DASHBOARDQA: Benchmarking Multimodal Agents for Question Answering on Interactive Dashboards

Aaryaman Kartha[♣] [*], **Ahmed Masry**[♣] [*], **Mohammed Saidul Islam**[♣], **Thinh Lang**[♣], **Shadikur Rahman**[♣], **Ridwan Mahbub**[♣], **Mizanur Rahman**[♣★], **Mahir Ahmed**[♣], **Md Rizwan Parvez**[§], **Enamul Hoque**[♣], **Shafiq Joty**[◇△]

[♣]York University, Canada, [★]RBC, Canada, [§]Qatar Computing Research Institute (QCRI)
[◇]Nanyang Technological University, Singapore, [△]Salesforce Research, USA
{aarykary, masry20, saidulis, tlang46, shadikur, mrahmed, rmahbub, enamulh}@yorku.ca
{mparvez@hbku.edu.qa, sjoty@salesforce.com}

Abstract

Dashboards are powerful visualization tools for data-driven decision-making, integrating multiple interactive views that allow users to explore, filter, and navigate data. Unlike static charts, dashboards support rich interactivity, which is essential for uncovering insights in real-world analytical workflows. However, existing question-answering benchmarks for data visualizations largely overlook this interactivity, focusing instead on static charts. This limitation severely constrains their ability to evaluate the capabilities of modern multimodal agents designed for GUI-based reasoning. To address this gap, we introduce DASHBOARDQA, the first benchmark explicitly designed to assess how vision-language GUI agents comprehend and interact with real-world dashboards. The benchmark includes 292 tasks on 112 interactive dashboards, encompassing 405 question answer pairs overall. These questions span five categories: multiple-choice, factoid, hypothetical, multi-dashboard, and conversational. By assessing a variety of leading closed- and open-source GUI agents, our analysis reveals their key limitations, particularly in grounding dashboard elements, planning interaction trajectories, and performing reasoning. Our findings indicate that interactive dashboard reasoning is a challenging task overall for all the VLMs evaluated. Even the top-performing agents struggle; for instance, the best agent based on Gemini-Pro-2.5 achieves only 38.69% accuracy, while the OpenAI CUA agent reaches just 22.69%, demonstrating the benchmark's significant difficulty. We release DASHBOARDQA at https://github.com/vis-nlp/DashboardQA.

1 Introduction

Using visualizations to explore data and answer questions is fundamental to many decision-making tasks (Kim et al., 2020; Hoque et al., 2022). While static charts can present snapshots of specific insights, they fall short in capturing the dynamic, iterative nature of real-world data exploration. In practice, analysts continuously refine queries, switch views, and filter subsets of data as their investigative goals evolve (Hoque et al., 2018). Interactive dashboards support these workflows by allowing users to adjust visual and query parameters via UI controls (e.g., filters, sliders, dropdowns, tabs), while linking coordinated views so that interactions in one view propagate to others—facilitating navigation through complex datasets and extraction of deeper insights. Effective question answering over dashboards also broadens access to data-driven insights(Chowdhury et al., 2020), supporting applications such as accessibility for users with visual impairments (Alam et al., 2023).

However, automatic question answering over such dashboards presents unique challenges. Answering even a single question may require a sequence of GUI operations such as setting filters, switching tabs, zooming into a chart, then computing a ratio or identifying an outlier across views, while maintaining and interpreting intermediate states (see an example in Figure 1). This multi-step process is cognitively demanding for humans, incurring high mental load as they make efforts to manage linked visualizations and interactive constraints (Convertino et al., 2003). These challenges raise a fundamental research question: *Can we build AI agents that answer questions by interacting with dashboards as humans do—through a sequence of informed, goal-directed GUI operations over dynamic visual states?*

While interest in benchmarking reasoning over data visualizations has grown over the past few years, existing work has consistently overlooked the critical element of interactivity. Early benchmarks focused on synthetic static charts (FigureQA (Kahou et al., 2017), PlotQA (Methani et al., 2020)), and while later work introduced real-

[*] Equal contribution.

Findings of the Association for Computational Linguistics: EACL 2026, pages 3385–3407
March 24-29, 2026

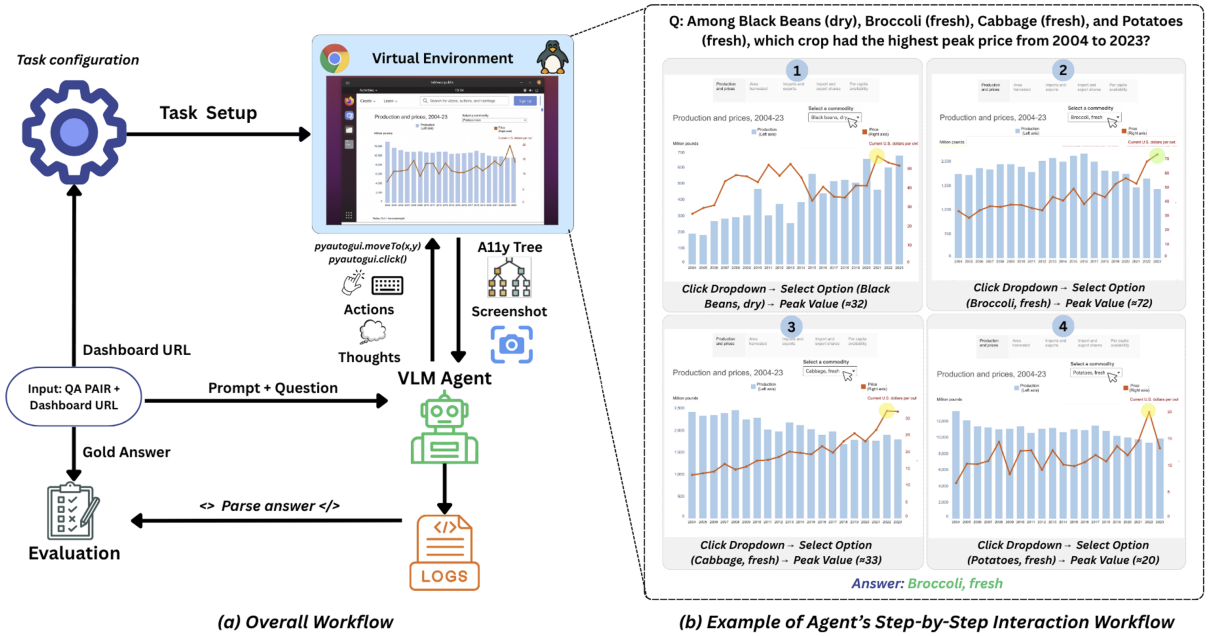


Figure 1: An overview of the DASHBOARDQA task execution pipeline within the OSWorld environment, where (a) illustrates the overall setup of the VLM agent, and (b) shows an interaction trajectory to answer a question. The VLM, acting within an Ubuntu virtual environment, takes visual (Screenshot) or structural (A11y tree) input and generates `pyautogui` commands (e.g., mouse clicks, movements, scrolling) to guide navigation at every step.

world data and more complex reasoning (ChartQA (Masry et al., 2022), ChartQAPro (Masry et al., 2025a)), the visualizations remained static. In parallel, vision-language models (VLMs) (OpenAI et al., 2024; Comanici et al., 2025), particularly those adapted to chart-centric tasks through fine-tuning or fine-tuning followed by reinforcement learning (Masry et al., 2023, 2024a,b, 2025b), have significantly advanced static chart understanding (Islam et al., 2024); however, their application to interactive data analysis remains underexplored.

Concurrently, a wave of general-purpose GUI benchmarks like OSWorld (Xie et al., 2024a), AndroidWorld (Rawles et al., 2025), and OfficeBench (Wang et al., 2024a) has emerged, revealing that even state-of-the-art agents struggle with fundamental skills like grounding UI elements and planning long interaction sequences.

This has spurred the development of a new class of agentic approaches, from closed-source systems like OpenAI's Computer-Use Agent (ope) to open-source agents like UI-TARS (Qin et al., 2025) and hybrid strategies pairing powerful planners (e.g., Gemini Pro 2.5 (Comanici et al., 2025) or GPT-4o (OpenAI et al., 2024)) with specialized visual grounding modules like JEDI (Xie et al., 2025) or GTA1 (Yang et al., 2025). Yet, existing GUI benchmarks to evaluate these agents remain focused on utility-oriented tasks (e.g., clearing cookies, editing files), offering limited insights into agents' ability to perform analytical reasoning over visual data.

To bridge this gap, we introduce DASHBOARDQA, the first benchmark for agentic question answering over interactive dashboards. It comprises 112 diverse dashboards from Tableau Public and 292 expert-authored tasks including factoid, multiple-choice, hypothetical, conversational, and unanswerable questions to reflect realistic, complex analytical workflows. All together, 405 QA pairs are created, where all question categories are single turn expect for conversational, in which each task can comprise of 3-7 QA pairs. Answering these questions often requires multi-step interaction sequences such as filtering, switching views, interpreting updated charts, and analyzing various data visualization types present. We evaluate a range of open- and closed-source agents on this benchmark and find that even top-performing models struggle significantly; for example, the best-performing agent based on Gemini-Pro-2.5 achieves only 38.69% accuracy, underscoring the benchmark's difficulty and novelty.

In summary, our contributions are: (*i*) **A comprehensive benchmark** that requires complex, multi-step GUI interactions and more advanced visual reasoning with coordinated multi-view dashboards to answer questions. (*ii*) **Extensive evaluation** of open- and closed-source agents, revealing signifi-

cant performance drops on interactive dashboard QA tasks compared to static charts and utility-based benchmarks; (*iii*) **In-depth analysis and ablations** that reveal key limitations in grounding, planning, and visual reasoning, providing concrete directions for improving future multimodal agents.

2 Related Work

Data Visualization QA Benchmarks. Benchmarks designed to evaluate visual-language models (VLMs) through question answering on data visualizations have significantly evolved in complexity over recent years. Early datasets, such as FigureQA (Kahou et al., 2018), DVQA (Kafle et al., 2018), and PlotQA (Methani et al., 2020), primarily relied on synthetic data and simple charts. Later, benchmarks including ChartQA (Masry et al., 2022) InfoVQA (Mathew et al., 2021), Open-ended ChartQA (Kantharaj et al., 2022) and Text2Vis (Rahman et al., 2025) introduced real-world visualizations paired with human-authored questions, but still focused on static charts. However, rapid advancements in vision-language modeling quickly saturated performance on these datasets. To provide a greater challenge, recent benchmarks like CharXiv (Wang et al., 2024b) and ChartQAPro (Masry et al., 2025a) were developed, emphasizing more complex and diverse question types as well as richer visual content, including infographics and static dashboards.

Despite these advancements, existing benchmarks are fundamentally limited by their reliance on static images. Even recent work like MultiChartQA (Zhu et al., 2025), which incorporates multiple charts, still operates in a non-interactive setting. This stands in contrast to real-world analysis, which is often performed on interactive settings featuring coordinated views, filters, and other dynamic controls. DASHBOARDQA aims to bridge this critical gap by evaluating question answering on interactive dashboards, encompassing diverse question types and dynamic user interactions.

GUI Agents and Associated Benchmarks. Recent benchmarks such as OSWorld (Xie et al., 2024b), AndroidWorld (Rawles et al., 2025), WebArena (Zhou et al., 2024), and OfficeBench (Wang et al., 2024a) evaluate agent capabilities in grounding (locating and manipulating UI elements) and planning (sequencing actions) across desktop, mobile, web, and office environments. Specialized grounding benchmarks such as ScreenSpot and ScreenSpot Pro further isolate the grounding challenge (Cheng et al., 2024; Li et al., 2025).

Existing GUI agent approaches typically either use separate specialized models for grounding and planning or unified models for both tasks. Closed-source models are generally strong in planning but may struggle with grounding, leading to the development of open-source models explicitly trained for grounding, such as JEDI, AgentS2, and GTA1 (Xie et al., 2025; Agashe et al., 2025; Yang et al., 2025). Conversely, unified models include fully open-source solutions like UI-TARS (Qin et al., 2025) and proprietary models such as CUA, Gemini, GPT-4o, and Claude (ope; Georgiev et al., 2024; OpenAI et al., 2024; Anthropic, 2024)

Existing benchmarks do not evaluate the end-to-end analytical reasoning required for interactive dashboards—skills like grounding linked views, maintaining state across interactions, and performing visual mathematical calculations. DASHBOARDQA is designed specifically to test GUI agents and VLMs on these critical capabilities.

3 THE DASHBOARDQA BENCHMARK

As illustrated in Figure 1, an agent is given a dashboard interface and a natural language question about it. To arrive at the correct answer, the agent must reason and plan a sequence of actions–navigate the interface by making selections and applying filters, extract relevant information from the views, synthesizing information across views and perform any necessary calculations.

3.1 Dataset Construction

The dataset construction process comprises three key stages (Figure 2): (*i*) Interactive Dashboard curation, (*ii*) Question-Answer pair collection, and (*iii*) Question-Answer Review.

Phase 1 – Dashboard Curation: We selected Tableau Public (Tableau Software, 2025) as our primary source due to its large and diverse repository and support for rich interactivity features such as coordinated views, cross-filtering, and cross-highlighting—critical for evaluating agents on dynamic visual updates. We curated dashboards with multiple linked views and meaningful text–chart integration to mirror real-world analytical workflows. Using keyword and topic searches (e.g., "interactive dashboards," "analytics dashboards," "health," "renewable energy"), we manually screened dashboards for relevance and popularity. We excluded

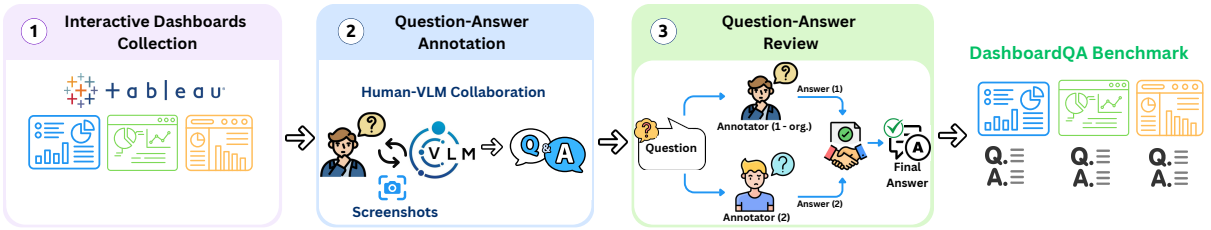


Figure 2: DASHBOARDQA benchmark construction process. We collect interactive dashboards from the Tableau Public website. Annotators collaborate with VLMs to generate question–answer pairs by providing dashboard screenshots. Each QA pair is then reviewed by a second annotator to verify and resolve the final answer.

static or single-view dashboards and infographics, retaining only those with rich interactivity (e.g., dropdowns, filters, tabs), pertinent modern-day topics, varied chart types, and visually distinct layouts. In the end, we selected 112 high-quality dashboards for our benchmark. Example dashboard categorized by topics are displayed in 11, and by source in 10.

Phase 2 - QA pair creation: DASHBOARDQA includes five question categories: (*i*) Factoid, (*ii*) MCQ, (*iii*) Hypothetical, (*iv*) Conversational, and (*v*) Multi-Dashboard (see example QA pair walkthroughs for each category are given in 5–9). A team of eight annotators—comprising university students and professors experienced in data visualization tasks—collaboratively created the QA pairs, with two annotators assigned per category to ensure semantic diversity. To enhance coverage and reduce bias, we employed a three-step human–VLM collaboration process:

- **Curating Seed QA Pairs:** Annotators crafted a diverse set of seed QA pairs covering different question types that required complex reasoning, interactive operations (filters, tabs), and realistic analytical workflows.
- **VLM-Assisted Expansion:** To enrich the dataset and reduce potential bias, we used multiple VLMs (GPT-4o, Gemini, Claude) to expand the initial seed QA pairs. Each model received a seed pair and was instructed to generate five new QA pairs across multiple dashboard views. This interactive prompting strategy encouraged greater diversity and novelty in the generated questions. Prompts used are shown in 13 and 12.
- **Human Refinement:** Finally, annotators reviewed the generated questions to remove incorrect or overly simple ones(e.g., direct data retrieval from charts) and revised those that were unclear, hallucinatory or ambiguous.

During the QA creation process, we prioritized interactivity to distinguish our benchmark from traditional chart QA tasks (Masry et al., 2022). Conventional chart QA datasets typically use static charts, allowing answers to be derived from a single visualization without interaction. In contrast, DASHBOARDQA focuses on dynamic dashboards that require multi-step reasoning through navigation across linked views and the application of interactive controls (e.g., filters, tabs). This design evaluates not only answer correctness but also navigational efficiency, reflecting realistic analytical workflows and imposing higher cognitive demands.

Below we present a brief description of each question type, with corresponding visual examples shown from 5 to 9:

Factoid: These questions focus on information retrieval, pattern recognition, and comparative reasoning, requiring the model to extract key facts, identify patterns, and make informed inferences from contextual cues.

MCQ: MCQs test a VLMs ability to evaluate various candidate options for a question. We specifically focus on scenarios where the VLM Agent is presented with four or five answer options and must synthesize information from multiple views within a dashboard in order to select the correct response. Often, options feature values very close to each other, so it would test the ability to extract data values from visualizations as accurately as possible when used in intermediate calculations for the answer.

Multi-Dashboard: These questions require VLM agents to reason across two or more dashboards, introducing significantly greater cognitive complexity. Beyond understanding a single dashboard, the agent must interpret, compare, and synthesize information across diverse visual and textual contexts. This tests cross-dashboard navigation, multi-view integration, and higher-order reasoning—making such questions inherently more demanding than single-dashboard tasks.

Conversational: These questions consist of multiple interrelated QA pairs based on one or more dashboard views. Each question builds logically on

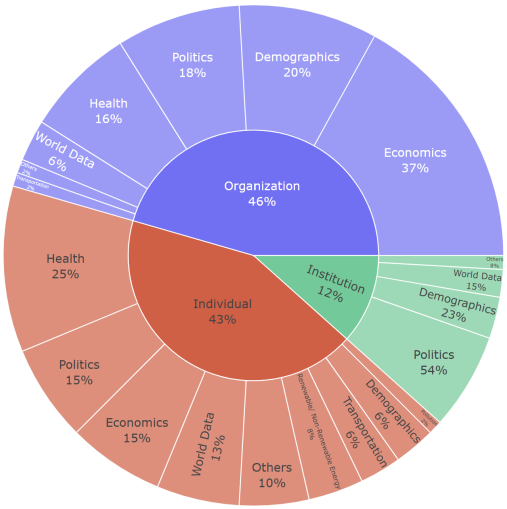


Figure 3: Dashboard topic breakdown by author type. For the 8 dashboard topics present, the nested pie chart shows the entire composition of the topics (Outer loop) to their respective author source type (Inner loop).

the previous ones, forming a coherent dialogue that tests the model's ability to track context and reason over a sequence. This setup evaluates the model's grasp of contextual dependencies, including coreference resolution, multi-step reasoning, and both logical and arithmetic inference.

Phase 3 - Question-Answer Review: To ensure quality and accuracy, each annotator reviewed questions from a different category than they originally worked on, then cross-checked their assessments with the original author. Discrepancies were resolved through discussion, and unresolved cases were discarded. This cross-category review ensured factual correctness and helped eliminate ambiguity or unnecessary complexity. We also examined the interaction states associated with each QA pair to confirm that the optimal navigation trajectory (i.e., number of steps) was used. For estimation questions, we accepted responses within a 0.5% margin of the reference value. The inter-annotator agreement rate was 74.93%, with 25.07% requiring further revision.

3.2 Dataset Analysis

We analyze DASHBOARDQA along two dimensions: dataset composition, covering the diversity of visualizations, topics, and question types, and interaction diversity, representing UI controls and the multi-step complexity of tasks.

3.2.1 Dataset Composition

The dataset dashboards were carefully curated to maximize high visual diversity, including 13 visualization types (Table 8). Fundamental chart types dominate: 48.51% of dashboards include at least one line chart and 42.57% include at least one bar chart. Other common charts in data science and analytics, such as maps, pie charts, scatter plots, and area charts, also appear with moderate frequency. The dataset also exhibits strong domain diversity, covering a wide range of real-world topics and maintaining an almost equal distribution of dashboards created by organizations (Figure 3). Finally, we also prioritized QA diversity, including five question categories: factoid, hypothetical, MCQ, multi-dashboard, and conversational. These categories are designed to test distinct reasoning skills, from direct retrieval to multi-step numerical and logical inference.

Navigation Tool	Prevalence	Count	Avg. per dashboard
Dropdown Menus	92.31%	266	2.38
Radio Buttons	7.69%	26	0.23
Tabs	23.08%	103	0.92
Range Sliders	7.69%	9	0.08

Table 1: Prevalence, count, and average number of respective navigation tools across the dashboards curated.

Question Category	Difficulty Level			
	Easy	Medium	Hard	Total
MCQ	38	32	10	80
Factoid	31	34	9	74
Hypothetical	31	26	8	65
Multi-Dashboard	5	24	17	46
Conversational	7	14	6	27
Total	**112**	**130**	**50**	**292**

Table 2: Distribution of difficulty levels across different question categories in DASHBOARDQA

3.2.2 Interaction Diversity

We emphasize dynamic navigation, requiring agents to perform multi-step interactions rather than just single-view information extraction. To facilitate this, DASHBOARDQA dashboards incorporate four primary UI controls: dropdown menus (92.31%), tabs (23.08%), radio buttons (7.69%), and range sliders (7.69%). Dropdown menus are the most prevalent, appearing in nearly all dashboards with an average of 2.38 each (Table 1).

Table 2 summarizes question diversity by task difficulty. We define task difficulty by the number of interactions. or "actions", an agent must perform to get the correct answer. We approximate this by counting the steps (e.g., mouse clicks, scrolls) a human would take to answer each question. We cate-

gorize difficulty into three levels: EASY, MEDIUM, and HARD, corresponding to 1–2, 3–5, and 6+ required steps, respectively. While MCQ, Factoid, and Hypothetical questions primarily fall into the Easy and Medium levels, Multi-Dashboard and Conversational questions are more skewed toward the Medium and Hard levels. This is because Multi-Dashboard questions often require navigating across multiple visualizations, while Conversational questions involve referencing previous answers, leading to additional interactions. A more granular distribution is shown in Table 7. Overall, DASHBOARDQA contains 112 Easy, 130 Medium, and 50 Hard questions, reflecting its focus on multi-step reasoning and cross-view planning.

4 Methodology

4.1 Task Definition

DASHBOARDQA frames the QA task as an interactive question-answering problem. Formally, given a dataset $\mathcal{D} = (d_i, q_i, a_i)_{i=1}^{N}$, each example comprises an interactive dashboard d_i, an associated question q_i, and a ground truth verifiable answer a_i. The autonomous agent must navigate and interact with the dashboard d_i to accurately generate the correct answer a_i. Solving such tasks requires planning, state tracking, and logical reasoning in dynamic environments.

4.2 Overall Framework

We evaluate agents using an extended version of the OSWorld environment (Xie et al., 2024a), which facilitates interactive navigation within virtual machines. As illustrated in Figure 1, each evaluation begins by launching an Ubuntu OS virtual machine, opening the dashboard URL in a Chrome browser configured in full-screen mode at 1920×1080 resolution. An initial observation (state) of the environment is captured from the virtual environment and provided to the VLM agent, accompanied by a task instruction that includes an input prompt and the associated question from DASHBOARDQA. The VLM agent processes this information and responds with intermediate reasoning steps ("thoughts") and predicted actions, which are executed in the environment. This cycle repeats with updated observations until task completion. Below, we formally define the observation and action spaces.

Observation Space. We employ two types of observations to represent the environment state to the VLM agent: *(i)* **Screenshot**: a raw pixel-level capture of the dashboard interface, mimicking human perception, and *(ii)* **Screenshot + Accessibility Tree (A11y)**: which augments the visual input with structured UI element information. This hybrid input is particularly beneficial for addressing grounding limitations commonly observed in VLMs when relying solely on raw visual data (Xie et al., 2024a). We extract the A11y tree using ATSPI[1].

Action Space. Following the OSWorld framework (Xie et al., 2024a), we define the action space to encompass a wide range of GUI operations, including mouse movements (e.g., clicking, dragging, scrolling) and keyboard inputs (e.g., typing text, pressing specific keys). Additionally, we incorporate three special meta-actions: **WAIT**, **FAIL**, and **DONE**. These allow the agent to *(i)* pause and wait for the environment to update, *(ii)* terminate the episode when the task is deemed unsolvable, or *(iii)* signal task completion, respectively.

All responses from the agent are logged during this process. The process terminates when one of the following conditions is met: *(i)*: the agent successfully completes the task and outputs the **DONE** meta-action, *(ii)* the agent admits failure and outputs the **FAIL** meta-action, or *(iii)* the agent reaches the max number of allowed steps, which is set to 25 steps. Finally, we extract the final answers from the VLM agent's logs (responses). The complete prompts used for model evaluation is presented in Appendix A.5 and Table 14.

5 Evaluation

5.1 Experimental Setup

We conduct experiments under two observations types: dashboard screenshots (primary setting) and screenshots with accessibility (a11y) trees using ATSPI[2]. Evaluations are conducted on Google Cloud, deploying open-source models on either L4 or A100 GPU machines based on model size, utilizing the vLLM framework (Kwon et al., 2023).

5.2 Baselines

We evaluate a comprehensive list of 12 agents categorized into three groups: closed-source, open-source, and hybrid. Closed-source models include GPT4o (OpenAI et al., 2024), GPT5-mini (OpenAI, 2025), O4-mini (OpenAI et al., 2024), Gemini Pro

[1] https://docs.gtk.org/atspi2/
[2] https://docs.gtk.org/atspi2/

Model	Screenshot						Screenshot + A11y Accessibility Tree					
	Factoid	MCQ	Convers.	Hypoth.	Multidash.	Overall	Factoid	MCQ	Convers.	Hypoth	Multidash	Overall
Closed-Source Models												
GPT4-o	8.88	20.00	14.81	**7.69**	4.35	11.50	27.25	35.00	17.86	13.85	10.87	22.94
O4-mini	1.35	1.25	0.00	0.00	0.00	0.68	7.92	22.50	0.00	2.82	2.17	9.14
GPT5-mini	1.35	2.50	0.00	0.00	0.00	1.03	13.11	27.50	7.41	16.55	0.00	15.28
Gemini-Pro-2.5	10.33	15.00	22.22	**7.69**	**8.70**	11.86	**40.03**	**46.25**	**51.20**	**39.27**	**15.22**	**38.69**
OpenAI CUA	**26.45**	**42.50**	**29.63**	7.22	0.00	**22.69**	22.24	38.75	16.93	3.08	0.00	18.50
Open-Source Models												
Jedi-3B w/GPT4o	32.70	40.00	44.85	23.41	**6.67**	29.73	N/A	N/A	N/A	N/A	N/A	N/A
Jedi-7B w/GPT4o	**35.33**	**42.50**	**49.52**	**34.00**	2.17	**33.09**	N/A	N/A	N/A	N/A	N/A	N/A
UI-Tars-2B	0.75	0.00	0.00	0.00	0.00	0.19	N/A	N/A	N/A	N/A	N/A	N/A
UI-Tars-1.5-7B	8.98	**12.50**	0.00	2.40	0.00	6.23	N/A	N/A	N/A	N/A	N/A	N/A
GUI-OWL-7B	**17.12**	1.25	**25.93**	**17.75**	**0.37**	**11.12**	N/A	N/A	N/A	N/A	N/A	N/A

Table 3: Accuracy (%) on DASHBOARDQA across different Observation Types (main headers) and Question Categories (sub-headers). Each Observation block includes five question types along with an Overall score. Color shading indicates model category: closed-source models, hybrid models, and fully open-source models. The highest score within each category is shown in bold.

2.5 (Comanici et al., 2025), and OpenAI's Computer Use Agent (CUA) (ope). Hybrid models combine open-source grounding models with proprietary planning models and include Jedi-3B and 7B w/GPT4o (Xie et al., 2025). Open-source models include UI-TARS 2B and 7B (Qin et al., 2025), and GUI-OWL-7B (Ye et al., 2025).

5.3 Evaluation Metric

We evaluate performance using the enhanced relaxed accuracy metric (Masry et al., 2025a), which applies a 5% tolerance for numerical answers, exact matching for years, and ANLS (Biten et al., 2019) for textual responses.

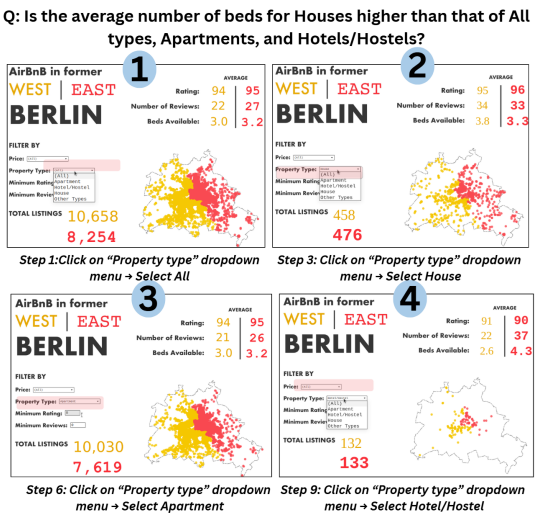
5.4 Main Results

Table 3 presents accuracy across two observation settings: Screenshots and Screenshots with Accessibility Tree (A11y), and five question categories: Factoid, MCQ, Conversational, Hypothetical, and Multi-Dashboard. With A11Y, Gemini Pro 2.5 achieves the strongest overall accuracy (38.69%), outperforming other closed-source agents (GPT-4o: 22.94%, OpenAI CUA: 18.50%). Under the SCREENSHOTS-ONLY setting, all closed-source agents remain below 23%, indicating grounding and navigation challenges when relying solely on visual input. In this setting, the best overall result comes from the hybrid setup Jedi-7B with GPT-4o (33.09%), well above OpenAI CUA (22.69%). These gaps highlight both the value of structured UI representations (A11y) and the advantage of specialized hybrid pipelines for UI grounding. Among the question categories, Multi-Dashboard remains the hardest category: even with A11y, the best accuracy is 15.22% (Gemini Pro 2.5), reflecting persistent difficulties with cross-view reasoning and long-horizon planning. Overall, DASHBOARDQA exposes significant weaknesses in grounding, planning, and multi-view visual reasoning.

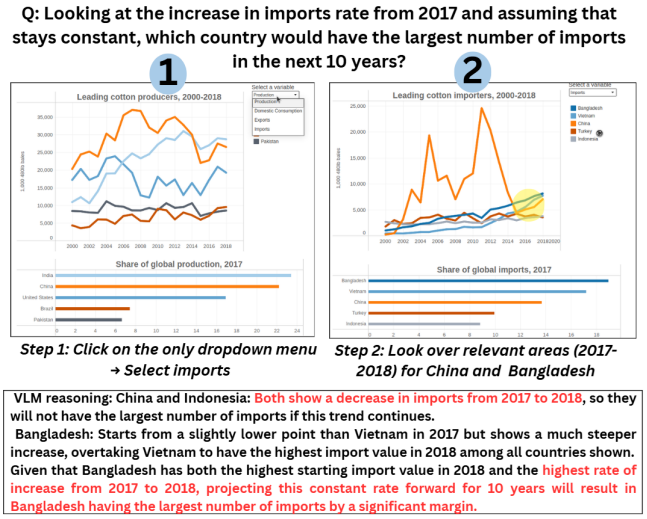
Task Difficulty	Gemini Pro 2.5 (Screenshot + A11y)	Jedi-7B w/ GPT4o (Screenshot)
Easy	50.64	40.40
Medium	32.17	32.31
Hard	28.86	18.74
Overall	38.69	33.09

Table 4: Ablation results on DASHBOARDQA across task difficulty levels: EASY, MEDIUM, and HARD.

We further conducted a quantitative comparison of the reasoning behaviors exhibited by the strongest model under each setup: Gemini-Pro-2.5 (Screenshot + A11y setup) and Jedi-7B w/ GPT-4o (Screenshot setup). Specifically, we analyzed four key metrics: *response lengths, step distributions, frequency of reaching the maximum step limit, and action type frequencies*. Figure 15 illustrates the distributions of steps and reasoning thought lengths for both models. To avoid skewing the results, we excluded 11 extreme outliers in Gemini's thought lengths (exceeding 10K characters). For Jedi-7B, the reasoning traces are relatively short, with all actions restricted to simple `click` operations. Nevertheless, the model frequently fails to complete tasks within the allowed trajectory length, reaching the maximum 25-step limit in 56 cases. In contrast, Gemini exhibits more diverse interaction patterns, with actions distributed across three categories: `click` (2,920 instances), `moveTo` (3,796 instances), and `hotkey` (876 instances). Despite this richer action space, Gemini Pro 2.5 hits the 25-step ceiling in 77 cases.



(a) Plan Tracking Error



(b) Visual Reasoning Error

Figure 4: Reasoning errors: (a) Plan tracking: the VLM follows the correct states but fails to use them, looping until step limit. (b) Visual reasoning: the VLM misinterprets visual information in a multi-line chart.

5.5 Ablation Studies

Table 10 presents the ablation results on DASHBOARDQA across three task difficulty levels defined in Table 2: EASY, MEDIUM, and HARD. We analyze Gemini Pro 2.5 and Jedi-7B w/ GPT4o, the two strongest models from our main results under the *Screenshot + A11y* and *Screenshot* setups, respectively. Both models achieve higher accuracy on easier tasks, with performance gradually declining as task complexity increases. Gemini Pro 2.5 surpasses Jedi-7B on the Easy and Hard categories, while both perform comparably on Medium tasks. Overall, these results highlight the need for future models to strengthen multi-step reasoning, long-horizon planning, and context retention to handle complex dashboard interactions more effectively.

5.6 Qualitative Analysis

We examined 20 randomly sampled cases from the strongest open-source and closed-source models **Jedi-7B w/GPT-4o (Screenshot)** and **Gemini-Pro-2.5 (Screenshot + A11y Accessibility Tree)**. These were drawn from instances where at least one model failed or reached the 25-step limit. Our review revealed recurring error patterns.

Plan Tracking Errors: Both models frequently lost track of their initial reasoning plans. Steps that had already been completed were redundantly repeated, causing loops that often exhausted the 25-step budget. In Figure 4(a), for example, the model had all necessary details by step 9 but kept re-cycling through earlier actions until termination.

Visual Reasoning Errors: Models also made blatant reasoning mistakes or hallucinations. Figure 4(b) illustrates a misinterpretation of a multi-line chart that led to an incorrect answer.

Failure analysis	Gemini Pro 2.5 (Screenshot + A11y)	Jedi-7B w/ GPT4o (Screenshot)
Navigation Tool Usage errors	66	29
Visualization Tool Usage errors	43	10
Planning failure errors	25	19
Overall	134	58

Table 5: Failure quantification for SOTA models on DASHBOARDQA.

Information Retention Errors: Models sometimes misremembered or clung to incorrect details. Figure 13(c) shows a case where a year from an earlier step was misrecalled; in other cases, screenshots were taken before the dashboard fully updated, caused by underestimating the required waiting time, locking in the wrong state.

Table 5 depicts the number of tasks the best performing models encountered certain error types in. Notably, the hybrid model approach was less prone to all error types, but still wasn't able to match the performance of Gemini Pro 2.5.

6 Conclusion

We introduced DASHBOARDQA, the first benchmark for question answering over interactive dashboards, aimed at advancing the capabilities of vision-language and GUI agents in realistic, exploratory data analysis tasks. Unlike prior benchmarks focused on static charts or utility-based GUI tasks, DASHBOARDQA requires agents to plan and execute complex sequences of GUI interactions to answer diverse question types including factoid,

multiple-choice, hypothetical, and conversational queries. Our experiments on this benchmark expose critical limitations in current models, with even the strongest agents experiencing substantial performance drops. Through extensive evaluation and analysis, we identify key challenges in grounding, planning, and visual reasoning, highlighting open problems in interactive multimodal understanding. We hope DASHBOARDQA catalyzes future research toward more robust, grounded, and reasoning-capable agents for dynamic, real-world visual environments.

Limitations

Firstly, this is a Tableau-centric dataset. This limits the visual and topic diversity that would exist otherwise if dashboards from alternative visualization platforms such as Power BI and Zoho analytics were included. In addition, the website URLS on tableau public were utilized in this dataset. Future work on interactive dashboard question answering can focus more on essential GUI grounding tasks on the desktop app.

Ethical Considerations

During the dashboard curation process authors were instructed to not collect controversial, divisive, or harmful dashboards, and to rather focus on more neutral and academic sources. This includes of sources including of United Nation subgroups and university institutions (i.e Oxford University). All dashboards were taken from Tableau Public, where they are publicly available and permitted for use in developing AI models[3]

All QA pairs were either fully authored by the authors or generated by vision-language models (VLMs) and subsequently reviewed and validated by the authors. Each author has expertise in data visualization comprehension tasks. The annotators were also informed that their annotations would be released with the dataset for research purposes. Additionally, AI-writing tools were employed to improve the clarity, structure, and overall writing quality of the paper.

Acknowledgement

This work was supported by the Natural Sciences Engineering Research Council (NSERC) of Canada, Canada Foundation for Innovation (CFI), and York University's Catalyzing Interdisciplinary Research Clusters fund. It also received support through a Google Cloud Platform (GCP) credits award from Google's PaliGemma Academic Program.

[3] https://public.tableau.com/app/data-policy

References

Computer-Using Agent — openai.com. https://openai.com/index/computer-using-agent/. [Accessed 14-07-2025].

Saaket Agashe, Kyle Wong, Vincent Tu, Jiachen Yang, Ang Li, and Xin Eric Wang. 2025. Agent s2: A compositional generalist-specialist framework for computer use agents. *Preprint*, arXiv:2504.00906.

Md Zubair Ibne Alam, Shehnaz Islam, and Enamul Hoque. 2023. Seechart: Enabling accessible visualizations through interactive natural language interface for people with visual impairments. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 46–64.

Anthropic. 2024. Introducing the next generation of claude.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. *Preprint*, arXiv:1905.13648.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing gui grounding for advanced visual gui agents. *Preprint*, arXiv:2401.10935.

Imran Chowdhury, Abdul Moeid, Enamul Hoque, Muhammad Ashad Kabir, Md Sabir Hossain, and Mohammad Mainul Islam. 2020. Designing and evaluating multimodal interactions for facilitating visual analysis with dashboards. *Ieee Access*, 9:60–71.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, and et al. Inderjit Dhillon. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Gregorio Convertino, Jian Chen, Beth Yost, Y-S Ryu, and Chris North. 2003. Exploring context switching and cognition in dual-view coordinated visualizations. In *Proceedings International Conference on Coordinated and Multiple Views in Exploratory Visualization-CMV 2003-*, pages 55–62. IEEE.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig,

Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, and 1117 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Enamul Hoque, Parsa Kavehzadeh, and Ahmed Masry. 2022. Chart question answering: State of the art and future directions. *Computer Graphics Forum*, 41(3):555–572.

Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2018. Applying pragmatics principles for interaction with visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):309–318.

Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are large vision language models up to the challenge of chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3334–3368, Miami, Florida, USA. Association for Computational Linguistics.

Kushal Kafle, Scott Cohen, Brian L. Price, and Christopher Kanan. 2018. DVQA: understanding data visualizations via question answering. *CoRR*, abs/1801.08163.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. *Preprint*, arXiv:1710.07300.

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022. Opencqa: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11817–11837.

Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. 2025. Screenspot-pro: Gui grounding for professional high-resolution computer use. *Preprint*, arXiv:2504.07981.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025a. ChartQAPro: A more diverse and challenging benchmark for chart question answering. *Preprint*, arXiv:2504.05506.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *Preprint*, arXiv:2305.14761.

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. *arXiv preprint arXiv:2203.10244*.

Ahmed Masry, Abhay Puri, Masoud Hashemi, Juan A. Rodriguez, Megh Thakkar, Khyati Mahajan, Vikas Yadav, Sathwik Tejaswi Madhusudhan, Alexandre Piché, Dzmitry Bahdanau, Christopher Pal, David Vazquez, Enamul Hoque, Perouz Taslakian, Sai Rajeswar, and Spandana Gella. 2025b. Bigcharts-r1: Enhanced chart reasoning with visual reinforcement finetuning. In *Second Conference on Language Modeling*.

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024a. Chartinstruct: Instruction tuning for chart comprehension and reasoning. *Preprint*, arXiv:2403.09028.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2024b. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. *Preprint*, arXiv:2407.04172.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021. Infographicvqa. *Preprint*, arXiv:2104.12756.

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

OpenAI. 2025. Introducing gpt-5. Accessed: 2025-10-01.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, and 16 others. 2025. Ui-tars: Pioneering automated gui interaction with native agents. *Preprint*, arXiv:2501.12326.

Mizanur Rahman, Md Tahmid Rahman Laskar, Shafiq Joty, and Enamul Hoque. 2025. Text2vis: A challenging and diverse benchmark for generating multimodal visualizations from text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31837–31862.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. 2025. Androidworld: A dynamic benchmarking environment for autonomous agents. *Preprint*, arXiv:2405.14573.

Tableau Software. 2025. Tableau public. https://www.tableau.com/products/public. Accessed: 2025-06-02.

Tableau Software. n.d. Discover tableau. Accessed: 2025-07-20.

Zilong Wang, Yuedong Cui, Li Zhong, Zimin Zhang, Da Yin, Bill Yuchen Lin, and Jingbo Shang. 2024a. Officebench: Benchmarking language agents across multiple applications for office automation. *Preprint*, arXiv:2407.19056.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Preprint*, arXiv:2406.18521.

Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Zekun Wang, Yiheng Xu, Junli Wang, Doyen Sahoo, Tao Yu, and Caiming Xiong. 2025. Scaling computer-use grounding via user interface decomposition and synthesis. *Preprint*, arXiv:2505.13227.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024a. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.

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. 2024b. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Preprint*, arXiv:2404.07972.

Yan Yang, Dongxu Li, Yutong Dai, Yuhao Yang, Ziyang Luo, Zirui Zhao, Zhiyuan Hu, Junzhe Huang, Amrita Saha, Zeyuan Chen, Ran Xu, Liyuan Pan, Caiming Xiong, and Junnan Li. 2025. Gta1: Gui test-time scaling agent. *Preprint*, arXiv:2507.05791.

Jiabo Ye, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Zhaoqing Zhu, Ziwei Zheng, Feiyu Gao, Junjie Cao, Zhengxi Lu, Jitong Liao, Qi Zheng, Fei Huang, Jingren Zhou, and Ming Yan. 2025. Mobile-agent-v3: Fundamental agents for gui automation. *Preprint*, arXiv:2508.15144.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. *Preprint*, arXiv:2307.13854.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. 2025. Multichartqa: Benchmarking vision-language models on multi-chart problems. *Preprint*, arXiv:2410.14179.

A Appendices

A.1 Dataset Analysis

Tableau (Tableau Software, n.d.) is a widely used interactive visualization tool and platform for creating, publishing, and exploring dashboards that allows users to freely share visualizations online and browse a large repository of dashboards created by others. We chose Tableau Public for the following reasons: (i) scale and topic and source diversity, with millions of dashboards covering finance, health, public policy, sports, and more (see Figure 3, 11, and 10); (ii) rich interaction patterns among the dashboards, including coordinated views, sliders, filters, and tooltips essential for multi-step reasoning; and (iii) reproducibility, since many dashboards expose dataset URLs or allow crosstab export, enabling deterministic ground-truth verification—something not reliably supported on other public platforms. To mitigate concerns about URL fragility, we release the extracted dashboards and their ground-truth crosstabs alongside trajectories.

We present detailed statistics on the visualization types and the navigation tool types in DASHBOARDQA dashboards in Table 8 and Table 1 respectively. Additionally, we show the distribution of the number of states traversed and steps needed to answer the questions in 6 and 7. A state represents the current configuration of selected navigation tool options of the dashboard. A step is defined as a round of mouse or keyboard actions that will either change the configuration of a navigation tool or drill down on a particular visualization within the dashboard.

Category Type	Views required to answer						
	1	2	3	4	5	6	7+
MCQ	0	41	25	7	3	4	0
Factoid	5	14	21	21	9	2	2
Hypothetical	4	30	19	7	2	2	1
Multi-Dashboard	0	9	17	13	0	5	2
Conversational	0	4	7	8	4	2	2

Table 6: Distribution of number of dashboard states required to traverse for the respective question categories.

Question Category	Easy		Medium			Hard			Total
	1	2	3	4	5	6	7	8+	
MCQ	14	24	15	11	6	7	3	0	80
Factoid	10	21	19	12	3	6	1	2	74
Hypothetical	18	13	12	8	6	4	3	1	65
Multi-Dashboard	3	2	7	8	9	8	1	8	46
Conversational	1	6	6	4	4	2	1	4	27
Total	**46**	**66**	**59**	**43**	**28**	**27**	**9**	**15**	**292**

Table 7: Distribution of difficulty levels and corresponding number of steps across different question categories in DASHBOARDQA.

Visualization	Prevalence	Count	Avg. per dashboard
Area Chart	16.83	31	0.28
Bar Chart	42.57	79	0.71
Bubble Chart	5.94	7	0.06
Bump Chart	0.99	1	0.01
Donut Chart	1.98	2	0.02
Heat Map	5.94	7	0.06
Histogram	1.98	2	0.02
Line Chart	48.51	86	0.77
Map	20.79	25	0.22
Pie Chart	6.93	9	0.08
Scatter Plot	10.89	30	0.27
Crosstab	10.89	12	0.11
Tree Map	4.95	7	0.06
Total	**100%**	**298**	**2.66**

Table 8: Prevalence, count, and average number of respective visualization types across all dashboards, along with overall average of all visualizations.

A.2 Question source makeup

Our QA generation pipeline was designed to balance VLM assistance with strong human oversight. We used multiple VLMs (GPT-4o, Gemini, and Claude) to ensure topic, wording, and question diversity and to avoid the bias that would arise from relying on a single model family. Most questions in the dataset are either human-authored or human-edited, and very few questions were accepted verbatim from a VLM. This substantially reduces the risk of leakage or model-specific bias. Table 11 depicts the distribution of the sources for the question answer pairs generated.

A.3 Prompts for QA pair generation

For reproducibility, we provide the exact prompts used to generate QA pairs in Tables 12 and 13.

A.4 Sample QA pairs for each category

We present multiple walkthroughs of sample questions from each category from 5 to 9.

A.5 Prompts and Step Limit for Models Evaluation

For reproducibility, we provide the prompts used for model evaluation, adapted from the ChartQAPro benchmark (Masry et al., 2025a) in Table 14. Additionally, our decision for setting a 25-step cap was chosen based on two practical considerations: (1) empirical human trajectories, where all tasks were solvable within 8–12 steps during annotation (see Table 6 in Appendix), and (2) practical inference cost, since a single full pass of Gemini Pro 2.5 with a11y input already costs 303 USD for 25 steps, making large-scale sweeps prohibitively expensive. To evaluate whether the step limit affected model performance, we conducted a focused ablation on 50 randomly sampled tasks, doubling the step budget from 25 $\rightarrow$ 50 for Gemini Pro 2.5 under the (A11y + screenshot) setting. The results (shown in the table 11) indicate no measurable improvement in accuracy when additional steps were allowed and additional steps primarily led to looping or repeated mis-grounding rather than progress toward the correct state.

A.6 Quantitative Analysis

Figure 15 shows the distribution of steps and thought lengths for both Gemini Pro 2.5 and Jedi 7B w/GPT4o. For clarity, we excluded 11 extreme outliers in Gemini's thought lengths (exceeding 10K characters) to avoid skewing the histograms.

Failure analysis	Gemini Pro 2.5 (Screenshot + A11y)	Jedi-7B w/ GPT4o (Screenshot)
tasks with dropdown menu errors	40	14
tasks with range slider errors	11	3
tasks with tabs errors	12	11
tasks with radio button errors	3	1
Overall	134	58

Table 9: Failure quantification for SOTA models on DASHBOARDQA by navigation tool type.

Gemini Pro 2.5	Accuracy (Screenshot + A11y)
25 steps	35.33%
50 steps	35.33%

Table 10: Step limits and accuracy rates for Gemini Pro 2.5 on DASHBOARDQA.

Question Category	Human-made	Assisted			Model-Only		
		GPT4o	Claude	Gemini	GPT4o	Claude	Gemini
MCQ	39	15	10	7	5	1	3
Factoid	34	11	0	6	3	0	13
Hypothetical	28	19	8	6	1	0	0
Multi-Dashboard	18	8	8	10	2	0	0
Conversational	9	7	6	6	0	0	0

Table 11: Counts of question creation methods through Human or VLMs.

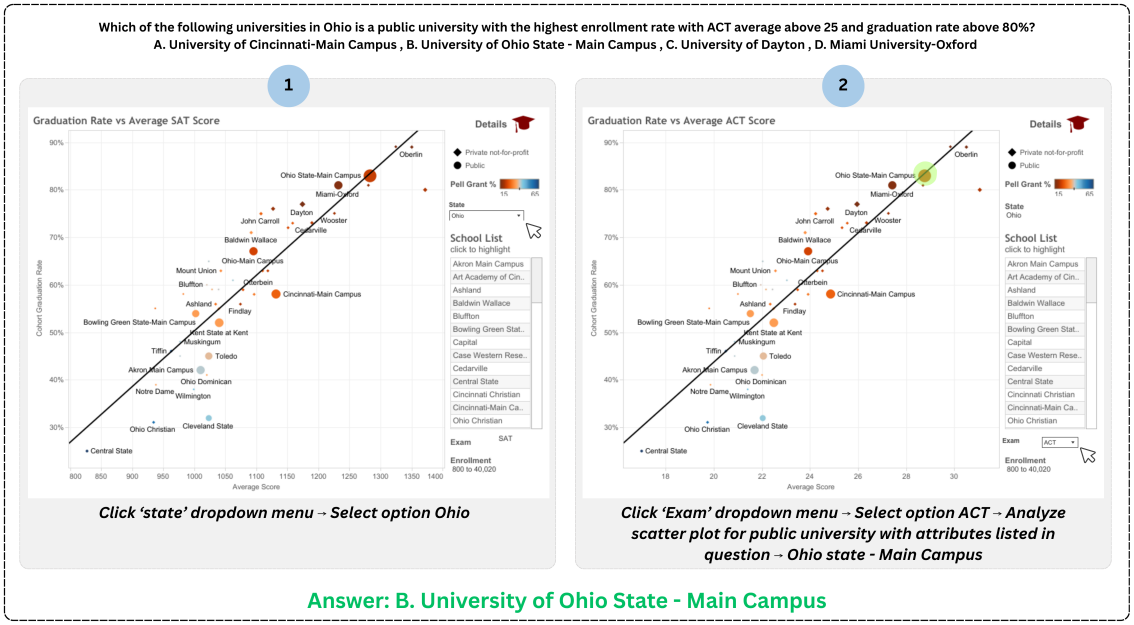


Figure 5: Example MCQ question in DashboardQA

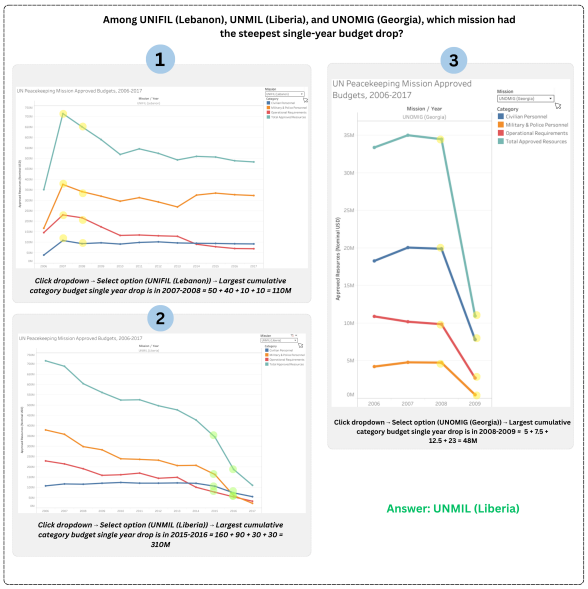


Figure 6: Example factoid question in DashboardQA

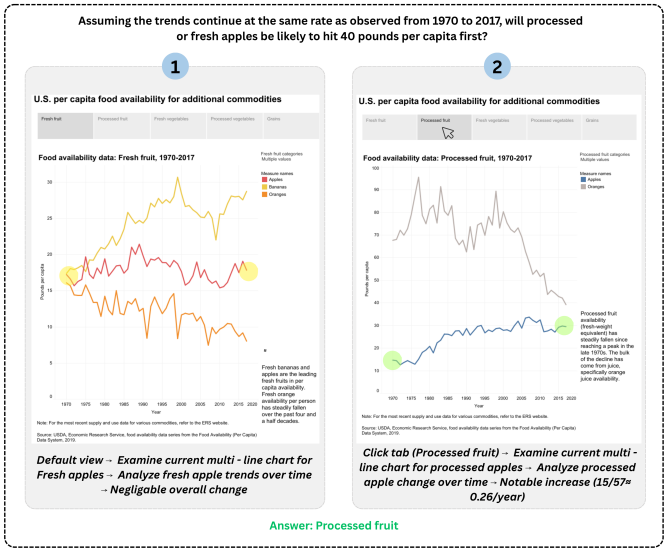


Figure 7: Example hypothetical question in DashboardQA

Figure 8: Example multiple dashboard question in DashboardQA

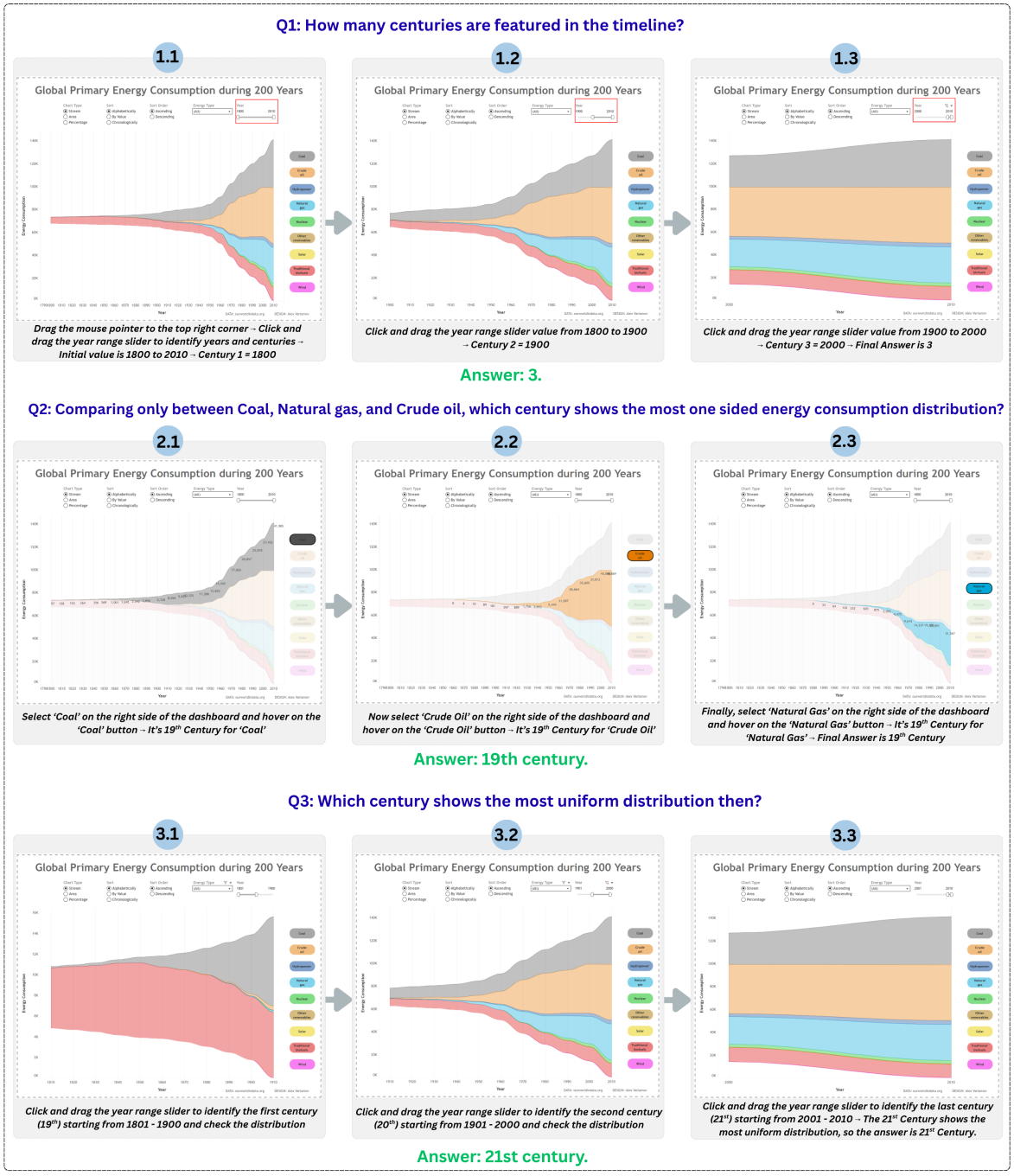


Figure 9: Example conversational question in DashboardQA

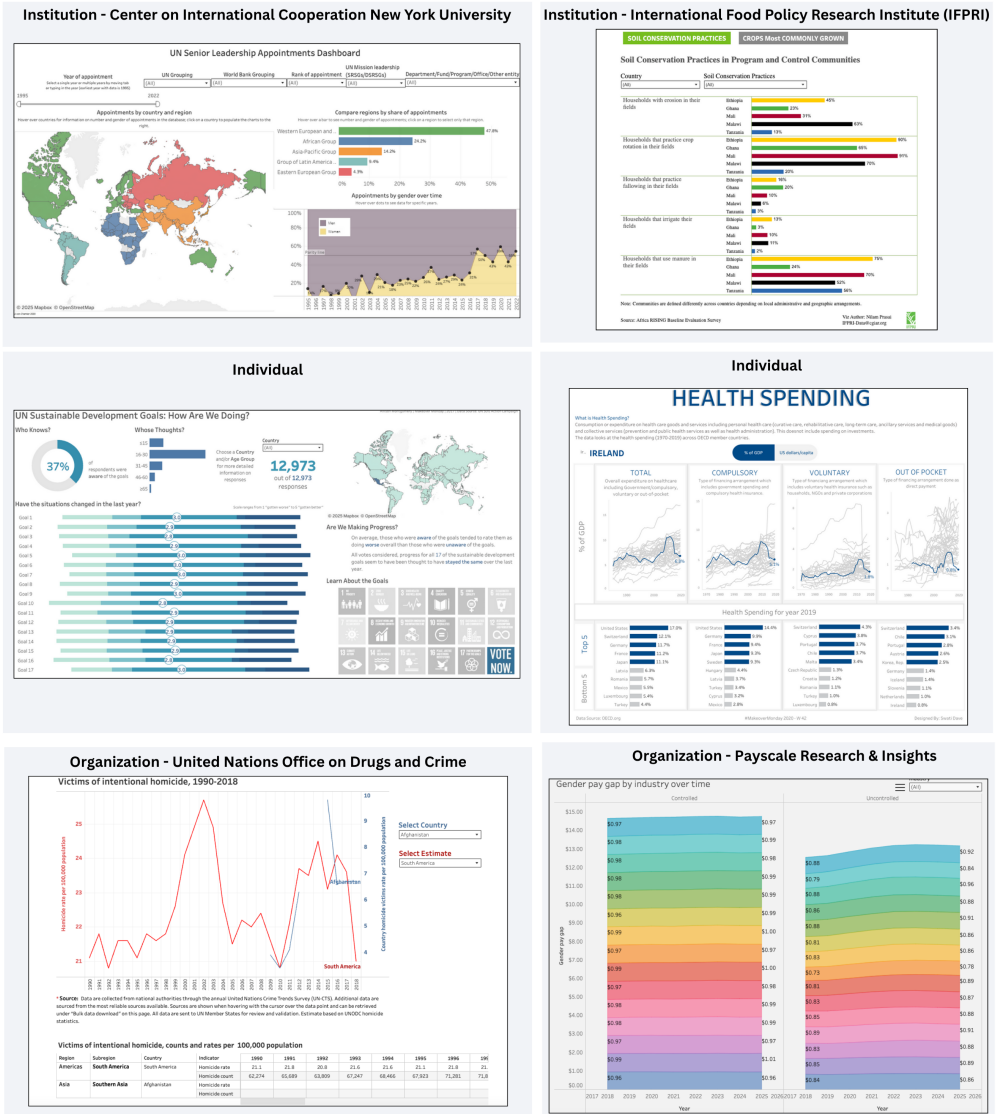


Figure 10: Example Tableau dashboards collected from various institutions,organizations,and individuals.

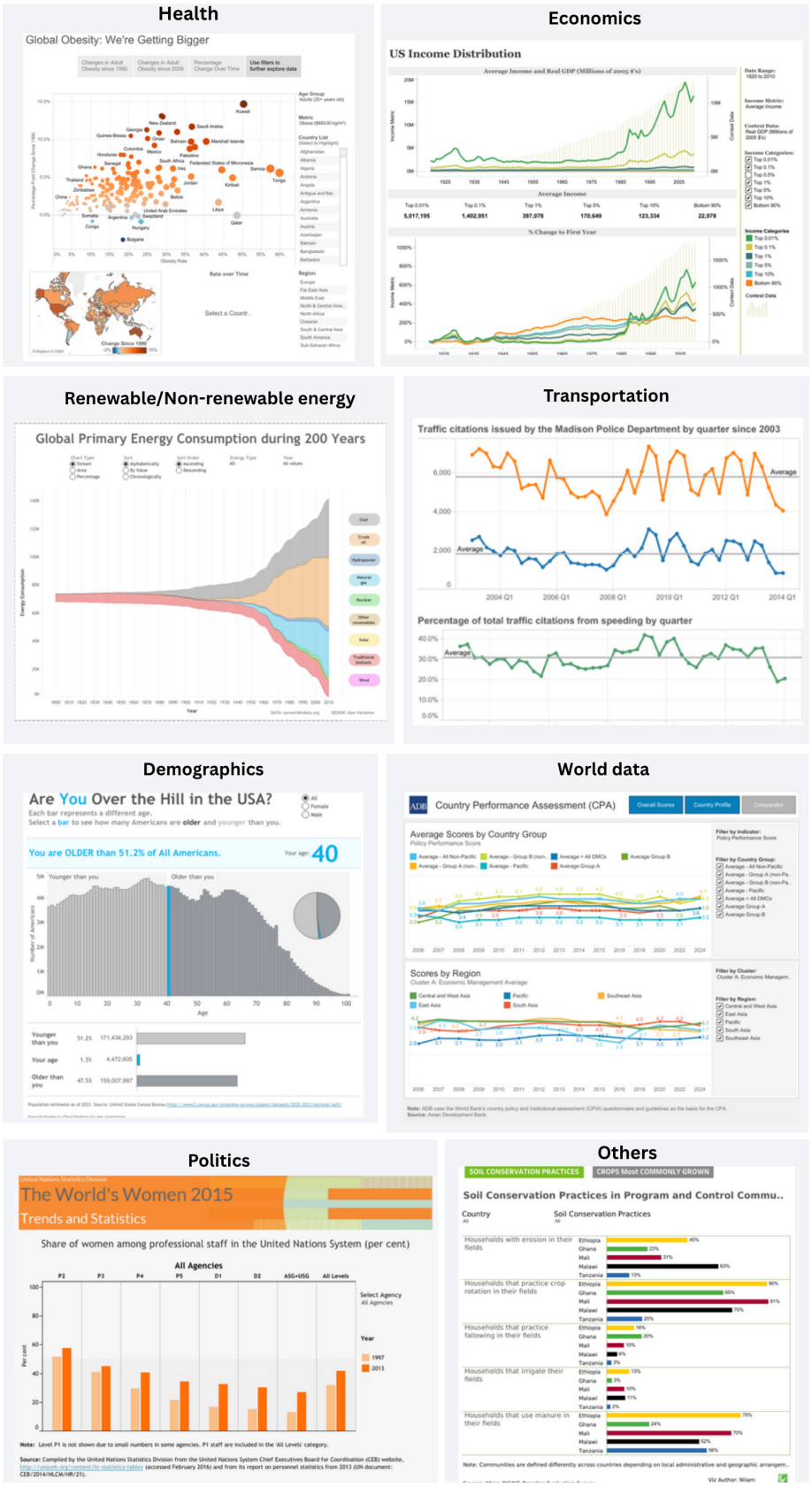


Figure 11: Sample dashboards chosen for all of the topics in DashboardQA

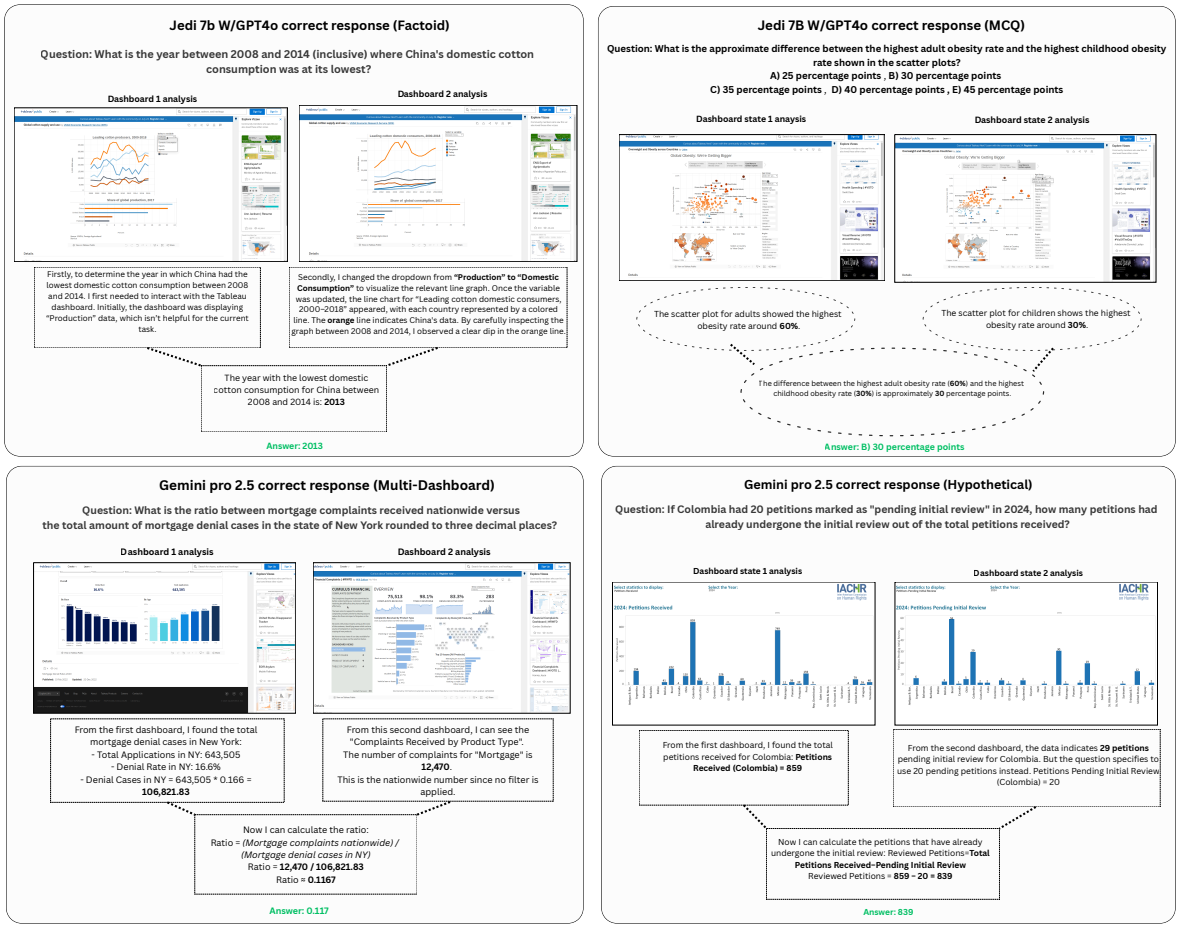


Figure 12: Examples of QA pairs that have received correct responses, and their corresponding dashboard and VLM.

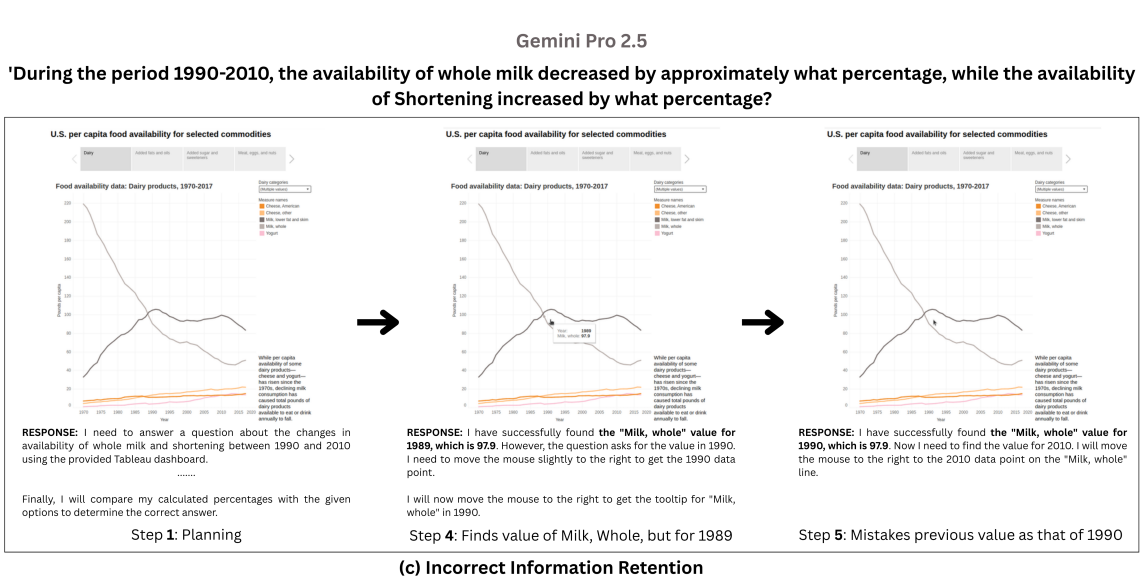


Figure 13: Additional error type seen from SOTA VLM and hybrid approaches.

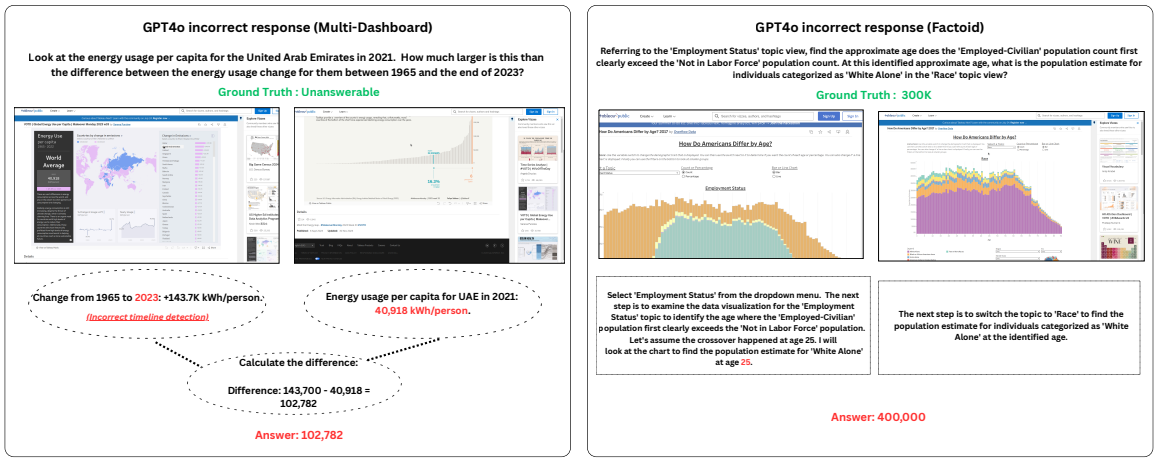
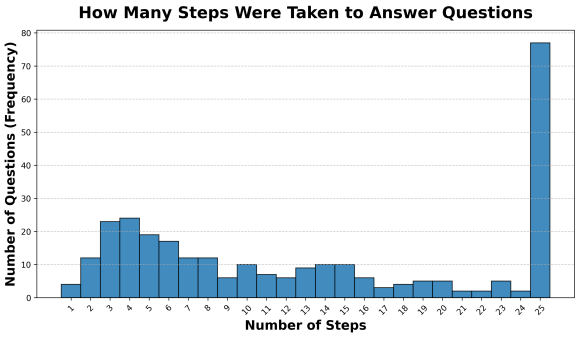
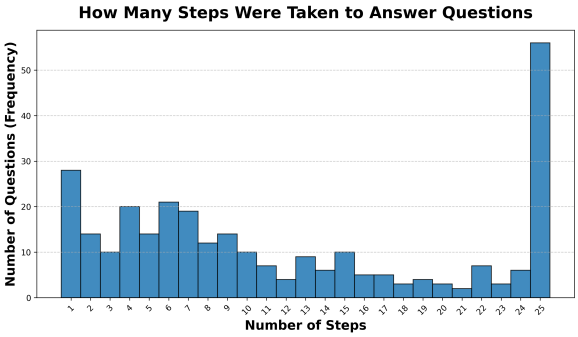


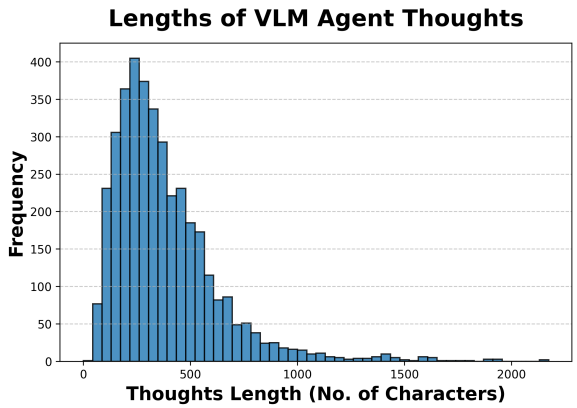
Figure 14: Additional examples of errors, as done by GPT4o. On the left an unanswerable question is given with an out of bounds date (2023), but GPT4o misses this. Additionally, an incorrect value is extracted from the second dashboard looked at. On the right for this multi-hop reasoning question we have a mistaken value extracted from the first dashboard and this error propagates to the information extracted from the next, resulting in an incorrect answer.



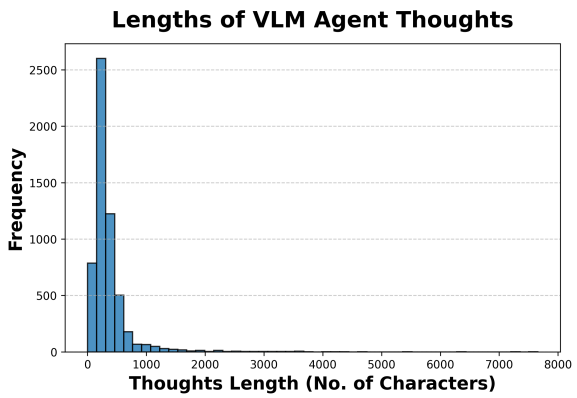
(a) Distribution of steps taken by **Gemini-Pro-2.5**.



(b) Distribution of steps taken by **Jedi-7B w/ GPT-4o**.



(c) Lengths of reasoning thoughts generated by **Gemini**.



(d) Lengths of reasoning thoughts generated by **Jedi-7B**.

Figure 15: Comparison of reasoning behaviors between Gemini-Pro-2.5 (left) and Jedi-7B w/ GPT-4o (right) on the DASHBOARDQA benchmark. The plots illustrate (a, b) the distribution of the number of interaction steps taken to answer questions and (c, d) the lengths of intermediate reasoning thoughts

Category	Prompt Template
Factoid	I am giving you multiple snapshots/views of multiple interactive dashboards, including of their respective default views, options available for navigation through the dashboard's navigation tools (i.e Dropdown menus, radio buttons, scrollers, etc.), and more views from when particular options are selected. Since not every possible snapshot / view of each dashboard is given, the selection options available and the example views for each dashboard should together help you guess how the views for the rest of the options would be structurally (i.e types and content of the charts or textual data available) for that particular dashboard. Using all of these, please generate 5 diverse and challenging factoid reasoning questions that involve arithmetic and logical reasoning. They must have multiple operators (i.e sum, ratio, etc.) in the question. Additionally, make sure the answers required are succinct, where they are either a numerical value, True or False, Yes or No, unanswerable or simply a label taken from the dashboard itself. Each question must require 3 to 5 views (Cumulative) of the dashboards to answer. Just the questions are to be given now from your end, in the next prompt the required views will be given for you to answer. Here are some sample questions to give some inspiration, but make sure to create the questions with lexical and semantic differences. Samples : <samples >
Multi Choice	I am giving you multiple snapshots/views of multiple interactive dashboards, including of their respective default views, options available for navigation through the dashboard's navigation tools (i.e Dropdown menus, radio buttons, scrollers, etc.), and more views from when particular options are selected. Since not every possible snapshot / view of each dashboard is given, the selection options available and the example views for each dashboard should together help you guess how the views for the rest of the options would be structurally (i.e types and content of the charts or textual data available) for that particular dashboard. Using all of these, please generate 5 diverse and challenging MCQ reasoning questions that involve arithmetic and logical reasoning. There may be 4-5 answer options available, but only one of them is correct. They must have multiple operators (i.e sum, ratio, etc.) in the question. Additionally, make sure the answers required are succinct, where they are either a numerical value, True or False, Yes or No, unanswerable or simply a label taken from the dashboard itself. Each question must require 2 to 5 views of the dashboard to answer. Just the questions are to be given now from your end, in the next prompt the required views will be given for you to answer. Remember to not be biased towards just the views given, and to reason as to how the others would look like as to help your questions. Here are some sample questions to give some inspiration, but make sure to create the questions with lexical and semantic differences. Samples : <samples >
Hypothetical	I am giving you multiple snapshots/views of multiple interactive dashboards, including of their respective default views, options available for navigation through the dashboard's navigation tools (i.e Dropdown menus, radio buttons, scrollers, etc.), and more views from when particular options are selected. Since not every possible snapshot / view of each dashboard is given, the selection options available and the example views for each dashboard should together help you guess how the views for the rest of the options would be structurally (i.e types and content of the charts or textual data available) for that particular dashboard. Using all of these, please generate 5 diverse and challenging hypothetical questions that involve arithmetic and logical reasoning. This may involve looking into future trends, extrapolation tasks, or looking at counterfactuals for example. They must have multiple operators (i.e sum, ratio, etc.) in the question. Additionally, make sure the answers required are succinct, where they are either a numerical value, True or False, Yes or No, unanswerable or simply a label taken from the dashboard itself. Each question must require from 2 to 5 views of the dashboard to answer. Just the questions are to be given now from your end, in the next prompt the required views will be given for you to answer. Here are some sample questions to give some inspiration, but make sure to create the questions with lexical and semantic differences. Samples : <samples >

Table 12: Prompt Templates for Factoid, MCQ, and Hypothetical question answer generation. These represent the single turn, single dashboard question types.

Category	Prompt Template
Multi-Dashboard	I am giving you multiple snapshots/views respectively for multiple interactive dashboards, including of their respective default views, options available for navigation through the dashboard's navigation tools (i.e Dropdown menus, radio buttons, scrollers, etc.), and more views from when particular options are selected. Since not every possible snapshot / view of each dashboard is given, the selection options available and the example views for each dashboard should together help you guess how the views for the rest of the options would be structurally (i.e types and content of the charts or textual data available) for that particular dashboard. Using all of these, please generate 5 diverse and high level questions surrounding all of the dashboards given that involve arithmetic and logical reasoning. This could include having multiple operators (i.e sum, ratio, etc.) in the question. Additionally, make sure the answers required are succinct, where they are either a numerical value, True or False, Yes or No, unanswerable or simply a label taken from the dashboard itself. Each question must require from 2 to 5 views of the dashboard to answer, and a majority of the follow up questions should require utilizing one or more of the previous questions to help. Just the questions are to be given now from your end, in the next prompt the required views will be given for you to answer. Here are some sample questions to give some inspiration, but make sure to create the questions with lexical and semantic differences. Samples : <samples >
Conversational	I am giving you multiple snapshots/views of an interactive dashboard, including the default view, the options available for navigation through the dashboard's navigation tools (i.e Dropdown menus, radio buttons, scrollers, etc.), and more views from when particular options are selected. Since not every possible snapshot / view are going to be given, the selection options available and the example views should together help you guess how the views for the rest of the options would be structurally (i.e types and content of the charts or textual data available for that particular dashboard). Using all of these, please generate a conversation of 4-7 questions that analyzes the contents of the dashboard and requires multiple operations (i.e sum, ratio etc.). With a focus on arithmetic and logical reasoning, make sure that a majority of the questions follow up on each other. Use ambiguous pronouns in the latter questions when making references to previous questions. Additionally, make sure the answers required are succinct, where they are either a numerical value, True or False, Yes or No, or simply a label taken from the dashboard itself. The conversation must require from 2 to 5 views of the conversation to answer, where each individual question can either require just one to three views to answer respectively. Just the questions are to be given now from your end, in the next prompt the required views will be given for you to answer. Here are some sample conversations to give some inspiration, but make sure to create the conversations with different lexical and semantic differences. The respective dashboard images are not given for these samples, they are just to get an idea (Not purely representative) of the style of conversations. Samples : <samples >

Table 13: Prompt Templates for Multi-Dashboard and conversational question answer generation. These represent the multi-turn or multi-dashboard question types.

Category	Prompt Template
Factoid	You are given a factoid question about an interactive Tableau dashboard that you need to navigate to answer the question. You need to think step-by-step, but your final answer should be a single word, number, or phrase. Do not generate units. But if numerical units such as million, m, billion, B, or K are required, use the exact notation shown in the dashboard. Remember to navigate the dashboard, think step-by-step, and put the final answer between these brackets <answer> </answer> Question: <question>
Multi Choice	You are given a question about an interactive Tableau dashboard along with different possible answers. You need to navigate the dashboard to select the correct answer from them. You need to think step-by-step, but your final answer should be one of the options letters only (without any additional text). Remember to navigate the dashboard, think step-by-step, and put the final answer between these brackets <answer> </answer>. Question: <question>
Hypothetical	You are given a hypothetical question about an interactive Tableau dashboard that you need to navigate to answer the question. You need to think step-by-step, but your final answer should be a single word, number, or phrase. Do not generate units. But if numerical units such as million, m, billion, B, or K are required, use the exact notation shown in the dashboard. Remember to navigate the dashboard, think step-by-step, and put the final answer between these brackets <answer> </answer> Question: <question>
Multi Dashboards	You are given a question about two interactive Tableau dashboards that are open in two tabs in the browser. You need to navigate them to answer the question. You need to think step-by-step, but your final answer should be a single word, number, or phrase. Do not generate units. But if numerical units such as million, m, billion, B, or K are required, use the exact notation shown in the dashboard. Remember to navigate the dashboard, think step-by-step, and put the final answer between these brackets <answer> </answer> Question: <question>
Conversational	You are given a multi-turn conversation, and your job is to answer the final question based on the conversation history and the information in the provided interactive Tableau dashboard that you need to navigate. You need to think step-by-step, but your final answer should be a single word, number, or phrase. Do not generate units. But if numerical units such as million, m, billion, B, or K are required, use the exact notation shown in the dashboard. Remember to navigate the dashboard, think step-by-step, and put the final answer between these brackets <answer> </answer> Question: <question with conversation history>

Table 14: Prompt Templates for Each Question Category in the Direct setup.