InSight: A Benchmark for Agentic Claim Verification in Interactive Visualizations

Maeve Hutchinson[1], Syed Mahbubul Huq[1], Mohammad Albinhassan[2], Radu Jianu[1], Aidan Slingsby[1], Pranava Madhyastha[1,3,2]

[1]City, University of London, [2]Imperial College London,
[3]The Alan Turing Institute

Correspondence: {maeve.hutchinson, pranava.madhyastha}@city.ac.uk

Abstract

Vision Language Models have demonstrated remarkable proficiency in interpreting static visual artifacts, but modern data analysis is inherently dynamic, requiring the active interrogation of interactive environments. Existing benchmarks are predominantly constrained to static imagery and one-shot question answering and fail to capture the epistemic demands of this domain, where evidence is frequently occluded, distributed across linked views, or conditionally revealed through user agency. In this paper, we introduce InSight, a benchmark for agentic claim verification over interactive visualizations. The dataset consists of 21,349 claims derived from human-authored analytical narratives and grounded in fully interactive web-based environments. Agents must navigate these environments to determine whether a natural language claim is supported, refuted or not verifiable given the available evidence. Unlike traditional evaluations, InSight treats interaction traces as intrinsic proxies for reasoning, enabling a rigorous audit of how models seek and synthesize visual evidence. We evaluate state-of-the-art models, revealing that interactive verification remains a non-trivial challenge. We release InSight at `https://github.com/maevehutch/insight`.

1 Introduction

The recent advances of vision-language models (VLMs) have shown impressive performance in chart interpretation and visual reasoning (e.g., Google, 2025b; OpenAI, 2025). These systems are being increasingly integrated into analytical workflows, evolving from mere curiosities into fully fledged intelligent assistants tasked with conducting visual analysis, generating design advice and captioning complex imagery (Kim et al., 2025; Kaur et al., 2025). However, while the deployment of these models has indeed accelerated, the premise that they possess a robust, ground-level understanding of visual data remains contentious. Recent empirical evidence suggests that despite their fluency, VLMs frequently lack fundamental *visualization literacy*, which is the ability to accurately read, understand, and interpret encoded data without relying on parametric knowledge. Hong et al. (2025) demonstrate that state-of-the-art models often fail to ground their answers in visual features, instead confabulating responses based on textual context or prior training data, particularly when confronted with modified or decontextualised charts.

This discrepancy between perceived capability and actual visual grounding is exacerbated by the limitations of current evaluation methodologies. The evaluation of these models has traditionally relied on static benchmarks, where the model is presented with a fixed image of a chart and queried via a question-answer pair (Methani et al., 2020; Masry et al., 2022) or a summarisation task or captioning task (Tang et al., 2023). While valuable for assessing perceptual acuity in closed settings, this static paradigm ignores the fundamental ecological validity of the process of data analysis where meaningful visualizations are rarely static artifacts. They are interactive environments, such as dashboards, faceted browsers, and exploratory tools, where information is conditionally revealed through user agency. In real-world analytical scenarios, evidence is frequently occluded behind tooltips, distributed across linked views, or accessible only through filtering and zooming.

In this paper, we introduce *InSight*, a benchmark for agentic claim verification over interactive visualizations. InSight reframes visual reasoning as a sequential decision-making process. We present an evaluation framework where an agent must actively navigate web-based visualization environments through executing actions such as clicking, hovering, and scrolling to verify data-related claims. Our work departs from most previous attempts in building visualization under-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 13968–13988
October 24-29, 2026 ©2026 Association for Computational Linguistics

standing benchmarks in two key ways. Firstly, unlike datasets that rely on synthetic templates or task setups (Methani et al., 2020; Masry et al., 2022), or crowd-sourced descriptions of static images (Akhtar et al., 2024), InSight is fundamentally grounded in human-authored analytical notebooks focused on real, analytical, data exploration. We extract claims directly from expert narratives and preserving the corresponding custom vega-lite visualizations (Satyanarayan et al., 2017), allowing us to preserve authentic analytical workflows rather than contrived proxies. Secondly, InSight treats the *interaction trace* as a first-class citizen. In traditional text-based fact-checking (Thorne et al., 2018; Aly et al., 2021) or static visual question answering (Methani et al., 2020; Masry et al., 2022), the reasoning process is often opaque, inferred only from post-hoc interpretability methods or chain of thought as a proxy. In InSight, the sequence of actions, such as, the mouse interaction or the navigation path taken, provides an explicit, traceable proxy for the model's attentional focus and reasoning strategy. This allows us to determine not only *if* a model is correct, but *whether* it engaged with the necessary evidence to justify its conclusion. Our experiments with state-of-the-art models reveal that interactive verification remains a difficult challenge for all models. We also demonstrate that interaction is particularly critical for falsification, where targeted visual evidence is required to overturn linguistically plausible but unsupported claims. We also observe that high-performing models exhibit distinct interaction signatures, engaging in deeper, more targeted exploration of the visual topology.

We make the following contributions: We formulate the task of *Interactive Visual Claim Verification*, extending the epistemic requirements of fact-checking to agentic, multimodal environments. We present *InSight*, an ecologically valid benchmark comprising over $21k$ claims grounded in human-authored interactive visualizations, explicitly designed to penalise passive perception. We provide a comprehensive evaluation of current VLMs, offering a granular analysis of how interaction strategies correlate with reasoning fidelity.

2 Background

2.1 Agentic Vision-Language Models

Recent work has explored VLMs deployed as agents that operate in interactive environments, such as web browsers, desktop GUIs, and simulated interfaces. Benchmarks such as OSWorld (Xie et al., 2024), MiniWoB++ (Furuta et al., 2024), VisualWebArena (Koh et al., 2024) and BrowserGym (Chezelles et al., 2025) frame perception and action as a sequential decision-making problem, where models must execute actions (e.g., clicking, scrolling, typing) to complete complex tasks. These efforts have driven rapid progress in general-purpose multimodal agents and highlighted the challenges of long-horizon planning, grounding, and robustness in realistic environments.

However, existing agentic benchmarks primarily evaluate task completion, with success defined by whether a goal state is reached. In contrast, our work focuses on a single, tightly scoped task, visual claim verification, where interaction is used explicitly to acquire evidence, enabling controlled analysis of how agents seek, evaluate, and integrate information across modalities.

2.2 Chart Understanding

A large body of work addresses question answering and reasoning over charts and visualizations. Early datasets such as FigureQA (Kahou et al., 2018) and DVQA (Kafle et al., 2018) rely on synthetically generated charts and templated questions, enabling controlled evaluation of specific reasoning skills but offering limited ecological validity. Later datasets, including PlotQA (Methani et al., 2020) and ChartQA (Masry et al., 2022), move toward greater realism by using real or semi-real data and more diverse visual styles, but still frame the task as one-shot question answering over static images.

Datasets such as ChartCheck (Akhtar et al., 2024) focus on chart reasoning through claim verification rather than QA, requiring models to determine whether a textual statement is supported or contradicted by a chart. This shift aligns more closely with real-world misinformation scenarios, but these datasets continue to rely on static visual inputs and typically involve claims written or mutated specifically for the benchmark. As a result, they do not capture how analysts actually interact with visualizations to uncover evidence, nor do they reflect the compositional, multi-view, and interactive nature of real analytical dashboards.

InSight differs from prior chart datasets in two key respects: it preserves interactive visualization environments rather than static renderings, and its claims are grounded in human-authored analytical narratives, rather than being synthetically constructed to fit a predefined schema. This allows the

benchmark to capture realistic tasks that are absent from existing chart understanding datasets.

2.3 Claim Verification and Fact-Checking

Claim verification has been extensively studied in text-only settings, most notably in datasets such as FEVER (Thorne et al., 2018), SciFact (Wadden et al., 2020), and FEVEROUS (Aly et al., 2021). These benchmarks emphasize evidence retrieval and textual entailment, introducing the Not Enough Information (NEI) label to capture epistemic uncertainty. In these datasets, relevant evidence is explicitly available in textual form and do not model the process of actively uncovering evidence and updating beliefs about claims.

Multimodal fact-checking datasets extend this paradigm to images. Datasets such as NewsCLIPpings (Luo et al., 2021) and AVerImaTeC (Cao et al., 2025) require models to assess the consistency across textual and image modalities often in closed-world settings where all evidence is contained in the input. These tasks emphasize cross-modal alignment and visual-textual reasoning but still rely on static inputs and do not support sequential evidence acquisition.

InSight bridges these lines of work by combining the epistemic framing of claim verification, including NEI supervision, with interactive visual environments. In this setting, the environment is partially observable with evidence distributed across views, or conditionally revealed through interaction, introducing challenges that are absent from both text-only verification and static multimodal fact-checking.

3 InSight: Task and Dataset

3.1 Task Definition

We define interactive claim verification over visualizations as a decision-making task in which a model or a human must determine the veracity of a natural language claim, c, in relation to an interactive visualization environment, $\mathcal{E}$.

Each instance in the dataset consists of a natural language claim made regarding the data presented in the environment, and the environment itself. The latter is rendered as a web-based interface containing multiple views (e.g., a dashboard) and interactive UI elements.

The objective of the task is to assign one of three labels to the claim: *True*, meaning the visualizations support the claim; *False*, meaning the visualizations contradict the claim; and *Not Enough Information (NEI)* meaning that the visualizations do not provide sufficient evidence to either verify or refute the claim.

The environment $\mathcal{E}$ is partially observable. Relevant visual evidence may be occluded behind interactions, distributed across multiple views or panels, or conditionally revealed based upon prior states (of interactions). Thus, to access this evidence, the model must execute a sequence of actions $a_1, ..., a_T$ from a fixed action space (e.g., clicking, hovering, scrolling) to update the environment state and generate new visual observations. The model operates within a turn-based interaction loop with a bounded budget of actions. We emphasise that no textual metadata regarding the underlying data is provided; all verification must be grounded in pixel-level evidence revealed through interaction.

We formalize this interactive claim verification task as a process of *sequential evidence acquisition*. As each action potentially exposes new information, the resulting interaction trace serves as an intrinsic proxy for the model's internal reasoning process, revealing what information the model deems relevant, how it navigates the topology of the visualization, and the stopping criteria it employs when judging evidence sufficiency.

3.2 Dataset Construction

InSight is derived from a corpus of human-authored analytical notebooks that combine interactive Vega-Lite (Satyanarayan et al., 2017) visualizations with natural language narratives describing the analytical insights they reveal (Wood et al., 2019). These notebooks were authored independently by analysts with formal training in data visualization, who selected their own datasets, formulated analytical questions, and designed custom visualizations. As a result, the corpus reflects authentic analytical reasoning rather than constructed descriptions. We apply rigorous filtering criteria, retaining $N = 297$ notebooks for downstream processing.

The construction pipeline proceeds in four stages. Full implementation details, including all prompts, thresholds, and lexicons, are provided in Appendix A.

Stage 1: Span Extraction. We extract candidate spans from each narrative using Lundgard and Satyanarayan (2022)'s four-level semantic model, retaining only spans at levels 2 (statistical insights) and 3 (visual/perceptual insights), which express

content verifiable against visualizations. To ensure extraction stability, we perform three independent LLM runs per text segment (Wang et al., 2023) and consolidate spans via a multi-stage agreement procedure based on trigram overlap. Spans extracted by only a single run are discarded, and the final span text is constructed from tokens selected by at least two runs. This procedure filters out hallucinated extractions and semantically irrelevant text.

Stage 2: Claim Decomposition. Retained spans may contain compound statements or implicit references. Following Metropolitansky and Larson (2025), we decompose each span into atomic, fully decontextualized propositions that can be independently verified against the visualizations. A self-verification step confirms that each decomposed claim is entailed by the source text (Weng et al., 2023), and semantic labels are reassigned via majority vote across three passes. Claims that fail grounding verification or exhibit unstable labeling are discarded. The resulting claims constitute the ground-truth TRUE set.

Stage 3: Claim Mutation. To construct a balanced verification benchmark, we generate FALSE and NEI claims through controlled mutation of the true claims, following the contrastive construction paradigm of Thorne et al. (2018). We apply three mutation strategies, each targeting a distinct source of semantic change:

- **Antonym substitution** reverses directional or quantitative language (e.g., *increased* $\rightarrow$ *decreased*) using a curated antonym lexicon, producing FALSE claims.
- **In-lexicon argument substitution** replaces data arguments (categorical values, numerical quantities) with alternatives from the same dataset, producing FALSE claims.
- **Out-of-lexicon argument substitution** replaces arguments with values or attributes absent from the dataset, producing NEI claims that refer to entities for which no visual evidence exists.

All mutations are constrained by XML-style tagging of substitutable spans to prevent lexical drift and preserve structural similarity with the source claim. Crucially, each mutated claim is validated using a pretrained NLI model (He et al., 2023): FALSE claims must exhibit high bidirectional contradiction scores (> 0.9) with the original, while NEI claims must exhibit low entailment *and* low contradiction (< 0.2). Claims failing validation are discarded.

Stage 4: Human Validation. We conduct a human validation study with 13 expert annotators, who produced 475 annotations across 294 sampled claims. Annotators performed the same interactive verification task defined in §3.1, blind to the original labels and construction process. Raw agreement with the dataset labels is 81.3%. A Bayesian Dawid–Skene model (Dawid and Skene, 1979; Paun et al., 2018) applied to the annotation data recovers class prevalences within 3 percentage points of the ground truth, with a posterior mean accuracy of 66.2% (95% CI: [59.5%, 71.8%]) and an MAP estimate of 77.6%. These results confirm that the controlled construction and NLI-based validation procedures produce labels broadly consistent with expert human judgment, despite the inherent difficulty of the task.

3.3 Dataset Analysis

InSight contains 21,349 claims derived from 297 human-authored analytical notebooks. Claims are distributed across the three verification labels, with 41.6% True, 45.0% False, and 13.4% Not Enough Information (NEI) cases.

Claim properties. Claims are concise but non-trivial, with an average length of 18 tokens (median 17). 56.8% correspond to semantic level 2, relating to statistical and numerical insights (e.g., value retrieval, extrema), and 43.2% to level 3, relating to visual and perceptual insights (e.g., trends, patterns). This distribution reflects a mix of precise data lookup and higher-order relational reasoning, both of which are common in real analytical workflows.

Visualization diversity. The visualizations are authored in Vega-Lite (Satyanarayan et al., 2017), a declarative grammar based on composable marks and encodings. As a result, the space of visualization forms is not bounded by a predefined taxonomy. Across the corpus, visualizations combine a wide range of marks—bars, lines, points, areas, geoshapes, text, and rules—and frequently involve layering and multi-view composition. On average, each notebook contains 4.5 views and 3.7 visualization specifications, with 67% of notebooks exhibiting explicit multi-view composition. This means that part of the verification task is determining *which* visualization is most relevant to the claim, a challenge absent from single-chart bench-

marks.

Interaction characteristics. Interactivity is intrinsic to the visualization environments in InSight. The most prevalent mechanism is hover-triggered tooltips, which conditionally reveal details-on-demand, such as precise values, labels, or additional information. Click-based and point-selection interactions are also common, enabling filtering, highlighting, or cross-view coordination. Less frequent primitives (interval selection, zoom/pan) nevertheless play an important role for claims requiring range-based or spatial reasoning. Importantly, these statistics capture interaction capabilities designed into the visualizations themselves; scrolling and basic mouse actions are additionally available as part of the agent action space (§4.1).

Partial observability and interaction necessity. Every environment in InSight is partially observable, and therefore every claim requires exploration of the environment. Environments contain on average 4.5 views and 3.7 specifications (Figure 3), and cannot be rendered in a single fixed viewport, so the initial screenshot never exposes the full set of visualizations. Even when the visualization relevant to a claim happens to appear first, an agent cannot know this without inspecting the remaining views: it must establish that no other view supplies stronger, contradicting, or disambiguating evidence, and for NEI claims it must establish that no view supplies evidence at all. In addition, the most prevalent interaction mechanism, hover-triggered tooltips, hides precise values behind an action, so even claims about a visible chart typically cannot be resolved from the static rendering.

Additionally, visualizations here were built by analysts to *explore* their own datasets, and the claims were derived from insights they reached through interactive use, not from what a single frame conveys. Consequently, no evaluation claim is verifiable from the initial viewport alone, and a correct answer produced without interaction should be attributed to parametric knowledge or guessing rather than to visual grounding. This property is what allows InSight to penalize passive perception, and it motivates the design of the Interaction Efficiency Score in 4.2.

Comparison with prior datasets. Unlike existing chart understanding and claim verification benchmarks, which typically rely on static images or pre-extracted data, InSight jointly supports: (i) fully interactive visual environments, (ii) claims grounded in real-world analytical provenance rather than synthetic construction, (iii) explicit NEI supervision, and (iv) the collection of action traces during verification. See Appendix C for additional distributional analyses, including mark-type frequencies and specification density histograms.

4 Evaluation and Behavioural Analysis

4.1 Experimental Setup

Evaluation Environment. Each visualization environment $\mathcal{E}$ is rendered as an HTML page containing one or more interactive visualizations authored in Vega-Lite and executed in a Chromium-based headless browser using Playwright with a fixed viewport. This setup renders the visualization notebooks exactly as a human expert would see them and ensures consistent rendering and interaction behavior across models. The fixed viewport is deliberately smaller than the rendered page: because notebooks contain multiple views (§3.3), only a portion of the environment is visible at any time, and the remainder must be reached by scrolling, while details-on-demand (tooltips, selections) must be triggered by mouse actions. Every environment is therefore partially observable, and no claim in the evaluation set can be verified from the initial screenshot alone, regardless of which visualization contains the relevant evidence.

Crucially, we treat the web page as a purely visual surface. The model is provided only with the RGB screenshot of the current viewport; it does not have access to the underlying DOM tree, HTML source code, or accessibility tree. This design choice aligns with recent work in GUI agents (e.g. Qin et al., 2025; Luo et al., 2025), ensuring that the model must ground its reasoning in the visual rendering of data rather than parsing textual data representations embedded in the code.

Agent Action Space. The model interacts with the environment through a defined high-level action space $\mathcal{A}$, drawing on prior work in web and UI agent benchmarks (Xie et al., 2024; Qin et al., 2025, interalia). The action space mimics standard mouse and keyboard inputs, requiring the model to predict pixel coordinates (x, y).

We define the action space $\mathcal{A}$ as follows: *Navigation:* `scroll(dx,dy)` for precise pixel scrolling, and macro-actions such as `page_up(n)`, `page_down(n)`, and directional arrow keys (e.g., `arrow_right(n)`) to navigate the webpage. *Mouse Interaction:* `drag((x1,y1),(x2,y2))`, `click(x,y)`, `shift_click(x,y)`, and

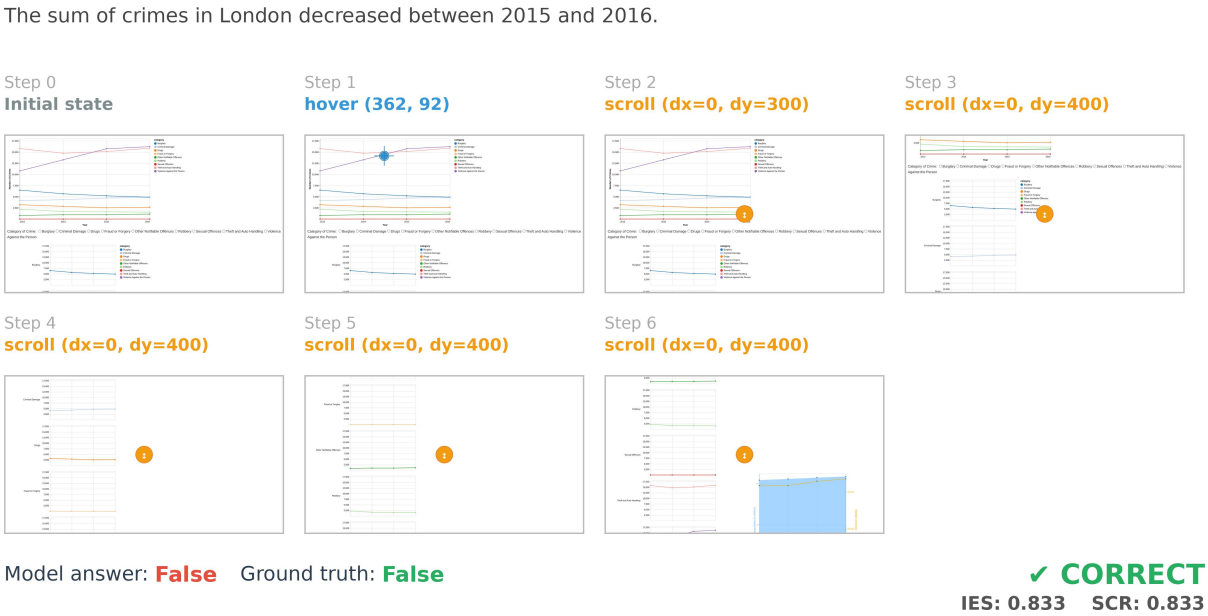

Figure 1: Interaction trace for Gemini 3.5 Flash verifying a false claim about London crime trends. The model hovers to inspect a tooltip, then scrolls through the multi-view notebook to locate relevant evidence, correctly predicting FALSE with an IES of 0.833.

`hover(x,y)`. These are essential for interacting with visualizations to reveal more evidence, such as tooltip inspection (hover), range selection (drag), or multi-selection (shift-click). *Termination:* `answer(label)`, where label $\in \{\text{True}, \text{False}, \text{NEI}\}$. This action terminates the episode and submits the verification decision.

Interaction Protocol. Models interact with the environment through a standard turn-based protocol. At each turn $t \leq T_{\text{max}}$, the model receives the current screenshot s_t and the claim c, and must output a single action $a_t \in \mathcal{A}$. If a_t is not a terminal `answer` action, the environment executes a_t, updates its state, and provides the next screenshot s_{t+1}. The process continues until the model outputs an `answer` action or the maximum turn budget T_{max} is reached.

4.2 Evaluation Metrics

We introduce an *Interaction Efficiency Score* (IES), designed to capture whether agents interact productively with the visualization environment while performing claim verification. Existing agentic benchmarks primarily evaluate task completion alone, whereas InSight explicitly treats interaction traces as part of the reasoning process. Our metric is conceptually related to Success weighted by Path Length (SPL) (Anderson et al., 2018) used in embodied navigation benchmarks, but adapted to browser-based visual evidence acquisition.

For an episode i, let $A_i \in [0, 1]$ denote answer accuracy, where $A_i = 1$ if the predicted label is correct and $A_i = 0$ otherwise. Let T_i denote the total number of actions executed by the agent, and let E_i denote the number of effective actions. We define an effective action as an action that produces an observable change in the environment state, i.e., interactions that reveal new visual evidence.

Formally, we compute effective actions as:

$$E_i = \sum_{t=1}^{T_i} \mathbf{1}[s_{t+1} \neq s_t] \tag{1}$$

where s_t denotes the observable environment state after action t, and $\mathbf{1}[\cdot]$ is the indicator function.

We then define the Interaction Efficiency Score as:

$$\text{IES}_i = A_i \cdot \frac{E_i}{T_i} \tag{2}$$

where $\frac{E_i}{T_i}$ is the state change ratio (SCR) and we report the dataset-level score as:

$$\text{IES} = \frac{1}{N} \sum_{i=1}^{N} A_i \frac{E_i}{T_i} \tag{3}$$

This formulation rewards agents that both answer correctly and interact productively with the environment. Incorrect answers receive a score of

0 regardless of interaction behavior, while correct predictions are additionally differentiated by the proportion of actions that meaningfully advance the observable state of the environment. IES thus provides a behavioural signal indicating whether an agent's interaction trajectory acquires visual evidence during verification, complementing accuracy rather than replacing it.

Correct answers without interaction. An agent that submits an answer at the first turn has $T_i = 0$ and $E_i = 0$; we define $E_i/T_i = 0$ in this case, so the episode receives IES $= 0$ even if the answer is correct. This is deliberate: since no claim is resolvable from the initial screenshot, a correct answer with no interaction reflects parametric knowledge or chance rather than visual grounding, which is precisely the failure mode IES is meant to expose. Qwen 3.5 0.8B, which never interacts, receives IES $= 0$ for this reason.

Scrolling. Scrolling almost always changes the rendered pixels and is therefore almost always counted as effective. This is by design, because environments contain more views than fit in a viewport, scrolling is the action by which an agent discovers where relevant evidence resides and by which it returns to a relevant view after surveying the alternatives.

NEI claims. For an NEI claim, the ideal trajectory is to explore the environment and find nothing. Such exploration still changes the observable state and is counted as effective. Since for NEI claims there is by construction no evidence directly relevant to the claim's veracity, no metric can score whether the answer *depended on* revealed evidence, therefore IES scores whether the agent looked. The same holds to a lesser degree for True and False claims, because the agent cannot know in advance which view is relevant, actions that reveal irrelevant views are still useful exploratory work in a partially observable environment, and we intentionally do not penalize them.

Limitation. IES credits any state-changing action regardless of whether the final answer used the revealed evidence. It therefore cannot fully rule out correct-but-ungrounded predictions by agents that interact but ignore what they see. IES identifies agents that do not interact at all, and ranks interacting agents by how much of their budget produces new observations. Finer-grained grounding would require claim-specific evidence annotations and is a natural extension we leave for future work.

4.3 Model Results

We benchmark a selection of closed-source and open-weight state-of-the-art models. We evaluate models on a stratified subset of 500 claims, a similar sample size to equivalent multi-turn agentic benchmarking (Xie et al., 2024). The subset consists of 167 True, 167 False, and 166 NEI claims. We do not benchmark across the whole dataset to avoid potential data contamination issues and reserve samples for future rigorous benchmarking (Deng et al., 2024).

We evaluate fourteen model configurations spanning four model families: Gemini, GPT, Gemma, and Qwen. For all models we use an interaction budget of 10, following other similar CUA benchmark action budgets (Xie et al., 2024). We additionally do ablations with interaction budgets of ($T_{\max} \in \{1, 25\}$) for some models to compare behavior across action budgets. Table 1 reports accuracy, interaction volume (T/N, the mean number of actions per episode), the effective action ratio (E/T), and the Interaction Efficiency Score (IES). Both accuracy and IES are reported overall and disaggregated by claim label.

Overall Performance At $T_{max} = 10$, GPT-5.5 achieves the highest overall accuracy at 57.2%, followed by Gemma 4 31B (47.2%) and Qwen 3.5 27B (45.0%). These results are notable given that random performance on a balanced three-class task is 33.3%, indicating that all models meaningfully exceed chance, but none approach human-level reliability. A clear scaling trend is visible within model families: Qwen accuracy degrades monotonically from 44.8% at 35B parameters to 34.0% at 0.8B, and Gemma drops from 47.2% (31B) to 31.2% (E2B). At the smallest scales, models barely exceed random baselines, suggesting that interactive visual reasoning imposes a minimum capacity threshold that sub-billion parameter models consistently fail to meet.

The Role of Interaction The $T_{\max} = 1$ setting serves as a *no-interaction baseline*. In this ablation, the model receives the claim and the initial viewport and must answer at the first turn, which reduces the task to static visual question answering over a partial view of the environment. Performance in this setting reflects what the model can infer from parametric knowledge, linguistic plausibility of the claim, and whatever happens to be visible in the first screenshot. The results are mixed. Gemini 3.5 Flash is slightly more accurate without

Table 1: Comparison of model efficiency and interaction metrics across model families. Accuracy and IES are reported overall and disaggregated by claim label.

Model	T_{max}	T/N	**Accuracy**	**Acc$_{\text{False}}$**	**Acc$_{\text{NEI}}$**	**Acc$_{\text{True}}$**	E/T	**IES**	**IES$_{\text{False}}$**	**IES$_{\text{NEI}}$**	**IES$_{\text{True}}$**
Gemini 3.5 Flash	25	9.33	50.0%	54.8%	50.6%	44.6%	56.7%	31.0%	29.87%	36.10%	27.03%
Gemini 3.5 Flash	10	6.27	41.6%	45.8%	45.8%	33.3%	61.3%	24.93%	23.62%	32.63%	18.63%
Gemini 3.5 Flash	1	-	44.2%	33.1%	73.5%	26.2%	-	-	-	-	-
GPT-5.5	10	3.13	57.2%	59.6%	59.0%	53.0%	45.4%	26.98%	29.24%	23.98%	27.70%
Gemma 4 31B	10	3.16	47.2%	40.4%	63.9%	37.5%	43.8%	20.57%	15.12%	28.48%	18.13%
Gemma 4 31B	1	-	45.4%	26.5%	84.9%	25.0%	-	-	-	-	-
Gemma 4 E4B	10	0.89	43.2%	18.7%	88.0%	23.2%	18.3%	6.87%	2.81%	15.66%	2.18%
Gemma 4 E2B	10	0.16	31.2%	7.8%	71.7%	14.3%	5.0%	0.90%	0.00%	2.71%	0.00%
Qwen 3.6 35B	10	1.78	44.8%	42.8%	62.7%	29.2%	21.9%	9.07%	7.54%	12.67%	7.01%
Qwen 3.5 27B	10	1.57	45.0%	38.0%	66.3%	31.0%	17.7%	7.57%	5.65%	10.20%	6.87%
Qwen 3.5 27B	1	-	46.2%	35.5%	79.5%	23.8%	-	-	-	-	-
Qwen 3.5 9B	10	0.50	38.0%	50.6%	26.5%	36.9%	9.56%	3.52%	4.67%	3.82%	2.08%
Qwen 3.5 2B	10	0.24	37.0%	30.7%	51.8%	28.6%	4.10%	1.17%	1.20%	0.24%	2.06%
Qwen 3.5 0.8B	10	0.00	34.0%	21.7%	54.8%	25.6%	0.00%	0.00%	0.00%	0.00%	0.00%

interaction (44.2% vs. 41.6%), as is Qwen 3.5 27B (46.2% vs. 45.0%), whereas Gemma 4 31B drops from 47.2% to 45.4% when interaction is removed. For the two models that do not benefit, the gains at $T_{\max} = 1$ are within a few points and suggest that unrestricted interaction can be counterproductive when a model lacks a plan for exploration and may commit on partial evidence, or lose track of the claim across turns. That the pattern does not hold for Gemma indicates the effect is model-specific rather than a property of the task, and it is worth noting that the accuracy gap between the two settings is small for all three models, so parametric knowledge and the initial viewport account for a large share of the accuracy that interaction-capable models achieve.

The Gemini budget ablation is informative here because it is non-monotonic: 44.2% at $T_{\max} = 1$, 41.6% at 10, and 50.0% at 25. Under the 10-action budget Gemini is the most active model, and many of its trajectories are still mid-exploration when the budget forces an answer or when the model chooses to stop. A partially completed survey of a multi-view environment appears to be worse than none, presumably because the model has seen enough to abandon its prior but not enough to replace it. Given 25 actions it uses 9.33 on average, completes more of its exploration, and reaches the highest accuracy of any Gemini configuration, with IES rising from 24.93% to 31.0% even though its effective action ratio falls slightly, consistent with later actions being used to return to and inspect views already discovered.

The interaction traces reveal a pronounced disparity in how models allocate their action budgets. Gemini 3.5 Flash is the most active agent, averaging 6.27 actions per episode with an effective action ratio of 61.3%, indicating that a majority of its interactions produce observable state changes. GPT-5.5 is comparatively economical, averaging 3.13 actions but achieving the highest accuracy, suggesting more deliberate and targeted exploration. At the other extreme, Qwen 3.5 0.8B averages zero actions per episode, defaulting to immediate answer submission without any environmental interaction, a behaviour that reduces the task to uninformed guessing from the initial viewport.

Interaction Efficiency The IES metric disentangles accuracy from interaction quality, rewarding models that both answer correctly and interact productively. GPT-5.5 leads with an IES of 26.98%, followed closely by Gemini 3.5 Flash (24.93%) and Gemma 4 31B (20.57%). The gap between accuracy and IES rankings is informative: while Gemma 4 31B achieves comparable accuracy to Gemini 3.5 Flash, its lower effective action ratio reduces its IES, indicating that a substantial proportion of its actions fail to advance the observable environment state, meaning it is not making use of the visual information.

Small models exhibit near-zero IES scores. Gemma 4 E2B achieves 0.90% and Qwen 3.5 0.8B achieves 0.00%, indicating that even when these models occasionally produce correct answers, they do so without meaningful environmental engagement. This finding validates the design of IES as a metric that penalizes correct-but-ungrounded predictions, which accuracy alone cannot detect.

Label-Specific Analysis Disaggregating IES by claim label reveals asymmetric difficulty across verification categories. For GPT-5.5, IES is relatively balanced across labels, suggesting robust interaction strategies across all claim types. In contrast, most other models exhibit a characteristic pattern where NEI claims yield the highest IES. Gemini 3.5 Flash achieves IES_{NEI} = 32.63% but only IES_{True} = 18.63%, and this gap is even more pronounced in smaller models such as Gemma 4 E4B (IES_{NEI} = 15.66%, IES_{True} = 2.18%).

This pattern admits a plausible explanation: verifying a True claim requires locating and correctly interpreting specific visual evidence, whereas recognizing that evidence is insufficient for an NEI claim may require less precise interaction; the model needs only to fail to find confirming or contradicting evidence across the environment. Conversely, False claims, which require discovering evidence that contradicts a linguistically plausible statement, are particularly challenging for weaker models, indicating difficulty to productively acquire disconfirming evidence. This is consistent with prior observations in textual fact-checking that falsification is cognitively harder than verification, and extends that finding to the agentic multimodal setting (Thorne et al., 2018).

Actions and Behavioral Signatures Analyzing the actions taken by the models reveals qualitative differences in exploration strategies. Gemini 3.5 Flash employs the broadest action vocabulary, with substantial use of scroll (2.51 per sample), hover (1.42), click (1.15), and page_down (0.73), as well as non-trivial use of arrow keys and drag actions. This diverse repertoire enables it to achieve the highest effective action ratio. GPT-5.5 similarly prioritises hover (1.07) and scroll (1.02) but uses fewer total actions, consistent with a more efficient exploration strategy.

In contrast, smaller models exhibit a smaller range of action types taken. Gemma 4 E2B and Qwen 3.5 0.8B use almost exclusively zero actions across all types, while intermediate models such as Qwen 3.5 9B show limited use of scroll and click but negligible use of hover, which is a critical action for tooltip inspection. Given that hover-triggered tooltips are the most prevalent interaction mechanism in the InSight environments (Appendix C), this omission likely explains these models' inability to access conditionally revealed evidence.

Failure Modes Several systematic failure patterns emerge from the interaction traces. First, *premature commitment*: smaller models frequently terminate after one or two actions, submitting answers before engaging with the environment. Second, *unfocused exploration*: some incorrect predictions by larger models involve many actions with low effective action ratios, suggesting exploration that fails to target relevant visual elements. Third, *action-type rigidity*: mid-range models tend to default to scrolling and clicking without employing hover or drag actions, missing evidence encoded in tooltips or selection-based interactions.

These findings collectively demonstrate that interactive claim verification is not simply a harder version of static visual question answering. It requires coordinated perception, hypothesis formation, and action selection, which are capabilities that scale with model capacity but remain far from saturated even at the frontier.

5 Conclusion

We have presented InSight, a benchmark for agentic claim verification over interactive visualizations comprising 21,349 claims grounded in human-authored analytical notebooks. By requiring models to actively navigate web-based environments to gather evidence, InSight moves beyond static visual question answering and captures the sequential, exploratory nature of real analytical workflows.

Our evaluation reveals several findings. First, interactive claim verification remains challenging: the best-performing model, GPT-5.5, achieves only 57.2% accuracy. Second, IES exposes models that answer correctly without engaging with the environment at all, which indicates ungrounded prediction rather than evidence-based reasoning. Third, falsification is disproportionately difficult, requiring targeted evidence-seeking interaction that most models struggle to produce, extending prior observations from textual fact-checking to the agentic multimodal setting.

These findings point to several directions for future work. The gap between static and interactive performance suggests that explicit training for evidence-seeking behavior, rather than one-shot visual reasoning, may be necessary. The difficulty of falsification motivates investigation into how models form and update hypotheses during multi-turn visual exploration. More broadly, InSight provides a foundation for studying multimodal argumentation in agentic systems, where the reasoning process is not opaque but traceable through interaction.

6 Limitations

InSight has several limitations that point to directions for future work. The benchmark assumes a fixed, high-level action space that, while grounded in prior work on agentic benchmarks, abstracts common mouse and navigation interactions. While this enables controlled comparison across models, it may not capture all interaction modalities used by humans, such as complex gesture-based interactions or semantic shortcuts provided by specific interfaces. Different action space designs may yield different interaction strategies and performance profiles.

Although InSight emphasizes interaction as a proxy for reasoning, interaction traces do not fully capture internal model deliberation. A model may perform correct actions for the wrong reasons or fail despite internally plausible reasoning. Thus, interaction traces should be viewed as complementary, not exhaustive, evidence of reasoning behavior. However, we conjecture that these traces still provide evidence for examining model behavior.

7 Ethical Considerations

The construction of this corpus was subject to strict institutional governance and was formally ratified by the university's Research Ethics Committee (REC). The corpus comprises artefacts originating from the capstone projects of a graduate-level module in Advanced Data Visualisation. Unlike uncurated web-scraped datasets, which often suffer from variable quality and noisy alignment, our data collection protocol imposed a rigorous high-pass quality filter. The inclusion criteria were determined by a two-stage validation process. Each submission was assessed by a senior academic specialising in data visualisation and visual analytics. This evaluation served as our primary signal of quality; only submissions demonstrating substantial technical and design proficiency were selected for inclusion. The initial assessment was subsequently audited and validated by the departmental Academic Board, ensuring that the selection serves as a gold-standard representation of domain concepts. Following ethical approval, we engaged in a retrospective consent acquisition process. We contacted graduates of the programme to articulate the study's objectives, specifically the contribution of their work to the advancement of multimodal machine learning. The annotation procedure was also subject to ratification by the REC. We recruited data science master's students via email to participate in the study. Participation was voluntary and participants were not paid.

Acknowledgments

This work was supported in part by the Alan Turing Institute under Fundamental Research Project No. PP00029.

References

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. ChartCheck: Explainable Fact-Checking over Real-World Chart Images. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13921–13937, Bangkok, Thailand. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. 2018. On evaluation of embodied navigation agents. *Preprint*, arXiv:1807.06757.

Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2025. AVerImaTeC: A Dataset for Automatic Verification of Image-Text Claims with Evidence from the Web.

Thibault Le Sellier De Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Quentin Cappart, Graham Neubig, Ruslan Salakhutdinov, Nicolas Chapados, and Alexandre Lacoste. 2025. The BrowserGym Ecosystem for Web Agent Research. *arXiv preprint*. ArXiv:2412.05467 [cs].

A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28. Publisher: [Royal Statistical Society, Oxford University Press].

Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. 2024. Unveiling the Spectrum of Data Contamination in Language Model: A Survey from Detection to Remediation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16078–16092,

Bangkok, Thailand. Association for Computational Linguistics.

Hiroki Furuta, Kuang-Huei Lee, Ofir Nachum, Yutaka Matsuo, Aleksandra Faust, Shixiang Shane Gu, and Izzeddin Gur. 2024. Multimodal Web Navigation with Instruction-Finetuned Foundation Models. *arXiv preprint*. ArXiv:2305.11854 [cs].

Google. 2025a. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. Technical Report, Google.

Google. 2025b. A new era of intelligence with Gemini 3.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv preprint*. ArXiv:2111.09543 [cs].

Jiayi Hong, Christian Seto, Arlen Fan, and Ross Maciejewski. 2025. Do LLMs Have Visualization Literacy? An Evaluation on Modified Visualizations to Test Generalization in Data Interpretation. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):7004–7018.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding Data Visualizations via Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656. ISSN: 2575-7075.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An Annotated Figure Dataset for Visual Reasoning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

Rachneet Kaur, Nishan Srishankar, Zhen Zeng, Sumitra Ganesh, and Manuela Veloso. 2025. ChartAgent: A Multimodal Agent for Visually Grounded Reasoning in Complex Chart Question Answering. *arXiv preprint*. ArXiv:2510.04514 [cs].

Nam Wook Kim, Yongsu Ahn, Grace Myers, and Benjamin Bach. 2025. How Good Is ChatGPT in Giving Advice on Your Visualization Design? *ACM Trans. Comput.-Hum. Interact.*, 32(5):46:1–46:33.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. 2024. VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 881–905, Bangkok, Thailand. Association for Computational Linguistics.

Alan Lundgard and Arvind Satyanarayan. 2022. Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1073–1083.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tiange Luo, Lajanugen Logeswaran, Justin Johnson, and Honglak Lee. 2025. Visual Test-time Scaling for GUI Agent Grounding. *arXiv preprint*. ArXiv:2505.00684 [cs].

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. PlotQA: Reasoning over Scientific Plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1516–1525. ISSN: 2642-9381.

Dasha Metropolitansky and Jonathan Larson. 2025. Towards Effective Extraction and Evaluation of Factual Claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6996–7045, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2025. Introducing GPT-5.2.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585. Place: Cambridge, MA Publisher: MIT Press.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, and 1 others. 2025. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*.

Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2017. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, Toronto, Canada. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. *arXiv preprint*. ArXiv:1803.05355 [cs].

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint*. ArXiv:2203.11171 [cs].

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large Language Models are Better Reasoners with Self-Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for Computational Linguistics.

Jo Wood, Alexander Kachkaev, and Jason Dykes. 2019. Design Exposition with Literate Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):759–768.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Preprint*, arXiv:2404.07972.

A Dataset Construction Details

This appendix provides full implementation details for the four-stage dataset construction pipeline summarized in §3.2.

A.1 Data Sources and Filtering

InSight is derived from human-authored analytical notebooks that combine interactive Vega-Lite (Satyanarayan et al., 2017) visualizations with rich natural language narratives (Wood et al., 2019). These notebooks were authored independently by analysts with formal training in data visualization. They selected their own datasets, formulated analytical questions, and designed custom visualizations to explore them. The resulting notebooks reflect authentic analytical reasoning about data rather than constructed descriptions.

We apply a series of filtering criteria to ensure quality and relevance, retaining $N = 297$ notebooks. We extract two components from each: the human-authored analytical narrative, which serves as the source of grounded claims, and the corresponding interactive visualization environment.

A.2 Span Extraction and Semantic Labeling

The analytical narratives contain a mixture of descriptive context, commentary, and claims grounded in visual evidence. We isolate spans that express verifiable claims about the underlying data using an agreement-based extraction procedure.

Candidate Span Extraction. We adopt Lundgard and Satyanarayan (2022)'s four-level semantic model of natural language about data visualizations, specifically levels 2 and 3. Level 2 captures statistical insights about the underlying datasets, such as extrema or value retrieval, while level 3 captures visual insights revealed by the visualizations, such as trends or patterns. This ensures that candidate spans relate to content that could be verified or refuted by visual evidence.

We prompt Gemini 2.5 Flash (Google, 2025a) to extract text spans from each analytical narrative containing level 2/3 semantic content. To reduce the potential impact of variability in model outputs, we perform three independent runs per text segment (Wang et al., 2023).

Agreement-Based Span Consolidation. We consolidate spans across the three model runs using a multi-stage agreement procedure. Any spans that are not extractive (i.e., not verbatim substrings of the original text) are discarded. This ensures that no new content is introduced by the LLM at this stage.

We group spans across runs based on lexical overlap by computing the word trigram F1 overlap score and greedily matching spans from different model runs with high lexical similarity, using a fixed minimum threshold for matching. Spans that are not grouped with any other span, meaning only one model run extracted that span, are discarded.

From the groups of matched spans, we construct a final span by retaining tokens chosen by at least two model runs. If this produces a non-contiguous sequence, we expand it to the minimal contiguous substring in the original text. The semantic label assigned to this span is the majority label across the runs. This procedure enforces cross-run stability, filtering out hallucinated extractions and semantically irrelevant text.

A.3 Span Decomposition into Verifiable Claims

The spans extracted in the previous stage may contain multiple propositions, implicit references, or compound statements that are not directly suitable for claim verification. We decompose each retained span into a set of atomic, verifiable claims that can be independently assessed against the visualizations. This stage is adapted from Claimify (Metropolitansky and Larson, 2025), tailored to interactive visual evidence.

We prompt an LLM to identify all specific and verifiable propositions expressed in each span. The model is instructed to decompose compound statements into their simplest discrete units and to discard non-verifiable content. Each claim must be fully decontextualized – interpretable in isolation, without access to the surrounding narrative – while preserving its original meaning. The model is provided with the span's immediate textual context and required to explicitly resolve all referents (e.g., pronouns, comparatives, implicit subjects). Claims are restricted to information present in the original span or its local context; no external knowledge is permitted.

To mitigate potential meaning shifts introduced by LLM-based decomposition, we apply a self-verification step in which the model is shown the original span and the decomposed claims and asked to verify that each claim is entailed by the source text (Weng et al., 2023). We also reassign semantic labels to each claim using the same semantic

schema, performing 3 passes and retaining the majority label. Claims for which semantic labeling is unstable or that fail grounding verification are discarded.

The claims produced by this stage constitute the set of ground-truth TRUE claims in InSight. Each is directly derived from human-authored analytical text, verified to preserve its original meaning after decomposition and decontextualization. These true claims form the semantic anchor of the dataset and are subsequently used as the source for controlled generation of FALSE and NEI variants.

A.4 Claim Mutation

We generate FALSE and NEI claims through controlled mutation of the true claims, applying minimal, semantics-preserving perturbations designed to alter veracity while retaining linguistic plausibility and structural similarity. This reduces annotation artifacts by ensuring that true, false, and NEI claims differ primarily in their semantic relation to the underlying visual evidence.

Antonym Substitution. This strategy targets claims whose veracity depends on scalar, directional, or quantitative language. We construct a lexicon of lemmas that have a clear antonym which, when substituted, induces a semantically opposed claim. The lexicon is curated from high-frequency lemmas in the true claim set, focusing on auxiliary verbs, determiners, adjectives, adverbs, and nouns expressing quantitative judgments (e.g., increase/decrease, all/none). Mappings are curated to avoid substitutions that merely weaken a claim or introduce ambiguity, ensuring clear semantic opposition.

We tag occurrences of lexicon entries in each claim using spaCy (Honnibal et al., 2020) and custom XML-style markers that encode the lemma identity. This constrains the mutation process, preventing lexical drift. We then prompt an LLM to generate a mutated false claim by substituting only the tagged tokens with an appropriate antonym, preserving all other content verbatim and making only minimal grammatical adjustments when strictly necessary.

Contradiction Validation. Each original–mutated claim pair is evaluated using a pretrained NLI model (He et al., 2023). We compute contradiction scores in both directions and retain only claims whose average contradiction score exceeds a threshold of 0.9. This filters out cases where antonym substitution fails to invert meaning cleanly or produces ambiguous statements.

In-Lexicon Argument Substitution. Many claims relate to specific data arguments—attributes, categorical values, or numerical quantities. We identify candidate arguments by generating a structured summary of each visualization's underlying dataset, including attribute names, categorical values, and numerical ranges. An LLM tags spans in each claim corresponding to attributes, values, and numbers using XML-style markers encoding the associated attribute name. This tagging procedure is shared across both in-lexicon and out-of-lexicon substitution strategies.

In-lexicon substitution replaces one or more tagged arguments with an alternative from the same dataset lexicon. We apply two forms: categorical value substitution, in which a categorical value is replaced with a different value from the same attribute; and numerical substitution, in which a numerical value is replaced with a different value within the observed range of the attribute. We do not substitute attribute names, as doing so can unintentionally produce true claims when attributes share similar distributions. Only mutated claims exhibiting high NLI contradiction scores (> 0.9) are retained and labeled FALSE.

Out-of-Lexicon Argument Substitution. This strategy generates NEI claims by replacing tagged arguments with values or attributes absent from the underlying dataset, introducing references that cannot be verified or refuted. We apply two types: attribute substitution, replacing an attribute name with one not observed in the dataset; and categorical value substitution, replacing a value with one not observed. We do not substitute numerical values, as this can only produce false claims.

NLI-based validation ensures that the resulting claim is neither supported nor contradicted by the original source text. We compute entailment and contradiction scores and retain only claims where both fall below a threshold of 0.2, removing cases where the substitution introduces an unintended true or false statement.

A.5 Human Validation Details

A total of 13 annotators with expertise in data visualization participated in the validation study, producing 475 annotations across a sample of 294 claims, with up to three independent annotations per claim. Annotators performed the same task

defined in §3.1: interacting with the visualizations as needed and assigning one of the three labels based solely on the available visual evidence. They were primed with example tasks and given feedback before starting annotation on real samples. Annotators were blind to the original dataset labels and to the construction process.

Raw agreement between human annotators and the original labels is 81.3%. We additionally applied a Bayesian Dawid–Skene model (Dawid and Skene, 1979; Paun et al., 2018) to the annotation data, inferring latent labels solely from annotator responses. The inferred class prevalence closely matches the dataset distribution: estimated proportions for TRUE and FALSE claims differ by less than 3 percentage points from the ground truth, while NEI prevalence is recovered within the model's uncertainty bounds. The posterior mean accuracy of inferred labels relative to the dataset labels is 66.2% (95% credible interval: [59.5%, 71.8%]), with an MAP estimate of 77.6%.

B Dataset Construction Method Example

This appendix illustrates the complete pipeline for constructing claims from human-authored analytical narratives, using a real example from the corpus.

B.1 Original Text

"...Firstly, the number of fatal traffic accidents per month has been decreased from 2005 to 2015. From the last bar chart, In 2005 to 2007, there were roughly about 350 records monthly. The number drops gradually, and starting from 2010, there were about 200 records in each month. This decrease in records may be a result of advancements in protective features of cars, legislation against mobile phone use while driving or other traffic safety related legislations..."

B.2 Span Extraction and Semantic Labeling

Candidate spans are extracted from the narrative and labeled according to their semantic level (Level 2: statistical/relational, Level 3: perceptual/cognitive):

Label 3: "Firstly, the number of fatal traffic accidents per month has been decreased from 2005 to 2015."

Label 2: "In 2005 to 2007, there were roughly about 350 records monthly."

Label 3: "The number drops gradually."

Label 2: "and starting from 2010, there were about 200 records in each month."

B.3 Decomposition and Decontextualization

Each span is decomposed into atomic, verifiable propositions with all implicit references resolved:

Span: "Firstly, the number of fatal traffic accidents per month has been decreased from 2005 to 2015."
`Proposition: "The number of fatal traffic accidents per month has been decreased from 2005 to 2015."`

Span: "In 2005 to 2007, there were roughly about 350 records monthly."
`Proposition: "In 2005 to 2007, there were roughly about 350 fatal traffic accidents monthly [as shown in the last bar chart]."`

Span: "The number drops gradually."
`Proposition: "The number of [fatal traffic accidents per month] drops gradually [from 2005 to 2015]."`

Span: "and starting from 2010, there were about 200 records in each month."
`Proposition: "Starting from 2010, there were about 200 [fatal traffic accident] records in each month."`

B.4 Claim Mutation: Contrary Swap

Antonym substitution generates false claims by reversing directional language. NLI contradiction scores validate semantic opposition:

Input: The number of fatal traffic accidents per month has been **decreased** from 2005 to 2015.
Mutated (False): The number of fatal traffic accidents per month has been **increased** from 2005 to 2015. *(contradiction score: 0.998)*

Input: The number of fatal traffic accidents per month **drops** gradually from 2005 to 2015.
Mutated (False): The number of fatal traffic accidents per month **rises** gradually from 2005 to 2015. *(contradiction score: 0.999)*

B.5 Claim Mutation: In-Lexicon Argument Swap

Arguments (dates, values) are replaced with alternatives from the dataset to generate false claims:

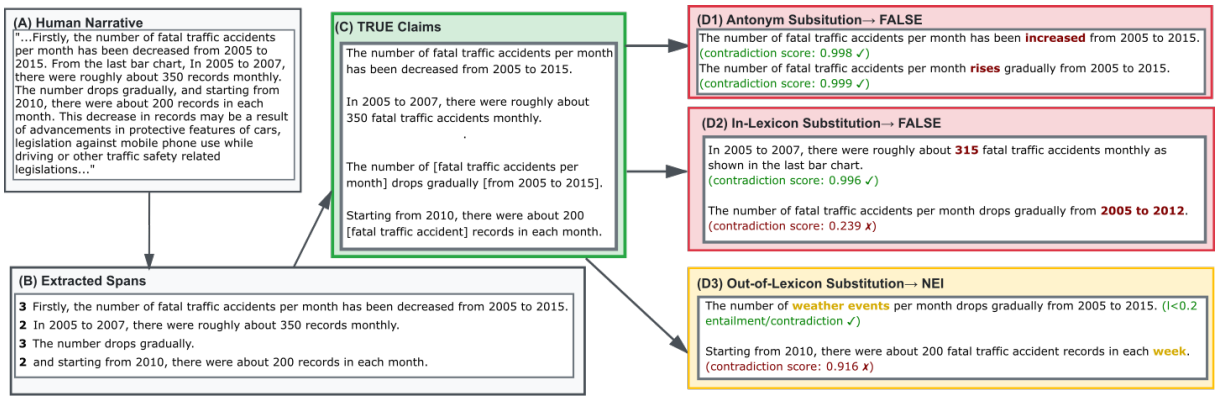

Figure 2: Worked example of the dataset construction pipeline. Starting from human-authored analytical text about traffic accidents (A), we extract semantically labeled spans (B), decompose them into atomic propositions with resolved references (C), and generate False claims through antonym substitution (D1) and in-lexicon argument replacement (D2), as well as NEI claims through out-of-lexicon substitution (D3). NLI scores validate that mutations produce the intended semantic relationships.

Mutated (False): In **2008 to 2007**, there were roughly about 350 fatal traffic accidents monthly as shown in the last bar chart. *(Contradiction score: 0.984)*

Mutated (False): In 2005 to 2007, there were roughly about **315** fatal traffic accidents monthly as shown in the last bar chart. *(Contradiction score: 0.996)*

Mutated (False): Starting from 2010, there were about **50** fatal traffic accident records in each month. *(Contradiction score: 0.068 - rejected)*

B.6 Claim Mutation: Out-of-Lexicon Argument Swap

Arguments not present in the dataset are substituted to generate NEI (Not Enough Information) claims:

Mutated (NEI): The number of **road incidents** per month drops gradually from 2005 to 2015. *(Entailment score: 0.885 - rejected)*

Mutated (NEI): Starting from 2010, there were about 200 fatal traffic accident records in each **week**. *(Entailment score: 0.010, Contradiction score: 0.916 - rejected)*

Note: Bold text indicates substituted elements. NLI scores validate that mutations produce the intended semantic relationship (high contradiction for false claims, low entailment/contradiction for NEI claims).

C Dataset Statistics

This appendix provides additional distributional detail complementing the dataset analysis in §3.2.

Mark types. The mark-type distribution (Figure 3, top right) illustrates that InSight is not dominated by a single visualization form. While bars are the most common mark, a substantial fraction of notebooks include lines, geoshapes, circles, text marks, and layered combinations thereof. This diversity is a natural consequence of using Vega-Lite specifications authored by analysts rather than enforcing a fixed chart taxonomy. Importantly, many claims require reasoning across multiple mark types within the same environment (e.g., relating a geographic map to a linked bar chart).

Specification and interaction density. The histograms of visualization specifications per file and interaction types per file (Figure 3, bottom row) show that most notebooks contain multiple specifications and multiple distinct interaction mechanisms. The long right tails indicate the presence of highly complex environments with many views and interaction affordances. This reinforces that InSight environments are not single-chart snapshots but multi-view analytical spaces in which part of the verification task is identifying where relevant evidence resides.

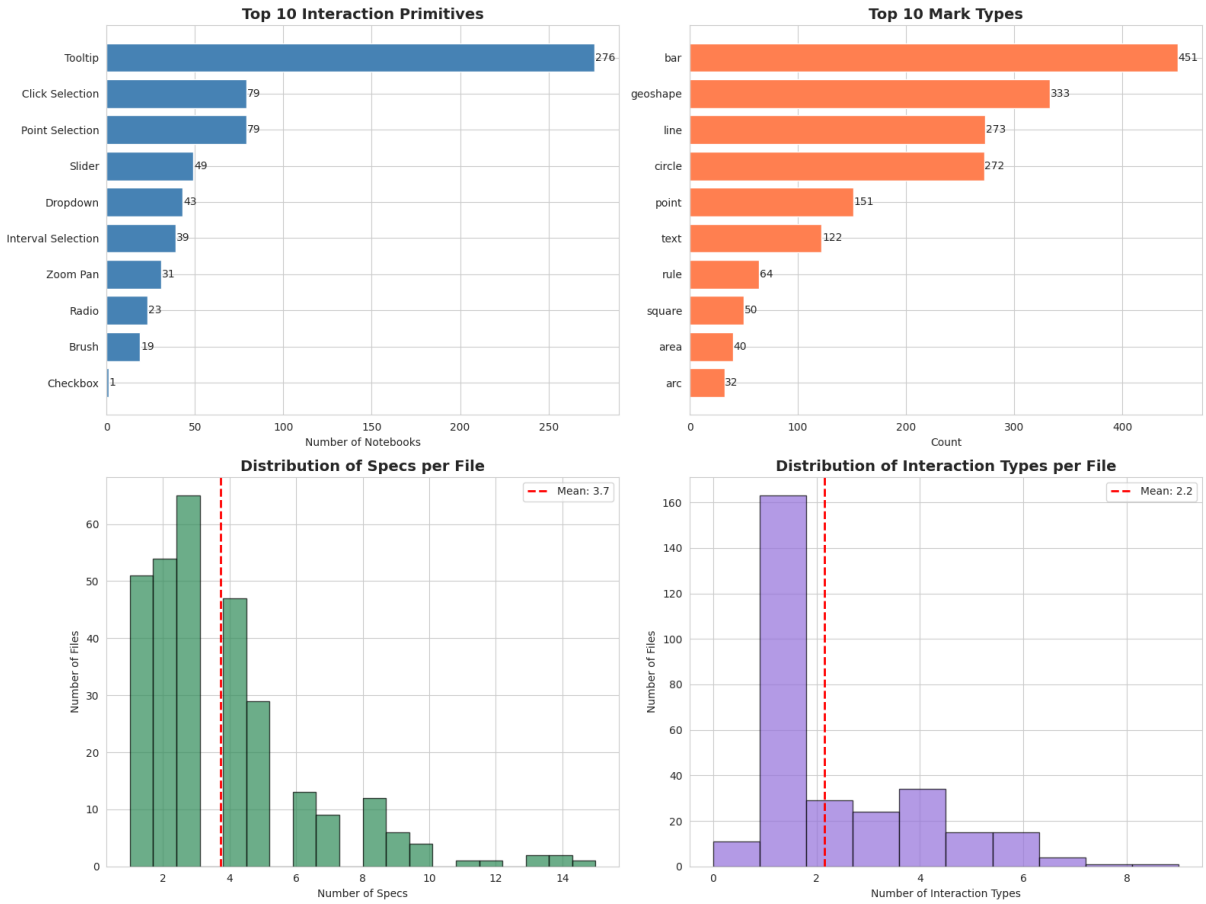

Figure 3: Dataset composition. Top left: distribution of interaction primitives across notebooks. Top right: distribution of Vega-Lite mark types. Bottom left: number of visualization specifications per notebook. Bottom right: number of distinct interaction types per notebook.

D Model Evaluation Examples

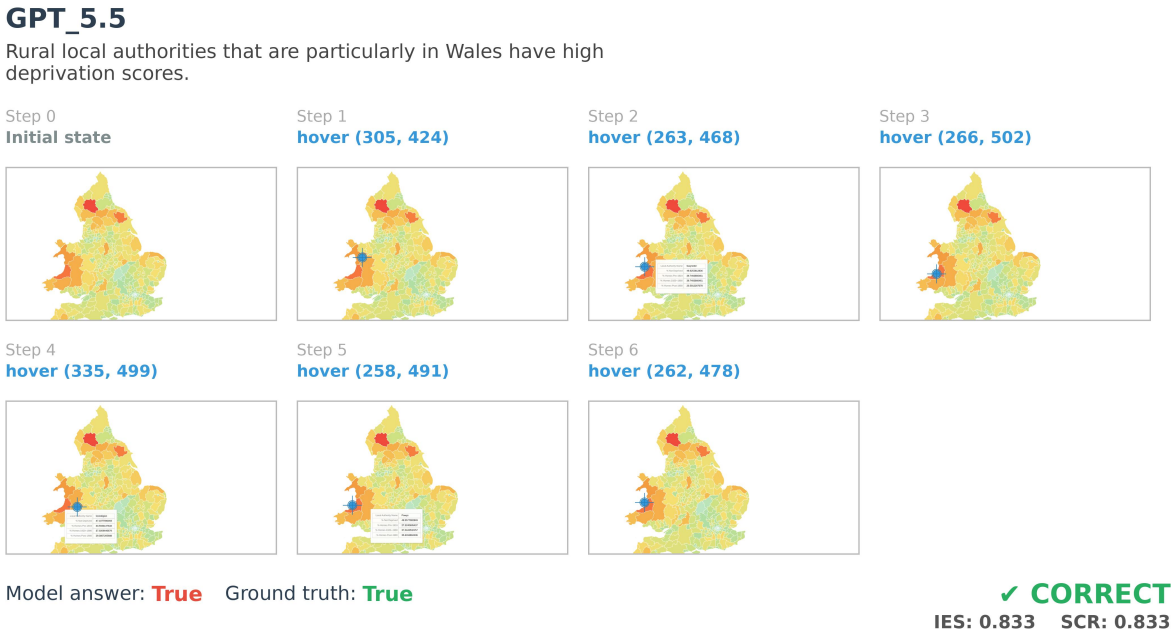

Figure 4: Interaction trace for GPT 5.5 correctly verifying a true claim

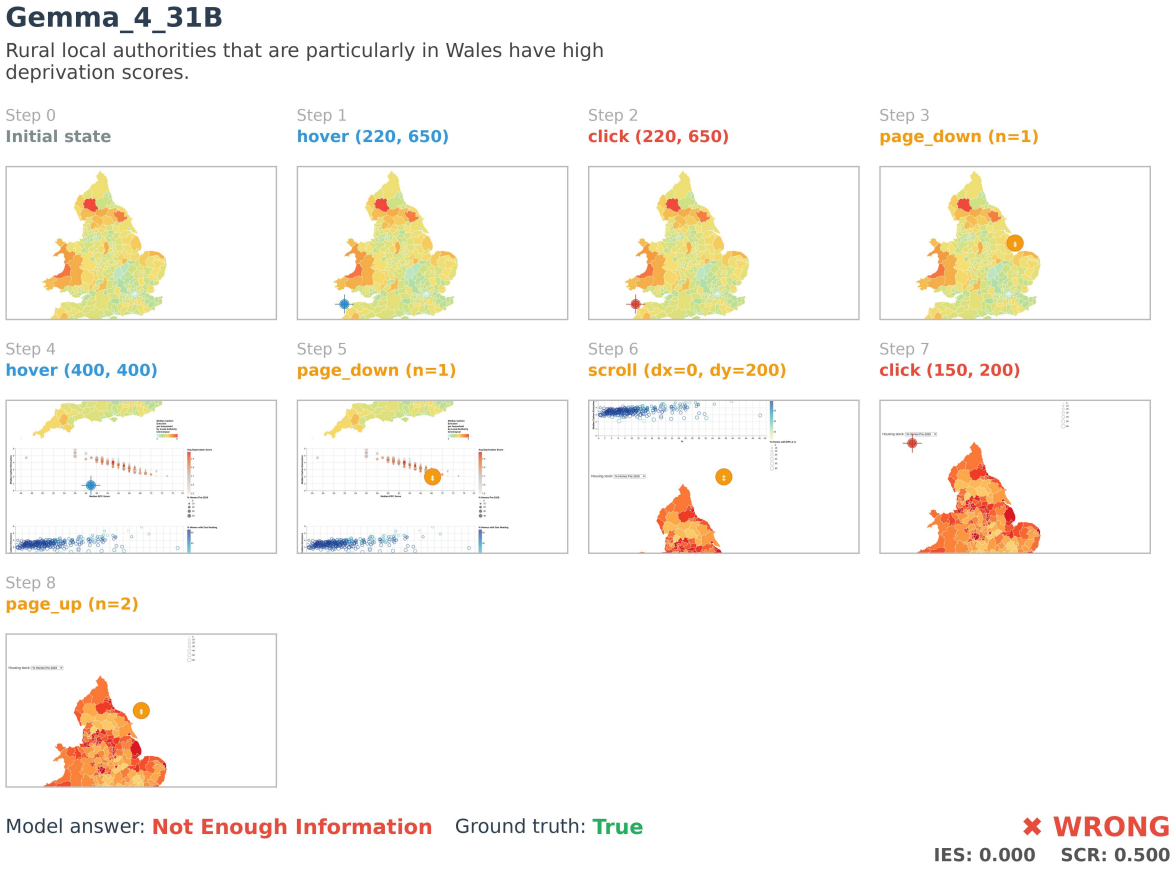

Figure 5: Interaction trace for Gemma 4 31B Flash verifying a true claim, incorrectly labelling the claim as NEI.

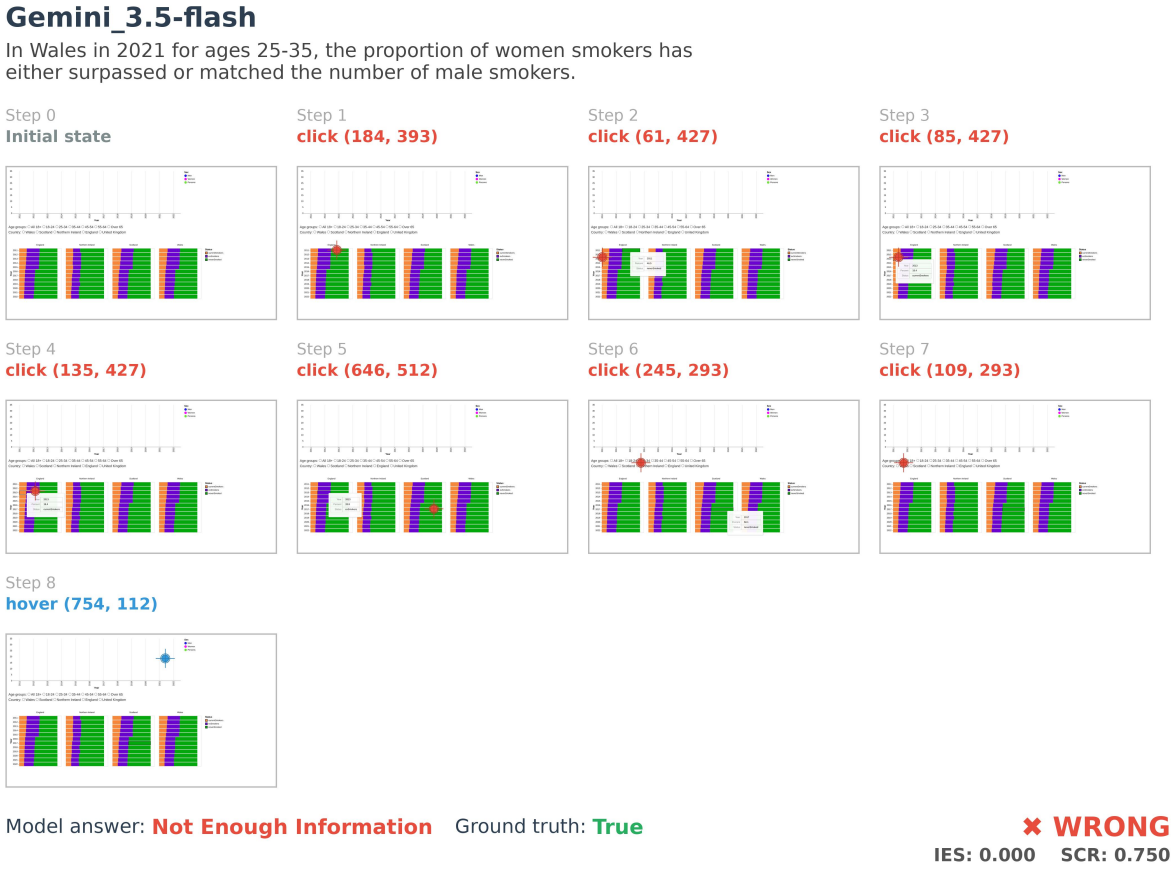

Figure 6: Interaction trace for Gemini 3.5 Flash verifying a true claim, incorrectly labelling the claim as NEI.

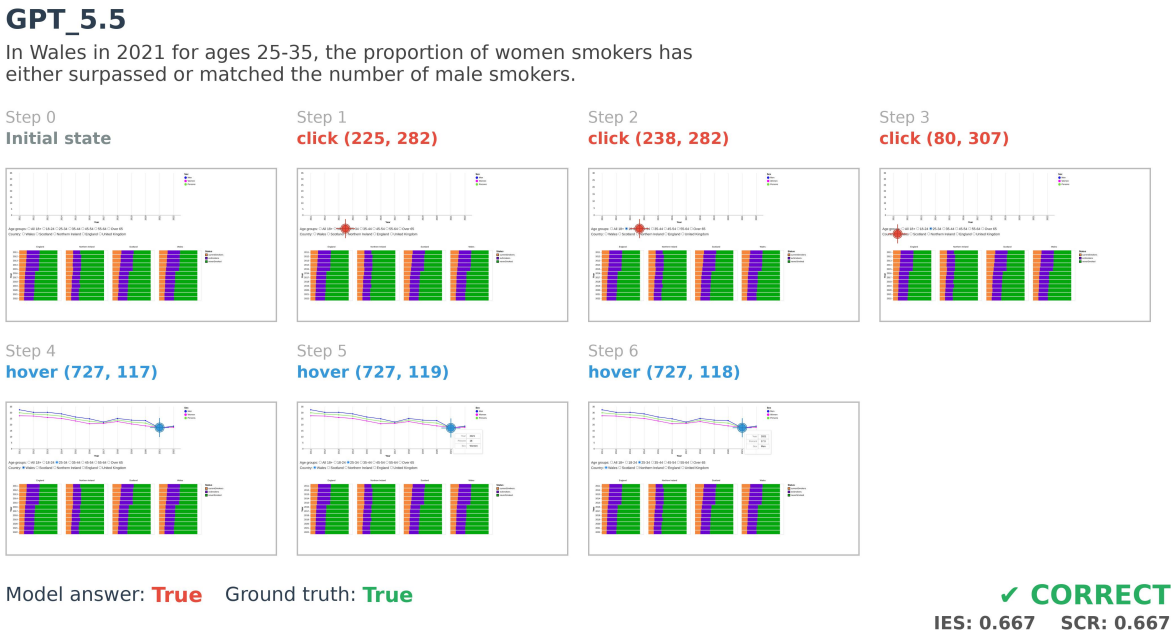

Figure 7: Interaction trace for GPT 5.5 correctly verifying a true claim.

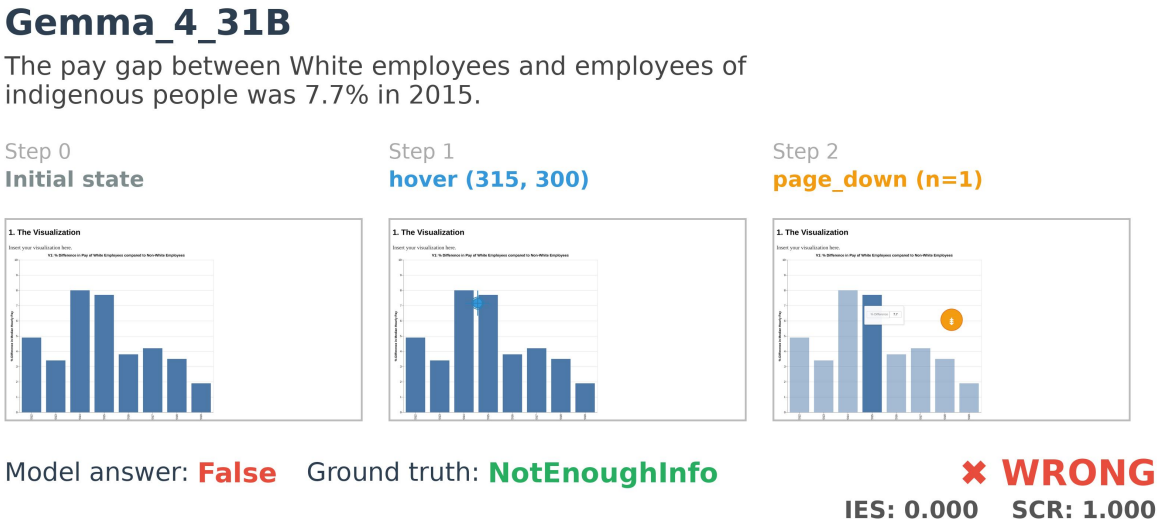

Figure 8: Interaction trace for Gemma 4 31B verifying an NEI claim, incorrectly labelling the claim as False.

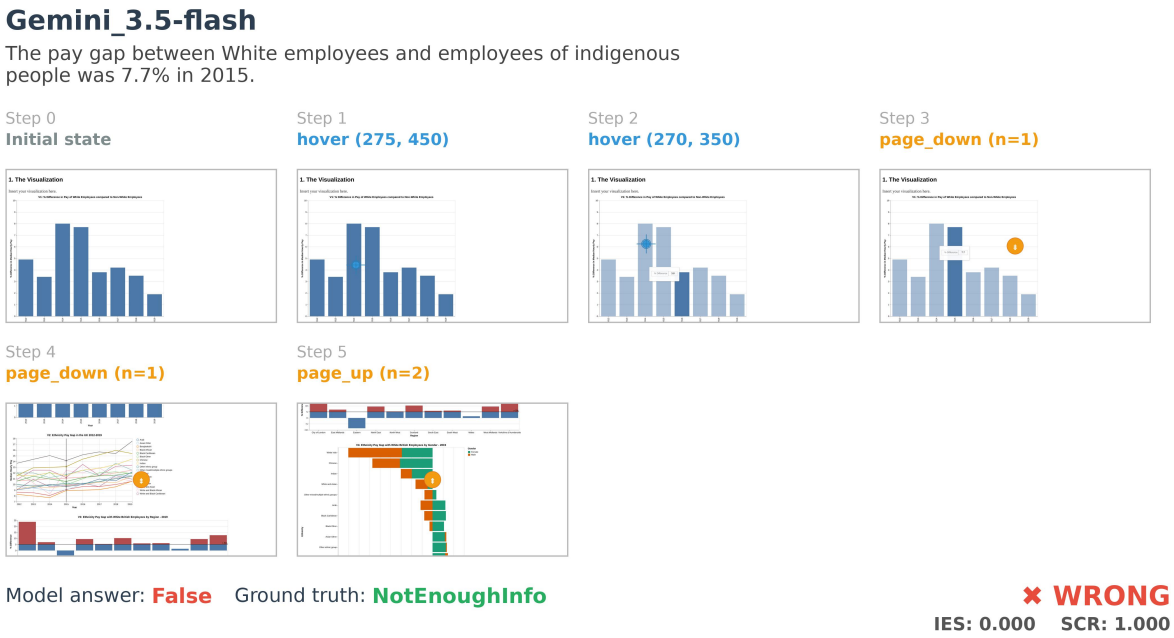

Figure 9: Interaction trace for Gemini 3.5 Flash verifying an NEI claim, incorrectly labelling the claim as False.