profile a dataset, run the right statistics, and return a defensible number with code that reproduces. Design agents [5, 6] generate visual artifacts (such as websites) from an instruction. *But can an agent serve as a journalist end to end, taking a raw data all the way to a finished, reader-facing article?*

However, building an agent for end to end data journalism is non-trivial. In practice, producing a data-driven article requires humans weeks of coordinated effort spanning data collection, analysis, writing, and design. The task is fundamentally **multi-disciplinary**, demanding the simultaneous exercise of analytical, narrative, design that rarely co-exist in a single contributor which is why new is typically the product of a coordinated newsroom team. Beyond

CAIS '26, San Jose, CA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

effort, the article must also be **trustworthy**: every number, quote, and visual asset should trace back to a program or a source URL. This is a particularly challenging requirement for language agents, which are prone to hallucination.

Motivated by this, we introduce **Data Journalist Agent (Agent-J)**, a multi-agent framework that orchestrates seven specialised roles into a single 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, most notably, an Inspector that traces every element 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) Evidence-traceable Inspector.** To ensure the output is grounded in verifiable evidence, we introduce a dedicated agent that links every element of the published article (*i.e.,* numbers, quotes, and visual assets) back to its provenance (*i.e.,* a specific line of code, a data source, or an external URL). This makes the resulting article *auditable*: a reader can verify any claim by tracing its provenance without having to trust the agent that produced it. **(ii) Multimedia generative article.** Rather than formatting articles as plain text or static documents, we argue that an article should be multimedia and interactive. Thus, an HTML like website is preferable, we let a designer reason about the topic and what its readers will want to see and interact with. For example, geographic data is rendered as an interactive map with zoom and per-region tooltips; music or comedy data is paired with an embedded audio or video clip; The result is a generative media that engages the reader in an easy way.

To validate the effectiveness of Data Journalist Agent, we collect 18 data samples from three representative publication sources (the Economist, the Pudding, and TidyTuesday), each paired with the corresponding expert-written piece. For a comprehensive assessment, we design metrics along four complementary axes. **(a) Reproducibility** uses a cross-family coding agent to validate claims by verifying every statement such as excutating code or search reference source. **(b) Human–agent alignment** uses an LLM extractor to pull factual claims from both sides, and reports similarity-matched coverage such as recall, precision and F1-score. **(c) Rubric evaluation with human judges** asks 53 participators to score paired versions blind on five rubric dimensions and pick the preferred one overall. **(d) Agent as judge** introduces a computer-use agent that navigates the rendered page like a reader and scores it on the same rubric.

Our experiments show that Agent-J produces multimedia-rich articles that are visually appealing and content-rich. On human studies, the resulting articles achieve rubric-based scores and side-by-side ratings on par with human-authored counterparts. Moreover, Agent-J makes the auditability of data and methods explicit and measurable, an attribute that human articles, however carefully crafted, usually do not natively provide. Our work shows the potential for agent-made articles to approach human-made ones in quality, while contributing an auditability dimension that human workflows rarely formalize. We position Agent-J as a *collaborator* rather than a replacement for human journalists: in the future, human provide the perspective, while agents take over the repetitive work of gathering, analysis, and rendering.

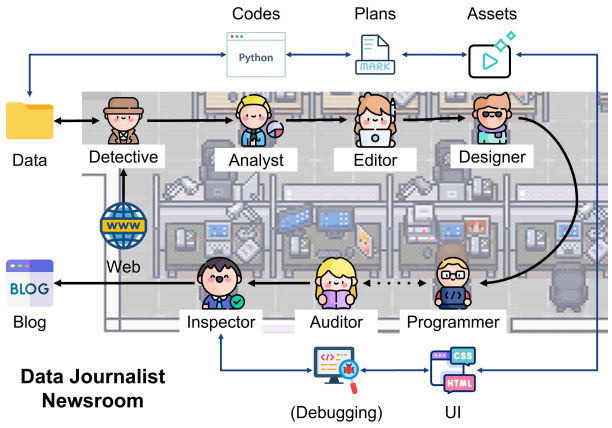


Figure 2: The Data Journalist Agent pipeline. A raw dataset flows through specialist roles: *Detective* (context), *Analyst* (statistics and code), *Editor* (narrative plan), *Designer* (visuals), and *Programmer* (HTML), then passes the *Auditor* (debugging) and the *Inspector* (evidence binding) before publication.

2 Data Journalist Agent

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 $\mathcal{R}$ via web search into an enriched corpus $\mathcal{R} \cup \widetilde{\mathcal{R}}$, where $\widetilde{\mathcal{R}} \xleftarrow{\text{Web search}} \mathcal{R}$ 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 often relies on dozens of statistics to support 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 findings $\mathcal{A} = \{a_i\}$ and supporting code $C = \{c_i\}$ with $a_i \xleftarrow{c_i} \mathcal{R} \cup \widetilde{\mathcal{R}}$, where each finding a_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 $\mathcal{S} \xleftarrow{\text{LLM}} \mathcal{A}$ that ranks each item by priority, selects the items worth keeping, and drafts a paragraph-level prose outline. Each statement s_i in $\mathcal{S}$ is annotated with the upstream items it draws on, $s_i \sim (a_i, c_i)$.

Designer. The right medium depends on the data: maps for geography, audio for music, video for events, interactive widgets for counter-intuitive findings. A pipeline that always emits a chart-and-prose template will fail on most datasets. The Designer focuses on choosing, for each section, the medium that best fits the data and the reader, recording the alternatives it considered and dropped so each choice is auditable rather than opportunistic. Given the Editor's section plan, it produces per-section visual specifications and the corresponding asset calls.

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 based on the previous artefacts, $\mathcal{U} \leftarrow \{\mathcal{S}, \mathcal{V}\}$. The Programmer does not generate any new facts or numbers; it simply quotes the upstream artefacts $\{\mathcal{S}, \mathcal{V}\}$ and assembles them into the complete interactive article $\mathcal{U}$.

Auditor. The rendered HTML may still harbour visual or structural defects: overlapping elements, broken charts, missing assets, or unresponsive interactions. These defects can quietly undermine an otherwise well-grounded story. The Auditor therefore inspects the rendered page, $\mathcal{U}^{\dagger} \xleftarrow{\text{Inspect}} \mathcal{U}$, flags such issues, and either sends the page back to the Programmer for repair or, once it passes, forwards the audited article $\mathcal{U}^{\dagger}$ to the Inspector.

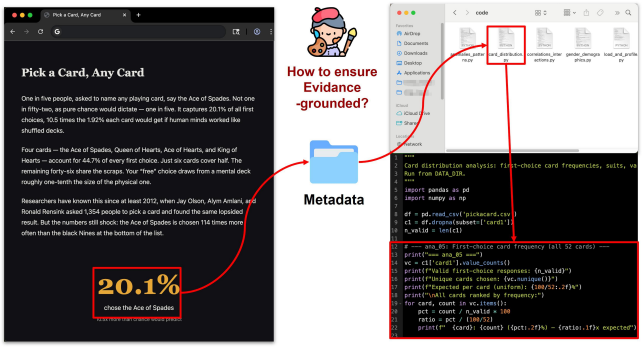


Figure 3: Illustration of the Inspector. It links every output statement back to its evidence, such as the source code file and the specific line that produced it.

How to ensure article reproducible: 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). We therefore introduce the Inspector, which closes this loop at the level of individual items.

The upstream agents each contribute atomic units of evidence. The Detective contributes a context $\widetilde{\mathcal{D}} = \{d_i\}$, where each d_i is a context item with a source URL. The Analyst contributes findings $\mathcal{A} = \{a_j\}$ paired one-to-one with code $\mathcal{C} = \{c_j\}$, so that every a_j is supported by the script c_j that produced it. The Editor contributes a statement $\mathcal{S} = \{s_k\}$, where each s_k is a paragraph with upstream pointers, and the Designer contributes specifications $\mathcal{V} = \{v_\ell\}$, where each v_ℓ is a per-section visual. Together these form the pool of upstream evidence $\mathcal{E} = \widetilde{\mathcal{D}} \cup \mathcal{A} \cup \mathcal{C} \cup \mathcal{S} \cup \mathcal{V}$. The Inspector decomposes the audited page into a set of visible statements $\mathcal{U}^{\dagger} = \{u_m\}$, where each u_m is a sentence, chart, or interactive element. It then binds every statement $u_m \in \mathcal{U}^{\dagger}$ to the evidence in $\mathcal{Z}$ that grounds it, and emits the bound page as a self-contained viewer (as shown in Figure 3).

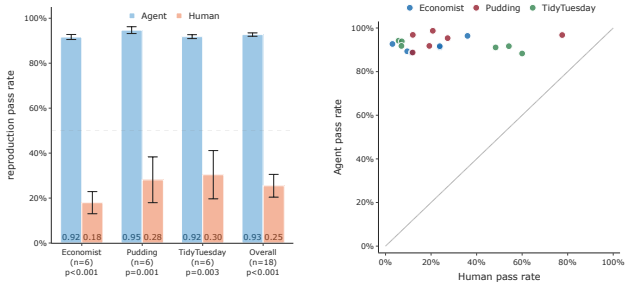
A skeptical reader can click any statement u_m to surface the upstream evidence behind it, or inversely click any piece of evidence $z \in \mathcal{Z}$ to surface the statements and visuals it supports. The result is a page where truthfulness is not asserted but *traceable*: any final number can be followed back through the Programmer, the Designer, and the Analyst, to the original data file and the line of code that produced it.

3 Evaluation

We evaluate Data Journalist Agent 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 following axes: domain (science, media, sports, politics, health, and others), temporal coverage (spanning 2018–2026), and data modality (time series, panel, and tabular etc). 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.

To comprehensive evaluate the Data Journalist Agent, we design four evaluation prototypes: (i) Reproducibility (ii) Human-agent alignment (iii) Rubric Evaluation with Human as Judge (iv) Agent as Judge. please refer section A for details.

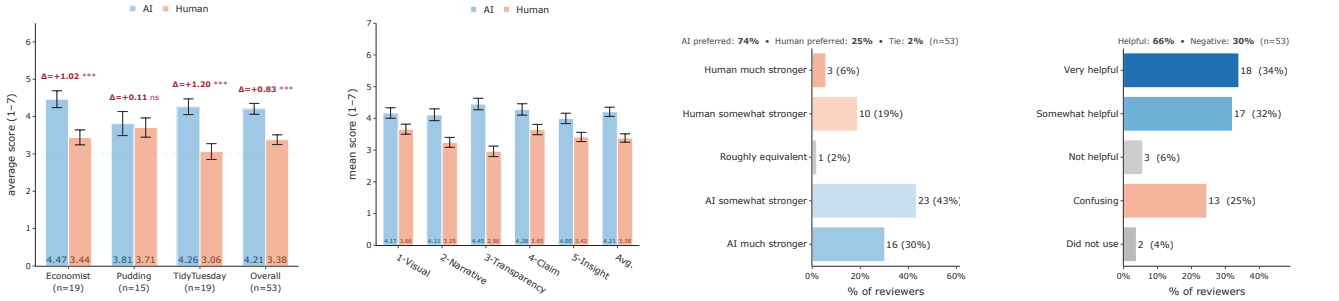
3.1 Experiment Results



(a) Reproduction pass rate. **(b) Distribution of pass rate.**

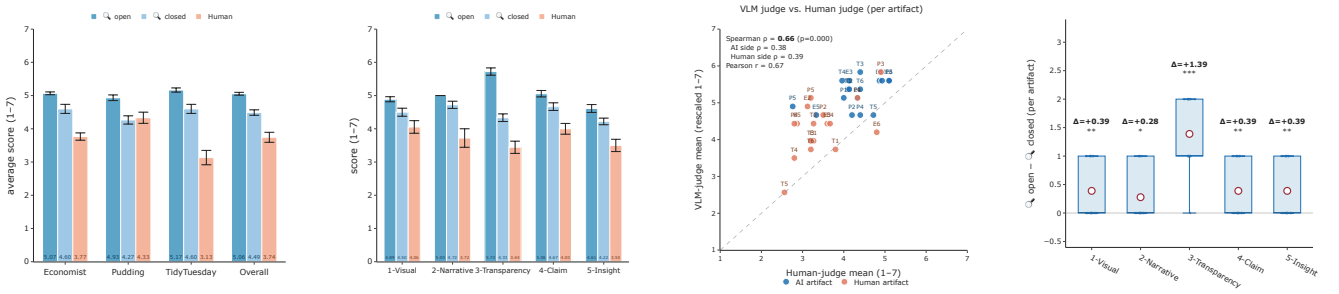
Figure 4: Reproduction vs. Human-Agent Alignment.

Reproduction Analysis. *The CodeX reproduces 93% of Agent's articles while the human's articles only 29% as without trajectories.* Figure 4a aggregates pass rates across the 18 outcomes. Pooled over all 1,706 agent attempts and 1,437 human attempts, the Data Journalist Agent article passes 1,581 checks (92.7%) against 484 for



(a) By source category. **(b) By rubric dimension.** **(c) Overall pairwise preference.** **(d) Human voting for inspector.**

Figure 5: Human as Judge (n=53 participators) across (a) overall score, (b) rubric dimension and (c) binary preference. We ask human participator to provide feedback for the inspector in (d).



(a) By source category. **(b) By rubric dimension.** **(c) VLM vs. human judge.** **(d) Gains by inspector.**

Figure 6: Agent as Judge under inspector open mode, close mode and human blog, reported by overall score (a) and rubric score (b). (c) Evaluation alignment between human and Agent artifact. (d) We report the rubric gain by inspector mode.

the human reference (33.7%); on per-pair means the gap is Δ=+0.64 (t=11.11, p<.001). The disparity is **not because human statements are wrong**—they are written for general readers and simply do not surface the same set of granular numerical facts that the pre-registered question bank probes. Moreover, we emphasise that this metric distinguishes **auditability** from **factual correctness**: professional data journalism rarely accompanies every claim with a line of code or a traceable source, so the reproducibility gap likely measures how openly evidence is exposed rather than whether the underlying claims are correct. The agent show the potential that it produces text whose claims align almost one-to-one with the underlying data, by construction of its pipeline.

The difference is significant on the most numerical Economist. Figure 4a also breaks the comparison down by source. *Economist* (Δ=+0.55, p<.001), *Pudding* (Δ=+0.76, p<.001), and *TidyTuesday* (Δ=+0.61, p<.01) all show a wide and significant gap. The *Pudding* margin is the widest because its scrollytelling pieces foreground a single editorial thesis with bespoke design and lean on qualitative framing rather than enumerating sub-population statistics — the very statistics our pre-registered questions ask about. *TidyTuesday* sits in the middle, and *Economist*'s briefing-style pieces, which are the most numerically explicit of the three sources, narrow the gap by raising the human floor.

When asked to pick one, most human participator pick agent-made. After viewing both versions, each reviewer gave a single overall preference (Figure 5c). 74% picked Data Journalist Agent as stronger, 25% picked the human, and 2% called it a tie (p<.001 under a sign test). The per-dimension lead is therefore large enough to flip the holistic "which is better" judgment at a 3:1 ratio.

The agent has consistent evaluation with humans. We also run a computer-use agent judge that navigates the rendered page and scores it on the same rubric, under three conditions: inspector-open, inspector-closed, and human reference (Figure 6). The score is highest with the inspector open (5.06), drops when it is closed (4.49), and is lowest on the human reference (3.74). The two judges agree on artifact ranking (Spearman ρ=0.66, p<.001), confirming the VLM judge is not hallucinating the gap.

The Inspector significantly improves perceived transparency. We isolate the contribution of the inspector panel from two angles (Figure 5d). *Subjective:* 66% of reviewers found the panel at least somewhat helpful, against 31% who did not. *Behavioral:* holding everything else constant, opening the inspector raises the agent judge's score by Δ=+1.39 on *Transparency* (p<.001) but only ∼ +0.35 on the other four dimensions, exactly the asymmetry expected if the panel's value is procedural credibility rather than overall polish.

References

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

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

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

[4] Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. 2024. DSBench: How Far Are Data Science Agents from Becoming Data Science Experts? *arXiv preprint arXiv:2409.07703* (2024).

[5] Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. 2025. 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)*. 3956–3974.

[6] Sukmin Yun, Haokun Lin, Rusiru Thushara, Mohammad Q Bhat, Yongxin Wang, Zutao Jiang, Mingkai Deng, Jinhong Wang, Tianhua Tao, Junbo Li, et al. 2024. Web2code: A large-scale webpage-to-code dataset and evaluation framework for multimodal llms. *Advances in neural information processing systems* 37 (2024), 112134–112157.

A Evaluation Metrics

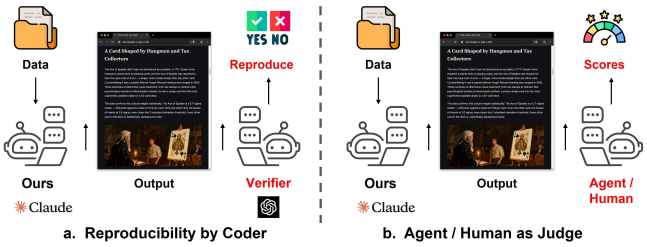


Figure 7: Two complementary evaluation protocols for Data Journalist Agent. (a) Reproducibility: a verifier agent attempts to reproduce Data Journalist Agent's output from the same input data, yielding a binary judgment of whether the article can be faithfully reproducible. (b) Agent / Human as Judge: an agent or human evaluator provide rubric scores Data Journalist Agent's output against the human-written reference, producing graded quality assessments.

Reproducibility. To verify that the published narrative is faithfully grounded in the underlying data, we replay every article with an independent coder (OpenAI `codex-GPT-5.4`). From each article, we extract the set of factual statements $\mathcal{S}$, which fall into two categories: (a) *computed 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, as shown in Fig. 7, left; and (b) *externally 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. We report the average reproducibility rate as pass verification.

Notably, in reproducibility experiments, verifier agents have access to the original dataset when evaluating human-written articles, rather than the article text alone. For agent-generated articles, verifiers additionally receive the full reasoning trajectory (by our Inspector) — a form of provenance made possible by Data Journalist Agent's evidence-grounded design.

Human-Agent Alignment. For every paired human–agent article, we measure how much overlap exists between the human-written reference and the Data Journalist Agent output. Using an LLM extractor, we pull out the set of factual claims from the human reference $\mathcal{H}$ and from the agent output $\mathcal{A}$ (one claim per statement, deduplicated). We then match claims across the two sides via a two-stage pipeline: `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 this check.

This gives us two directional coverage scores:

- **Human-in-Agent** $\mathrm{P}(\mathcal{A} \mid \mathcal{H})$: the fraction of human claims that the agent also surfaces. *Did the agent catch what a journalist would catch?*
- **Agent-in-Human** $\mathrm{P}(\mathcal{H} \mid \mathcal{A})$: the fraction of agent claims that also appear in the human article, indicating how anchored the agent stays to the human-curated angle.

Formally,

$$\mathrm{P}(\mathcal{A} \mid \mathcal{H}) = \frac{|\mathcal{H} \cap \mathcal{A}|}{|\mathcal{H}|}, \qquad \mathrm{P}(\mathcal{H} \mid \mathcal{A}) = \frac{|\mathcal{H} \cap \mathcal{A}|}{|\mathcal{A}|},$$

where $\mathcal{H} \cap \mathcal{A}$ denotes the set of claims matched across both sides. A high $\mathrm{P}(\mathcal{A} \mid \mathcal{H})$ means good recall against the reference; a low $\mathrm{P}(\mathcal{H} \mid \mathcal{A})$ means the agent strays from the human angle, which can reflect either useful breadth or off-topic noise. We report both per pair and an aggregate F1 over the pooled claim sets.

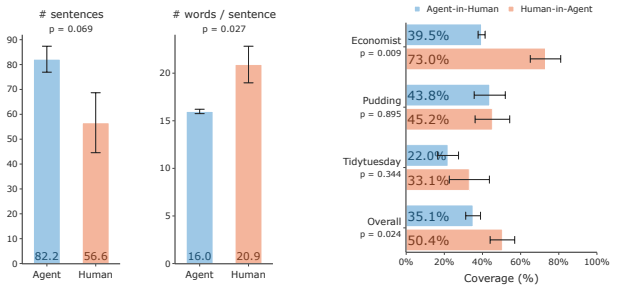
Rubric Evaluation & Human as Judge. A article is ultimately meant to be *read*, so the most direct evaluation is to put it in front of human reviewers. We recruit 53 reviewers via Prolific; each is assigned one Data Journalist Agent–human pair (presentation order randomised, source labels hidden) and scores both versions on five rubric dimensions on an integer 1–7 scale: **(i)** Visual Design, **(ii)** Narrative & Pacing, **(iii)** Data & Method Transparency, **(iv)** Claim–Data Alignment, and **(v)** Insight Value. After viewing both, each reviewer also expresses a binary preference indicating which version they prefer overall.

Agent as Judge. We also consider using a model as judge. However, a article is an interactive website: standard VLM judges only perceive 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 use a computer-use agent as judge, which navigates the rendered page like a human reader and scores it along the same rubric dimensions as our human studies.

B Distribution of article elements

Textual distribution. Before examining the article content, a natural first check is whether Data Journalist Agent writes at human scale. Across the 18 paired articles (Figure 8a), the agent uses 1.45× as many sentences but each is 0.77× as long, so total writing volume comes out comparable while the same material is broken into

shorter, more granular statements. Data Journalist Agent matches the human authors on overall volume but writes in finer chunks.



(a) Num. of sentence per article and Avg. words per sentence. **(b) Human-made v.s. agent-made article's claim coverage.**

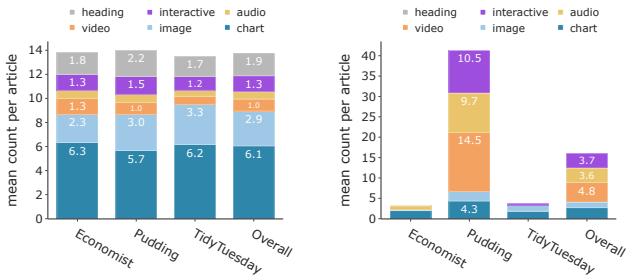
Figure 8: Textual distribution (left) and Content coverage (right) across 18 samples, reported by "mean ±SEM" with p value.

Claim coverage. Matching textual statistic is one thing; covering the same ground is another. As shown in Figure 8b, Claim-level coverage points clearly one way: about half of the human article (50.4%) lands in the agent's article, while only a third (35.1%) of agent sentences map back. The pattern is source-shaped, and each gap has a clear cause: it is widest on 'Economist' short briefings (73.0% vs. 39.5%), whose narrow single-topic scope (typically standard statistic or chart) makes them easy for the agent to predict and cover in full before writing well past the source; it stays uniformly lower on 'Pudding' and 'TidyTuesday', whose source articles either carry a single editorial thesis the agent does not fully reproduce ('Pudding's creative long-form scrollytelling) or span multiple subtopics in the same dataset across far greater length ('TidyTuesday'). Data Journalist Agent reliably absorbs and rewrites the source, but *reproducing a human author's narrative arc remains the harder problem.*

Multimedia assets distribution. Beyond text, every article might carries multimedia elements, leading the article visual diverge sharply. We classify multimedia elements by six categories: heading (big short title), interactive, audio, video, image, and chart. As illustrated in Figure 9a, Data Journalist Agent'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 9b 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. *Data Journalist Agent robustly produces every modality across topics, whereas human designers vary their distribution substantially with editorial style.*

C Qualitative Analysis

In this section, we visualize each example from different publication source to compare the articles made by Data Journalist Agent or by human.



(a) Data Journalist Agent. **(b) Human-made.**

Figure 9: Multimodal media assets distributions (*e.g.*, video, image, audio, interactive, etc) between Data Journalist Agent (left) and human (right).

Table 1: The Economist: The space race is dominated by new contenders

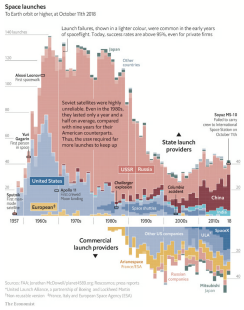
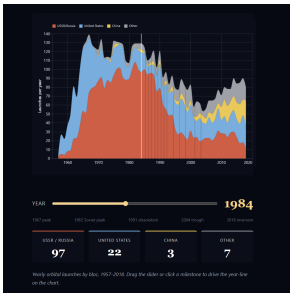
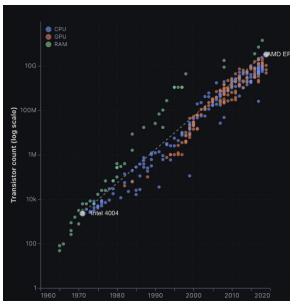
Data	The Great Launch Inversion — 1957 to 2018
Category	Audio Artifact + Visual Artifact
Human [link]	Title: *The space race is dominated by new contenders* 
Ours [link]	Title: *The Great Launch Inversion* 
Analyze	In the human-made version, a large amount of information is densely packed into a single 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 2: 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 3: 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 endpoints to anchor the regression. The agent surfaces between-class structure that the human design intentionally hides, at the cost of denser overplotting and higher reader effort.