IGenBench: Benchmarking the Reliability of Text-to-Infographic Generation

Yinghao Tang[1], Xueding Liu[2], Boyuan Zhang[1], Tingfeng Lan[3], Yupeng Xie[4], Jiale Lao[5], Yiyao Wang[1], Haoxuan Li[1], Tingting Gao[6], Bo Pan[1], Luoxuan Weng[1], Xiuqi Huang[1*], Minfeng Zhu[6], Yingchaojie Feng[7*], Yuyu Luo[4*], Wei Chen[1]

[1]*State Key Lab of CAD&CG, Zhejiang University*,
[2]*UESTC*, [3]*University of Virginia*, [4]*HKUST(GZ)*,
[5]*Cornell University*, [6]*Zhejiang University*, [7]*National University of Singapore*

Abstract

Infographics are composite visual artifacts that combine data visualizations with textual and illustrative elements to communicate information. While recent text-to-image (T2I) models can generate aesthetically appealing images, their reliability in generating infographics remains unclear. Generated infographics may appear correct at first glance but contain easily overlooked issues, such as distorted data encoding or incorrect textual content. We present IGenBench, the first benchmark for evaluating the reliability of text-to-infographic generation, comprising 600 curated test cases spanning 30 infographic types. We design an automated evaluation framework that decomposes reliability verification into atomic yes/no questions based on a taxonomy of 10 question types. We employ multimodal large language models (MLLMs) to verify each question, yielding question-level accuracy (Q-ACC) and infographic-level accuracy (I-ACC). We comprehensively evaluate 10 state-of-the-art T2I models on IGenBench. Our systematic analysis reveals key insights for future model development: (i) a three-tier performance hierarchy with the top model achieving Q-ACC of 0.90 but I-ACC of only 0.49; (ii) data-related dimensions emerging as universal bottlenecks (e.g., Data Completeness: 0.21); and (iii) the challenge of achieving end-to-end correctness across all models. We release IGenBench at `https://igen-bench.vercel.app/`.

1 Introduction

Infographics are composite visual artifacts that integrate data visualizations with textual and illustrative elements, such as pictograms, thematic icons, semantic text, and metaphorical imagery (Dur, 2014; Qin et al., 2020). By integrating data with visual narratives, infographics enhance expressive power and are widely used in journalism (Hamza, 2023), education (Traboco et al., 2022), and business analytics (Cui et al., 2019). Traditionally, creating high-quality infographics is labor-intensive and demands significant design expertise (Cui et al., 2019; Huang et al., 2024), often involving days of manual iteration (Huang et al., 2018).

Recently, advances in text-to-image (T2I) models, such as Nanobanana-Pro (Google, 2025) and GPT-Image (OpenAI Team, 2025), have enabled the generation of aesthetically appealing images with complex graphics and accurate text rendering. Leveraging these advanced T2I models for automated infographic generation has become a promising direction (Xiao et al., 2023; Peng et al., 2025; Google, 2025). However, T2I models suffer from inherent uncertainty (Franchi et al., 2025), raising doubts about their reliability in generating structured images such as infographics. Generated charts may appear good at first glance, but often contain easily overlooked issues that can mislead users (Zhuo et al., 2025; Zhu et al., 2025b), such as distorted data encoding (e.g., incorrect bar heights) or textual errors. Such shortcomings highlight the urgent need for systematic evaluation to identify potential issues in generated infographics.

However, to the best of our knowledge, no existing benchmark is specifically designed for this task. First, existing infographic-related datasets focus on question answering or visual reasoning (Xie et al., 2025a; Mathew et al., 2021), rather than on infographic generation. Second, there is currently no established method for evaluating the quality of infographics, and evaluation methods from related tasks do not transfer well to this setting. Prior work on text-to-image generation (Yu et al., 2022; Saharia et al., 2022; Huang et al., 2023) mainly evaluates prompt adherence for natural images. However, infographics require both semantic consistency between the prompt and visual elements, and accurate encoding of the underlying data values into corresponding visual repre-

*Xiuqi Huang, Yingchaojie Feng, and Yuyu Luo are corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36919–36954
July 2-7, 2026 ©2026 Association for Computational Linguistics

sentations. Moreover, existing evaluation methods for plain chart generation often rely on holistic scoring using multimodal large language models (MLLMs) (Xie et al., 2025b; Yang et al., 2024b). These methods offer limited interpretability and are unable to identify specific errors in the charts.

To address this gap, we introduce IGENBENCH, the first benchmark designed to evaluate the reliability of T2I models for text-to-infographic generation. As shown in Figure 1, IGENBENCH comprises 600 curated test cases spanning 30 distinct infographic types across 6 categories. We construct the dataset through a structured pipeline. We begin by collecting 40K real-world infographics, followed by clustering, sampling, and quality filtering to obtain 600 high-quality and diverse cases. For each case, we extract its design intent and underlying data, which are then used to synthesize prompts that serve as self-contained specifications for infographic generation.

We design and implement an evaluation framework that supports automatic and interpretable assessment of generated infographics, with a focus on semantic consistency and accurate data encoding. To enable fine-grained verification of infographic content, we first define a taxonomy of 10 question types that cover key elements, including visual components such as titles and chart types, as well as data-related aspects such as data marks. Next, we decompose the reliability verification of each generated infographic into atomic, self-contained yes/no questions, guided by the defined taxonomy. These verification questions are derived from two sources: (i) constraints explicitly specified in the prompt, and (ii) expert-informed seed dimensions, including data completeness, data ordering, and data encoding. The latter extends the coverage of question types beyond what is directly stated in the prompt. Finally, we use MLLMs to verify each question against the generated infographic, producing both question-level accuracy (Q-ACC) and infographic-level accuracy (I-ACC), where I-ACC reflects whether all specified constraints are simultaneously satisfied. This question-driven evaluation framework enables interpretable assessment of generated infographics, and shows strong agreement with human judgments (see Section 5.3).

We conduct a comprehensive evaluation of 10 state-of-the-art T2I models using IGENBENCH, uncovering key limitations and trends in current infographic generation. Our experiments reveal a clear three-tier performance hierarchy. The top-tier model, Nanobanana-Pro, achieves a Q-ACC of 0.90, significantly outperforming the second-tier models, which attain Q-ACC scores between 0.55 and 0.61. The remaining models fall below a Q-ACC of 0.48. This stratification highlights fundamental capability gaps in infographic generation across current T2I models. Moreover, even the best-performing model achieves an I-ACC of only 0.49, indicating that fewer than half of its generated infographics fully satisfy all specified constraints. This underscores that current T2I models are not yet reliable for autonomous infographic generation, and that human verification and post-editing remain necessary to ensure correctness. Additionally, we identify data-related dimensions as a universal bottleneck across all models—dimensions such as Data Completeness and Data Encoding emerge as the most challenging aspects. We summarize our contributions as follows:

- We present IGENBENCH, the first comprehensive benchmark for evaluating infographic generation fidelity, comprising 600 curated test cases spanning 30 infographic types and covering diverse real-world generation scenarios.
- We propose a systematic evaluation framework based on a taxonomy of 10 question types, which enables fine-grained assessment of infographic correctness using atomic yes/no verification questions. It supports both question-level accuracy (Q-ACC) and infographic-level accuracy (I-ACC) as evaluation metrics.
- We conduct extensive experiments on 10 state-of-the-art T2I models, revealing a three-tier performance hierarchy, the challenge of achieving end-to-end correctness, and data-related aspects as a universal bottleneck.

2 Related Work

Infographics Generation. Infographic creation traditionally requires professional designers and substantial manual effort. One direction for automation is the text-code-chart paradigm, where large language models generate visualization code (i.e., D3.js) that is then executed to produce infographics (Li et al., 2025, 2026). However, these code-based approaches require extensive manual template creation and are tightly coupled to their underlying asset libraries, limiting scalability and generalizability. They also struggle to freely render visual artifacts commonly found in infographics,

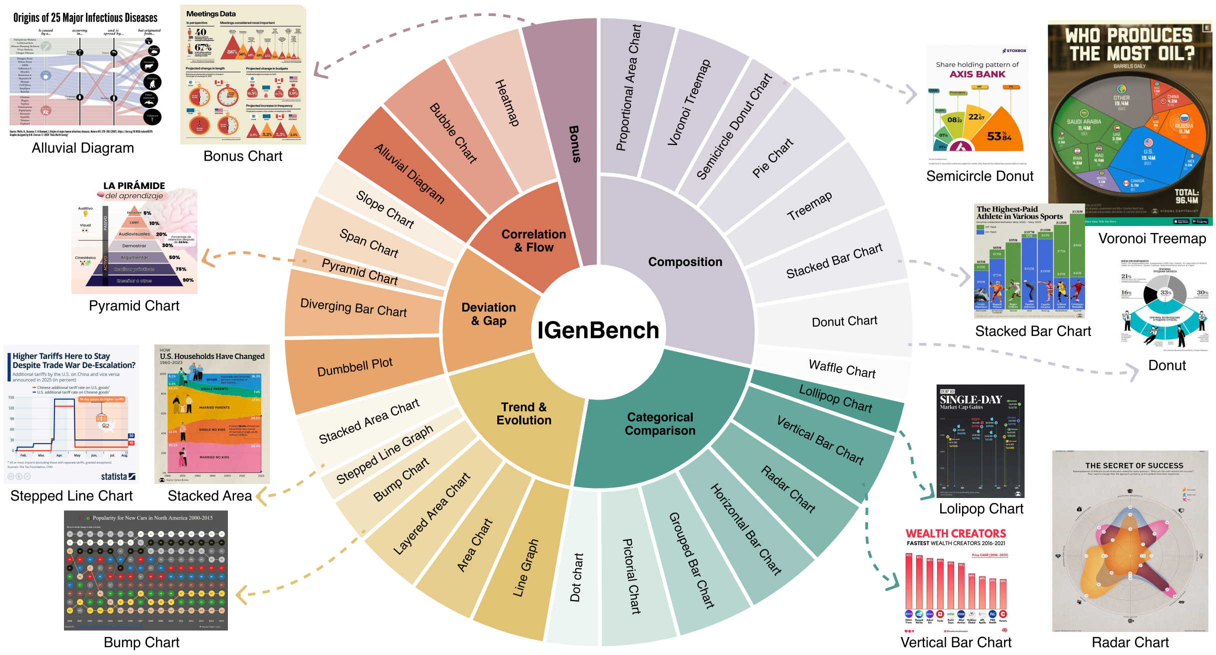


Figure 1: IGenBench overview. The benchmark covers 30 infographic types organized into 6 high-level categories: Composition, Trend & Evolution, Categorical Comparison, Deviation & Gap, Correlation & Flow, and Bonus.

such as pictograms, thematic icons, and metaphorical imagery. In contrast, text-to-image (T2I) approaches have emerged as a more promising and mainstream paradigm (Feng et al., 2023). Recent works (Xiao et al., 2023; Deo et al.; Dibia, 2023) explore embedding semantic context into infographics for better aesthetics, while BizGen (Peng et al., 2025) advances text-rich infographics generation using layout-guided cross-attention mechanisms. With the rapid advancement of increasingly powerful T2I models like Nanobanana-Pro (Google, 2025), public interest in using them for infographic creation has surged. Despite this momentum, the field lacks a dedicated benchmark that systematically evaluates T2I-generated infographics.

Benchmarks. Standardized benchmarks for infographics are lacking. Existing evaluation efforts largely focus on related but distinct tasks (Wu et al., 2026). For general text-to-image generation, benchmarks such as PartiPrompt (Yu et al., 2022), DrawBench (Saharia et al., 2022), TIFA (Hu et al., 2023), T2I-CompBench (Huang et al., 2023, 2025), MJHQ-30K (Li et al., 2024), and ArtiMuse (Cao et al., 2025) primarily assess prompt-following ability or visual aesthetics, often relying on vision-language models (e.g., CLIP (Radford et al., 2021) or modern multimodal LLMs (Bai et al., 2025)) as automatic evaluators. For chart and scientific visualization generation (Luo et al., 2018; Xie et al., 2024), benchmarks such as VisJudge-Bench (Xie et al., 2025b), VIS-Shepherd (Pan et al., 2025), Vividoc (Tang et al., 2026a,b) and MatplotBench (Yang et al., 2024b) evaluate data visulization by using MLLMs to perform holistic scoring. VISEval (Chen et al., 2024) introduces rule-based checks to assess the legality of generated charts. StructBench (Zhuo et al., 2025) and ChartMark (Chen et al., 2025) further study the generation and editing of structured scientific images. For chart understanding, benchmarks such as AskChart (Yang et al., 2024a) and ChartInsights (Wu et al., 2024) evaluate models' ability to interpret chart information. These benchmarks do not capture the unique challenges of infographic generation, which requires jointly evaluating semantic alignment and data encoding correctness. To address this gap, we propose IGenBench, a comprehensive benchmark that enables interpretable, fine-grained assessment of text-to-infographic generation reliability.

3 Dataset Construction

As shown in Stages 1 and 2 of Figure 2, we construct the dataset through Collection & Curation and Prompt Generation, which transform real-world infographic designs into infographic generation prompts. Our dataset follows two goals: (i) reflecting authentic infographic creation needs; and (ii) maximizing semantic and stylistic diversity.

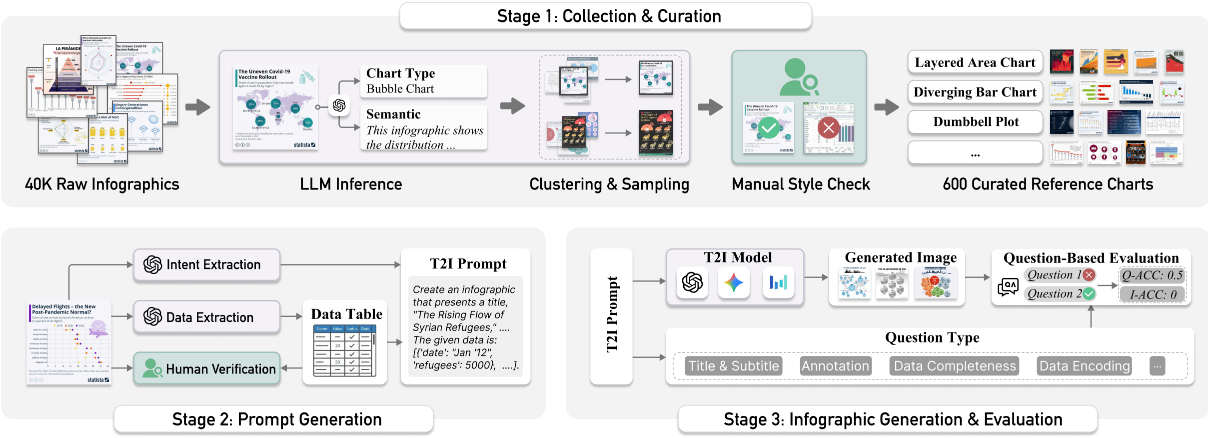


Figure 2: A three-stage pipeline for constructing IGENBENCH. Stage 1 collects and curates real-world infographics. Stage 2 generates self-contained T2I prompts via human-in-the-loop intent and data extraction. Stage 3 evaluates generated infographics by decomposing prompts into atomic yes/no verification questions assessed by MLLMs.

3.1 Infographic Collection & Curation

Real-world Infographic Collection. We begin by collecting a large pool of high-quality infographics from two mainstream visualization platforms, Statista[1] and Visual Capitalist[2], as well as the real-world portion of the ChartGalaxy dataset (Li et al., 2025). For ChartGalaxy, we exclude synthetic examples to ensure that IGENBENCH focuses on real-world usage scenarios. In total, we have collected 42,315 infographic charts from these sources.

Infographics Taxonomy. Following prior work on chart categorization (Li et al., 2025) and well-established visualization taxonomies such as the Data Viz Project (Ferdio, 2024), we construct a fine-grained yet interpretable taxonomy. We refine existing taxonomies by retaining only clearly defined types and merging visually similar variants. We additionally identify multi-panel layouts as a separate *bonus* category due to their unique generation challenges. This process results in a taxonomy of six high-level categories—*Composition*, *Trend & Evolution*, *Categorical Comparison*, *Deviation & Gap*, *Correlation & Flow*, and *Bonus*—comprising 30 types in total. Detailed taxonomy construction can be found in Appendix C.3.

Clustering & Sampling. Given the curated taxonomy, we first use a multimodal LLM (MLLM) to assign each infographic to its appropriate chart type and its high-level semantic description. To prevent any single semantic pattern from dominating the dataset, we perform intra-type deduplication to remove charts with substantial semantic overlap. Specifically, for each type, we apply a clustering approach to capture the distinct semantic subspaces. Within each cluster, we implement a stratified sampling procedure that filters out redundant samples based on cosine similarity to the cluster center embeddings. Finally, we manually check each sample and remove both low-quality and visually repetitive samples to further improve the quality of the dataset. Details can be found in Appendix C.

3.2 Human-in-the-Loop Prompt Generation

After obtaining a diverse set of reference infographics, we construct generation prompts through a two-step human-in-the-loop pipeline.

Intent and Data Extraction. We employ an MLLM to separately extract two critical components from each infographic: (i) a structural design description that captures the layout, chart type, data encoding, text placement, and decorative elements without referencing aesthetic details such as colors, fonts, or watermarks; and (ii) the underlying data table that encodes the numerical or categorical values represented in the visualization. For the design description, we prompt the MLLM to generate a single self-contained paragraph beginning with "Create an infographic that..." that captures structural visual elements. For data extraction, we instruct the model to output a structured table format. Both extraction outputs undergo manual verification to ensure accuracy and completeness, correcting hallucinations or omissions introduced by the MLLM.

Prompt Synthesis. We fuse the verified design description and data table into a single self-contained T2I prompt. The final prompt embeds the data table

[1] https://www.statista.com/
[2] https://www.visualcapitalist.com/

directly within the design description, ending with the sentence "The given data is: {data}." This format ensures that all necessary information—both structural intent and underlying data—is explicitly provided to the T2I model in a unified specification.

4 Evaluation Protocol

As shown in Stage 3 of Figure 2, we evaluate generated infographics through atomic yes/no verification questions. We then compute two complementary metrics: question-level accuracy (Q-ACC) and infographic-level accuracy (I-ACC).

4.1 Question Set Construction

Question Taxonomy. To systematically capture all critical aspects of infographic fidelity, we establish a taxonomy of 10 question types through expert consensus. Three visualization experts independently examined 300 randomly sampled infographic cases from our dataset, iteratively discussing and refining the categorization until reaching full agreement. The resulting taxonomy covers: (i) *Title & Subtitle*, verifying headings and title area text; (ii) *Chart/Diagram Type*, identifying the visualization form; (iii) *Decorative / Non-data Elements*, verifying icons and illustrations; (iv) *Annotations & Callouts*, checking numeric labels and explanatory text; (v) *Axes & Scales*, examining axis labels and tick marks; (vi) *Legend & Category Mapping*, validating color/shape/symbol keys; (vii) *Data Marks*, checking primary visual objects representing data; (viii) *Data Completeness*, ensuring all required data items appear; (ix) *Data Ordering*, verifying visual sequences match intended sorting logic; and (x) *Data Encoding*, verifying mappings from data to visual properties. The full definitions and examples can be found in Table 2.

Prompt Decomposition. Given an input prompt p, we first extract a set of prompt-derived verification questions $\mathcal{Q}_p(p) = \{q_1, \ldots, q_m\}$ by decomposing the prompt into atomic constraints. Specifically, we split the prompt into individual sentences, and then use an LLM to convert each sentence into a self-contained yes/no verification question. Each q_i is designed to be answerable solely by inspecting the generated infographic, without relying on external knowledge or implicit assumptions. Each question targets a specific visual or textual element based on our question taxonomy. For example, given a prompt requesting "an infographic with title 'Online dominiert den Versandhandel' and a horizontal grouped bar chart," we generate separate questions: one verifying the title presence (*Title & Subtitle*), and another confirming the chart type (*Chart/Diagram Type*). The full prompt used for question generation is provided in Appendix F.

Expert-Informed Augmentation. Beyond explicit prompt specifications, we augment the question set with expert-informed verification questions $\mathcal{Q}_e(p) = \{q'_1, \ldots, q'_n\}$ grounded in visualization best practices. These questions instantiate three critical requirements that may be implicit in prompts: *Data Completeness* (whether all required data items appear correctly), *Data Ordering* (whether the visual ordering follows the specified ranking), and *Data Encoding* (whether visual sizes and proportions accurately reflect the underlying data magnitudes). For each chart in the infographic, we instantiate these seed requirements into concrete, chart-specific yes/no questions following our question taxonomy. The final question set is the union: $\mathcal{Q}(p) = \mathcal{Q}_p(p) \cup \mathcal{Q}_e(p)$.

Verification. As illustrated in Figure 2 (Stage 3), given a generated infographic I, each question $q_i \in \mathcal{Q}(p)$ is evaluated with a strict binary correctness function:

$$\mathbb{I}(I, q_i) = \begin{cases} 1, & \text{if } q_i \text{ is clearly satisfied in } I, \\ 0, & \text{otherwise.} \end{cases}$$

Any ambiguity, partial satisfaction, or missing visual evidence yields a score of 0.

4.2 Reliability Metrics

Based on these question-level judgments, we report reliability at two aggregation levels.

Question-level Accuracy (Q-ACC). This metric measures the fraction of correctly satisfied verification questions across all evaluated infographics:

$$\text{Q-ACC} = \frac{1}{|\mathcal{Q}|} \sum_{q_i \in \mathcal{Q}} \mathbb{I}(I, q_i),$$

where $\mathcal{Q}$ denotes the union of all verification questions over the evaluated set.

Infographic-level Accuracy (I-ACC). This metric captures holistic correctness by measuring the fraction of infographics for which all associated verification questions are satisfied:

$$\text{I-ACC} = \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} \mathbb{I} \left(\sum_{q_i \in \mathcal{Q}(I)} \mathbb{I}(I, q_i) = |\mathcal{Q}(I)| \right),$$

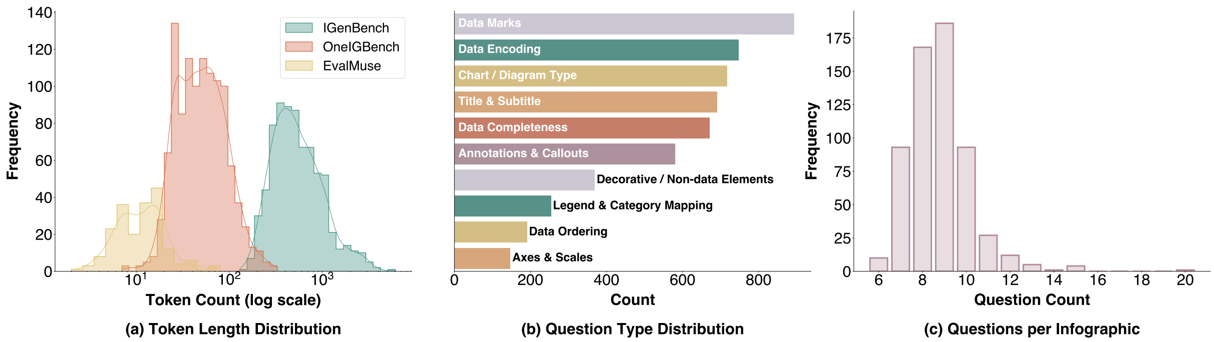


Figure 3: Statistical analysis of IGENBENCH. (a) Distribution of prompt token lengths (log scale), compared with typical T2I benchmarks. (b) Distribution of the 10 question types across the benchmark. (c) Distribution of the number of verification questions per infographic.

where $\mathcal{Q}(I)$ denotes the question set associated with infographic I.

Reporting both Q-ACC and I-ACC disentangles partial correctness from complete infographic correctness: high Q-ACC indicates that a model satisfies many individual constraints, whereas high I-ACC reflects whether the model can produce an entirely correct infographic. This distinction is critical for real-world usage, where a single error may invalidate the visualization as a whole.

4.3 Benchmark Statistics

Figure 3 presents key statistics of IGENBENCH. As shown in Figure 3(a), the prompt token lengths of IGENBENCH range from tens to several thousand tokens (log scale), which is 1–2 orders of magnitude longer than typical text-to-image prompts (OneIGBench (Chang et al., 2025) and EvalMuse(Han et al., 2024)). This reflects the inherent complexity of infographic generation. Figure 3(b) illustrates the distribution of our 10 question types across the benchmark. Figure 3(c) shows that most infographics are evaluated with 7–11 verification questions per case. Overall, IGENBENCH contains 600 test cases, including 600 curated prompts and 5259 verification questions.

5 Experiments

This section reports the main experimental results of IGENBENCH, including overall model performance, alignment with human evaluation, the selection of the evaluation model, and a case study. Extended experiments, including performance breakdown by chart type, data leakage analysis, and evaluator error analysis, are presented in Appendix D.

5.1 Experiment Setup

We evaluated 10 leading text-to-image models on IGENBENCH. Our evaluation includes 4 prominent open-source models: Qwen-Image (Wu et al., 2025), HiDream-I1 (Cai et al., 2025), FLUX.1-dev (Black Forest Labs, 2024) and Z-Image-Turbo (Team et al., 2025), as well as 6 leading closed-source models: Seedream 4.5 (ByteDance, 2025), Nanobanana (Google, 2025), Nanobanana-Pro (Google, 2025), GPT-Image-1.5 (OpenAI Team, 2025), Image-01 (MinimaxI, 2025), and P-Image (Pruna AI, 2025). For the main evaluation, we employ Gemini-2.5-Pro (Comanici et al., 2025) as the verification model to assess the generated infographics against corresponding questions.

5.2 Main Results

Tiered performance reveals fundamental capability gaps. As shown in Table 1, the experimental results reveal a clear three-tier performance hierarchy among the evaluated T2I models for infographic generation. The top-tier model, Nanobanana-Pro, achieves a Q-ACC of 0.90, significantly outperforming all other models. The second tier, consisting of Seedream-4.5 and GPT-Image-1.5, demonstrates moderate performance with Q-ACC scores of 0.61 and 0.55, respectively. The remaining models fall into a third tier with Q-ACC below 0.5, indicating fundamental difficulties in generating accurate infographics. This stratification is consistent across nearly all evaluation dimensions. For instance, in the Annotations & Callouts dimension, the performance gap between the top model (0.93) and the average (0.40) exceeds 0.50 points. Moreover, the average Q-ACC across all models is only 0.39, highlighting the difficulty of reliable infographic generation.

Table 1: Benchmark results of 10 state-of-the-art T2I models on IGenBench. Columns show per-category Q-ACC for each of the 10 question types (sorted by average difficulty from left to right), along with overall Q-ACC and I-ACC. Darker shading indicates higher accuracy.

Model	Question Type (sorted by average)										Overall	
	Comp.	Enc.	Order	Marks	Anno.	Axes	Leg.	Chart	Title	Deco.	Q-ACC↑	I-ACC↑
Nanobanana-Pro	**0.84**	**0.86**	**0.90**	**0.87**	**0.93**	**0.93**	**0.96**	**0.92**	**0.98**	**0.94**	**0.90**	0.49
Seedream-4.5	0.34	0.37	0.47	0.48	0.70	0.70	**0.81**	0.68	**0.95**	**0.84**	0.61	0.06
GPT-Image-1.5	0.38	0.48	0.44	0.57	0.50	0.54	0.57	0.68	0.60	0.80	0.55	0.12
Nanobanana	0.18	0.31	0.27	0.44	0.54	0.57	0.52	0.60	0.65	**0.81**	0.48	0.02
Qwen-Image	0.10	0.13	0.19	0.29	0.43	0.37	0.51	0.48	0.56	0.78	0.36	0.01
Z-Image-Turbo	0.10	0.16	0.16	0.25	0.38	0.31	0.58	0.42	0.61	0.73	0.35	0.00
P-Image	0.08	0.15	0.19	0.27	0.36	0.28	0.54	0.43	0.58	0.68	0.34	0.00
Image-01	0.01	0.05	0.04	0.10	0.10	0.14	0.03	0.22	0.14	0.47	0.13	0.00
HIDream-I1	0.01	0.03	0.03	0.10	0.07	0.14	0.10	0.26	0.19	0.20	0.11	0.00
FLUX.1-dev	0.00	0.03	0.01	0.08	0.06	0.06	0.01	0.24	0.09	0.39	0.10	0.00
Average	**0.21**	**0.26**	**0.27**	**0.35**	**0.40**	**0.40**	**0.46**	**0.49**	**0.54**	**0.66**	**0.39**	**0.07**

Data fidelity remains the primary bottleneck. As shown in Table 1, data-related evaluation dimensions consistently emerge as the most challenging aspects of infographic generation across all models. Data Completeness shows the lowest average performance at 0.21, followed by Ordering (0.27) and Data Encoding (0.26). Even the best-performing model, Nanobanana-Pro, achieves only 0.84 for Data Completeness and 0.86 for Data Encoding, leaving room for improvement.

This pattern reveals a fundamental limitation of current T2I models: while they excel at generating aesthetically pleasing visual layouts, chart types, titles, and decorative elements (average scores of 0.49, 0.54, and 0.46, respectively), they struggle with the precise rendering and faithful encoding of underlying data values—a capability that remains underdeveloped in current T2I models optimized primarily for natural image generation.

High Q-ACC does not imply reliable infographics. As shown in Table 1, there exists a dramatic gap between Q-ACC and I-ACC across all models, with I-ACC consistently much lower. The best-performing model achieves a Q-ACC of 0.90 but only 0.49 I-ACC. This difference is larger for second- and third-tier models: Seedream-4.5 and GPT-Image-1.5 drop from 0.61 and 0.55 Q-ACC to only 0.06 and 0.12 I-ACC, respectively, while most other models have I-ACC scores near or at zero.

The low I-ACC scores indicate that current T2I models show a "long-tail" failure mode: although they may correctly generate many aspects of an infographic, they often fail on at least one or two critical dimensions, which prevents end-to-end correctness. This suggests that for real-world use, particularly in high-stakes domains such as business analytics or education, where information accuracy is essential, current T2I models cannot yet be trusted to autonomously generate reliable infographics, even when their component-level metrics appear promising.

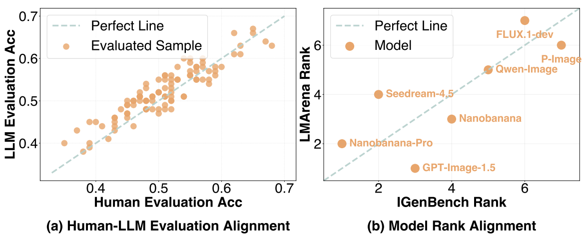
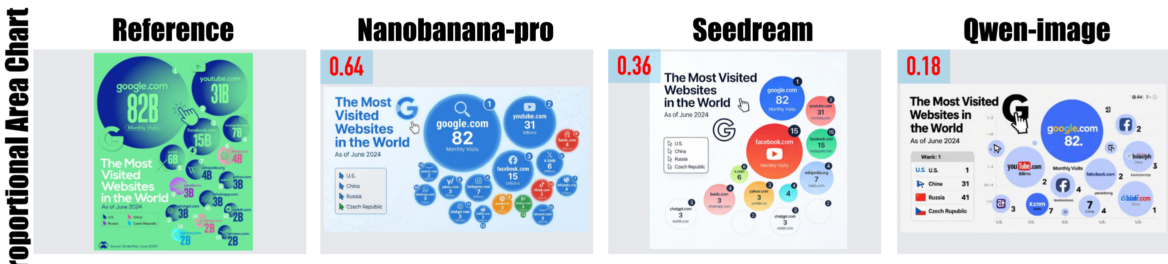


Figure 4: (a) Correlation between automatic evaluation (Gemini-2.5-Pro) and human judgments across 100 bootstrap samples of 25 questions each, showing strong Pearson correlation ($r = 0.90$). (b) Comparison of model rankings between IGenBench and LMArena, with Spearman $\rho = 0.78$.

5.3 Alignment with Human Evaluation and LMArena

Automatic evaluation aligns strongly with human judgment. Following prior work (Yang et al., 2024b), we assess the reliability of automatic evaluation by measuring its correlation with human judgments. Specifically, we use Gemini-2.5-Pro as the evaluator and recruit expert annotators to assess the same set of generated infographics. For each T2I model, we iteratively sample subsets of 25 examples from its generated outputs and compute av-



PROMPT: Create an infographic that features a title on the left reading 'The Most Visited Websites in the World' with a subtitle 'As of June 2024'. The main visualization is a bubble chart with 15 circles of varying sizes scattered across the canvas, where the size of each circle corresponds to the data value. Each bubble contains the website's domain name, a large number representing billions of visits, and has a rank number placed adjacent to it. The largest circle for the top-ranked site also includes the text 'Monthly Visits' below the number. Each circle is accompanied by a small, simplified icon representing the associated website. A legend in a rectangular box on the left side lists 'U.S.', 'China', 'Russia', and 'Czech Republic', each preceded by a cursor-shaped icon. A large, stylized letter 'G' is placed near the main title, and a hand cursor icon is positioned in the upper central area.
The given data is: [{'rank': 1, 'website': 'google.com', 'monthly_visits_billions': 82, 'country': 'U.S.'}, ...]

Figure 5: Case of Proportional Area Chart. The prompt specifies 15 bubbles with sizes proportional to website visit counts. Nanobanana-Pro generates 16 bubbles with some color encoding errors; Seedream and Qwen-Image exhibit more severe issues including incorrect ranking order and garbled text.

erage scores from both automatic and human evaluation. This process is repeated 100 times, yielding 100 data points (Figure 4a) for each evaluation type: $A = \{a_1, \cdots, a_{100}\}$ and $H = \{h_1, \cdots, h_{100}\}$, where a_i and h_i denote the average automatic and human scores on the i-th sampled subset, respectively. We obtain a Pearson correlation of $r = 0.90$ with $p = 7.54 \times 10^{-37}$. Given that $r > 0.8$ and $p < 0.05$, we conclude that automatic evaluation scores strongly correlate with human judgments, validating the reliability of IGENBENCH.

IGenBench rankings correlate with but diverge from natural image benchmarks. To further validate the characteristics of our benchmark, we compare model rankings on IGENBENCH with those on LMArena (Chiang et al., 2024), a widely-used arena for natural image generation. We select 7 models that are evaluated in both IGENBENCH and LMArena. We calculate the Spearman correlation between the two rankings and obtain $\rho = 0.78$ with $p = 0.04$. This moderate-to-strong positive correlation suggests that models with strong performance on natural image generation often also perform well on infographic generation.

At the same time, the observed ranking differences highlight challenges that are specific to infographic generation. For example, Seedream-4.5 ranks fourth on LMArena but places second on IGENBENCH, whereas GPT-Image-1.5 ranks first on LMArena but falls to third on IGENBENCH. These differences suggest that infographic generation requires more than photorealistic rendering, including accurate data encoding, compliance with structured layouts, and correct semantic alignment between visual elements and underlying data.

5.4 Selection of Evaluation Model

To identify an appropriate evaluator for the infographic assessment task, we examine the alignment between different MLLMs and human judgments. Following the setup in Experiment 5.3, we evaluate 9 leading open-source MLLMs (Llama-4-Maverick (Meta AI, 2025), Mistral-Small-3.2-24b (Mistral AI, 2025), Qwen3-VL-8b (Yang et al., 2025), Qwen3-VL-32b, Qwen2.5-VL-72b (Bai et al., 2025), GLM-4.5v (V Team et al., 2025), Gemma-3-27b (Google, 2025), Pixtral-12b (Agrawal et al., 2024), and InternVL3-78b (Zhu et al., 2025a)) and 3 closed-source models (Gemini-2.5-Pro (Comanici et al., 2025), GPT-5-mini (OpenAI, 2025), and Grok-4.1 (xAI, 2025)) by computing Pearson correlation with human evaluations on the same set of verification questions.

As shown in Figure 6, Gemini-2.5-Pro achieves a Pearson correlation of 0.90, making it the only model surpassing the strong correlation threshold of 0.8. GPT-5-mini (0.70) and GLM-4.5v (0.75) show moderate alignment, while most open-source models exhibit substantially lower correlations, with some falling below 0.5. Based on these results, we select Gemini-2.5-Pro as our automated evaluator throughout all experiments.

5.5 Case Study

As shown in Figure 5, we present a representative case involving a proportional area chart displaying “The Most Visited Websites in the World.” The prompt specifies 15 bubbles with sizes pro-

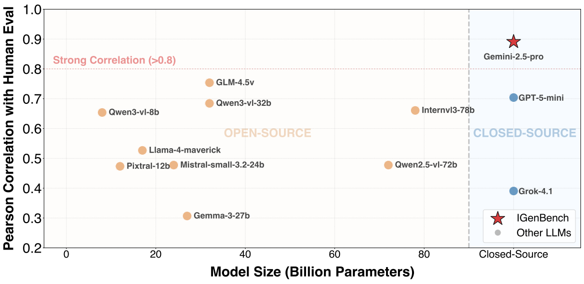


Figure 6: Pearson correlation between different MLLMs' automatic scores and human judgments. Gemini-2.5-Pro achieves the highest correlation ($r = 0.90$), the only model surpassing the strong correlation threshold of 0.8.

portional to visit counts. Nanobanana-Pro generates 16 bubbles instead of the required 15 and exhibits incorrect color encoding for certain data points. Seedream and Qwen-Image produce more severe errors, such as incorrect ranking order and garbled text in annotations. This case illustrates the difficulty of simultaneously satisfying multiple fine-grained constraints in infographic generation. Additional cases are provided in Appendix E.

6 Conclusion

We present IGENBENCH, the first benchmark for evaluating the reliability of text-to-infographic generation. Through 600 curated test cases spanning 30 infographic types and a question-driven evaluation framework, we assess current T2I models' capability to generate reliable infographics. Our evaluation of 10 state-of-the-art models reveals critical limitations of current T2I models and key insights for future model development.

7 Limitation

This work focuses on evaluating the reliability of text-to-infographic generation, with an emphasis on semantic consistency between the prompt and visual elements, as well as accurate encoding of the underlying data values into corresponding visual representations. As a result, IGENBENCH does not assess broader evaluation dimensions such as communicative effectiveness (whether the infographic successfully conveys its intended message), accessibility (e.g., colorblind-friendliness), or visual aesthetics (e.g., layout organization and stylistic creativity). We view reliability as a foundational prerequisite for these dimensions: accessibility presupposes correct visual encoding, and communicative effectiveness presupposes faithful data representation. Moreover, these dimensions require fundamentally different evaluation methodologies, such as user studies for communicative effectiveness and perceptual modeling for accessibility, which fall outside the scope of this work. We leave the systematic assessment of these complementary dimensions to future work. In addition, due to the high monetary cost of large-scale evaluation, we only include a selected set of representative state-of-the-art models. We view IGENBENCH as a living benchmark and plan to continuously incorporate more models.

8 Ethical Considerations

All infographic images used in this work are publicly available and obtained from open-access sources. To address potential copyright issues, we will not redistribute the original images and will only release their URLs as part of the dataset. We manually reviewed all collected images to verify that they do not contain harmful, illegal, or sensitive content. Some infographics include references to humans. These cases are limited to public figures appearing in widely distributed media content, such as news articles or public reports. The dataset does not involve private individuals or personal user data, and it does not include information intended for restricted or private use. As this work focuses on benchmarking and evaluation rather than deploying or enabling new generative systems, we do not foresee significant ethical or societal risks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62132017, 62502430, 62402409) and the Zhejiang Provincial Natural Science Foundation of China (LQ26F020004).

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, and et al. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Black Forest Labs. 2024. Flux. `https://github.com/black-forest-labs/flux`.

ByteDance. 2025. Seedream 4.5. `https://seed.bytedance.com/en/seedream4_5`.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, and 1 others. 2025. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. 2025. ArtiMuse: Fine-Grained Image Aesthetics Assessment with Joint Scoring and Expert-Level Understanding. *Preprint*, arXiv:2507.14533.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. 2025. OneIG-Bench: Omni-dimensional Nuanced Evaluation for Image Generation. *Preprint*, arXiv:2506.07977.

Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. 2024. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*.

Yiyu Chen, Yifan Wu, Shuyu Shen, Yupeng Xie, Leixian Shen, Hui Xiong, and Yuyu Luo. 2025. Chartmark: A structured grammar for chart annotation. In *2025 IEEE Visualization and Visual Analytics (VIS)*, pages 311–315. IEEE.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Gheorghe Comanici and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Weiwei Cui, Xiaoyu Zhang, Yun Wang, He Huang, Bei Chen, Lei Fang, Haidong Zhang, Jian-Guan Lou, and Dongmei Zhang. 2019. Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE transactions on visualization and computer graphics*, 26(1):906–916.

Anurag Deo, Savita Bhat, and Shirish Karande. VisualFusion: Enhancing Blog Content with Advanced Infographic Pipeline.

Victor Dibia. 2023. LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 113–126, Toronto, Canada. Association for Computational Linguistics.

Banu Inanc Uyan Dur. 2014. Data visualization and infographics in visual communication design education at the age of information. *Journal of arts and humanities*, 3(5):39–50.

Yingchaojie Feng, Xingbo Wang, Kam Kwai Wong, Sijia Wang, Yuhong Lu, Minfeng Zhu, Baicheng Wang, and Wei Chen. 2023. Promptmagician: Interactive prompt engineering for text-to-image creation. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):295–305.

Ferdio. 2024. Data viz project. https://datavizproject.com/.

Gianni Franchi, Nacim Belkhir, Dat Nguyen Trong, Guoxuan Xia, and Andrea Pilzer. 2025. Towards understanding and quantifying uncertainty for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8062–8072.

Google. 2025. Gemma 3. https://blog.google/technology/developers/gemma-3/. Accessed: 2026-01-02.

Google. 2025. Nano-banana. https://gemini.google/overview/image-generation/.

Google. 2025. Nanobanana-pro. https://blog.google/technology/ai/nano-banana-pro/.

Afifa Hamza. 2023. The importance of using infographics in field of journalism. *Journal El-Baheth in Human and Social Sciences Volume14-Number*, page 123.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. 2024. EvalMuse-40K: A Reliable and Fine-Grained Benchmark with Comprehensive Human Annotations for Text-to-Image Generation Model Evaluation. *Preprint*, arXiv:2412.18150.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *Preprint*, arXiv:2311.16483.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2025. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Qirui Huang, Min Lu, Joel Lanir, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2024. Graphimind: Llm-centric interface for information graphics design. *arXiv preprint arXiv:2401.13245*.

Simon Huang, Lynsey J Martin, Calvin H Yeh, Alvin Chin, Heather Murray, William B Sanderson, Rohit Mohindra, Teresa M Chan, and Brent Thoma. 2018. The effect of an infographic promotion on research dissemination and readership: a randomized controlled trial. *Canadian Journal of Emergency Medicine*, 20(6):826–833.

Boyan Li, Yiran Peng, Yupeng Xie, Sirong Lu, Yizhang Zhu, Xing Mu, Xinyu Liu, and Yuyu Luo. 2026. Deepeye: A steerable self-driving data agent system. *arXiv preprint arXiv:2603.28889*.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*.

Zhen Li, Duan Li, Yukai Guo, Xinyuan Guo, Bowen Li, Lanxi Xiao, Shenyu Qiao, Jiashu Chen, Zijian Wu, Hui Zhang, and 1 others. 2025. Chartgalaxy: A dataset for infographic chart understanding and generation. *arXiv preprint arXiv:2505.18668*.

Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. 2018. Deepeye: Towards automatic data visualization. In *2018 IEEE 34th international conference on data engineering (ICDE)*, pages 101–112. IEEE.

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

Meta AI. 2025. Llama 4: Multimodal intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

MinimaxI. 2025. Image-01. https://www.minimaxi.com/news/image-01.

Mistral AI. 2025. Mistral small 3.2. https://docs.mistral.ai/models/mistral-small-3-2-25-06.

OpenAI. 2025. Gpt-5. https://openai.com/gpt-5/. Accessed: 2026-01-02.

OpenAI Team. 2025. New chatgpt images is here. https://openai.com/index/new-chatgpt-images-is-here/.

Bo Pan, Yixiao Fu, Ke Wang, Junyu Lu, Lunke Pan, Ziyang Qian, Yuhan Chen, Guoliang Wang, Yitao Zhou, Li Zheng, Yinghao Tang, Zhen Wen, Yuchen Wu, Junhua Lu, Biao Zhu, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. Vis-shepherd: Constructing critic for llm-based data visualization generation. *Preprint*, arXiv:2506.13326.

Yuyang Peng, Shishi Xiao, Keming Wu, Qisheng Liao, Bohan Chen, Kevin Lin, Danqing Huang, Ji Li, and Yuhui Yuan. 2025. BizGen: Advancing Article-level Visual Text Rendering for Infographics Generation. *Preprint*, arXiv:2503.20672.

Pruna AI. 2025. Pruna ai. https://www.pruna.ai.

Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. 2020. Making data visualization more efficient and effective: a survey. *VLDB J.*, 29(1):93–117.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Yinghao Tang, Yupeng Xie, Yingchaojie Feng, Tingfeng Lan, and Wei Chen. 2026a. Demonstrating vividoc: Generating interactive documents through human-agent collaboration. *arXiv preprint arXiv:2603.01912*.

Yinghao Tang, Yupeng Xie, Yingchaojie Feng, Tingfeng Lan, Jiale Lao, Yue Cheng, and Wei Chen. 2026b. Vividoc: Generating interactive documents through human-agent collaboration. *arXiv preprint arXiv:2603.27991*.

Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, and 2 others. 2025. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *Preprint*, arXiv:2511.22699.

Lisa Traboco, Haridha Pandian, Elena Nikiphorou, and Latika Gupta. 2022. Designing infographics: visual representations for enhancing education, communication, and scientific research. *Journal of Korean medical science*, 37(27).

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, and et al. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Yifan Wu, Yiran Peng, Yiyu Chen, Jianhao Ruan, Zijie Zhuang, Cheng Yang, Jiayi Zhang, Man Chen, Yenchi Tseng, Zhaoyang Yu, Liang Chen, Yuyao Zhai, Bang Liu, Chenglin Wu, and Yuyu Luo. 2026. Autowebworld: Synthesizing infinite verifiable web environments via finite state machines. *Preprint*, arXiv:2602.14296.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. Chartinsights: Evaluating multimodal large language models for low-level chart question answering. *arXiv preprint arXiv:2405.07001*.

xAI. 2025. Grok 4.1. `https://x.ai/news/grok-4-1`. Accessed: 2026-01-02.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. 2025. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *Preprint*, arXiv:2402.12185.

Shishi Xiao, Suizi Huang, Yue Lin, Yilin Ye, and Wei Zeng. 2023. Let the chart spark: Embedding semantic context into chart with text-to-image generative model. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):284–294.

Tianchi Xie, Minzhi Lin, Mengchen Liu, Yilin Ye, Changjian Chen, and Shixia Liu. 2025a. InfoChartQA: A Benchmark for Multimodal Question Answering on Infographic Charts. *Preprint*, arXiv:2505.19028.

Yupeng Xie, Yuyu Luo, Guoliang Li, and Nan Tang. 2024. Haichart: Human and ai paired visualization system. *Proceedings of the VLDB Endowment*, 17(11):3178–3191.

Yupeng Xie, Zhiyang Zhang, Yifan Wu, Sirong Lu, Jiayi Zhang, Zhaoyang Yu, Jinlin Wang, Sirui Hong, Bang Liu, Chenglin Wu, and Yuyu Luo. 2025b. VisJudge-Bench: Aesthetics and Quality Assessment of Visualizations. *Preprint*, arXiv:2510.22373.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2023. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and et al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xudong Yang, Yifan Wu, Yizhang Zhu, Nan Tang, and Yuyu Luo. 2024a. Askchart: Universal chart understanding through textual enhancement. *arXiv preprint arXiv:2412.19146*.

Zhiyu Yang, Zihan Zhou, Shuo Wang, Xin Cong, Xu Han, Yukun Yan, Zhenghao Liu, Zhixing Tan, Pengyuan Liu, Dong Yu, Zhiyuan Liu, Xiaodong Shi, and Maosong Sun. 2024b. MatPlotAgent: Method and evaluation for LLM-based agentic scientific data visualization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11789–11804, Bangkok, Thailand. Association for Computational Linguistics.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, and 1 others. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, and et al. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Yizhang Zhu, Liangwei Wang, Chenyu Yang, Xiaotian Lin, Boyan Li, Wei Zhou, Xinyu Liu, Zhangyang Peng, Tianqi Luo, Yu Li, and 1 others. 2025b. A survey of data agents: Emerging paradigm or overstated hype? *arXiv preprint arXiv:2510.23587*.

Le Zhuo, Songhao Han, Yuandong Pu, Boxiang Qiu, Sayak Paul, Yue Liao, Yihao Liu, Jie Shao, Xi Chen, Si Liu, and Hongsheng Li. 2025. Factuality Matters: When Image Generation and Editing Meet Structured Visuals. *Preprint*, arXiv:2510.05091.

Appendix Overview

This appendix provides supplementary materials supporting the IGENBENCH benchmark. Section A describes the use of large language models (LLMs) in this work. Section B reports human involvement in this study. Section C details the dataset construction process, including clustering and sampling, manual quality filtering, as well as the resultant infographic and question taxonomies. Section D reports additional analyses beyond the main text, including a study of potential data leakage, an error analysis of the automated evaluation, an evaluator bias investigation across multiple providers, a reference-generation visual similarity analysis, a prompt length sensitivity study, and a question type independence analysis. Section E presents representative case studies covering all 30 chart types considered in this work. Section F documents all prompts used throughout the pipeline.

A LLM Usage

We used Claude-4.5-Sonnet for English grammar polishing of the paper. During dataset construction, we used MLLMs to assist with synthesizing text-to-infographic generation prompts, as part of a human-in-the-loop process rather than as autonomous decision makers. During evaluation, we used MLLMs to automatically answer atomic yes/no verification questions against generated infographics.

B Human Participation

Human involvement in this work included expert discussions during benchmark design and targeted annotation during dataset curation and evaluation. The taxonomy of question types was developed through iterative discussions among three coauthors with domain expertise to ensure coverage of key infographic elements. During dataset construction, human annotators filtered low-quality infographics to maintain dataset quality. For human evaluation, three undergraduate students majoring in computer science were recruited locally and compensated according to local wage standards. They were asked to answer the same atomic yes/no verification questions used in the automatic evaluation. The process was guided by an instruction, which is shown in Figure 41. The study did not involve sensitive personal data, and participants were not exposed to harmful or risky content.

C Benchmark Construction Detail

C.1 Clustering & Sampling Algorithm.

As shown in Algorithm 1, we employ a stratified cluster-then-sample strategy to ensure diversity and representativeness across different chart types. For each chart type c, we first apply k-means clustering on the semantic embeddings $e_i = f(x_i)$ extracted from the infographic descriptions or visual content, partitioning samples into C clusters. Within each cluster, we select the medoid (the sample closest to the cluster centroid) to capture the most representative instance, and optionally sample additional diverse instances. This process is repeated until K samples are selected for each chart type. This approach yields a balanced and comprehensive benchmark that reflects the full spectrum of real-world infographic complexity. K and C are set to 5 and 10, respectively.

Algorithm 1: Per-Type Sampling

Input: samples $D = \{(x_i, t_i)\}$, embeddings $e_i = f(x_i)$, samples per type K, clusters C
Output: selected sample indices S

$S \leftarrow \varnothing$;
foreach *chart type* c **do**
  Run k-means with C clusters on $\{e_i \mid t_i = c\}$;
  Let clusters be $\{I_1, \dots, I_C\}$ with centroids $\{\mu_1, \dots, \mu_C\}$;
  $S_c \leftarrow \varnothing$;
  for $j = 1$ **to** C **do**
    if $|S_c| \geq K$ **then**
      break;
    end
    $i^\star \leftarrow \arg\min_{i \in I_j} \mathrm{dist}(e_i, \mu_j)$; `// select medoid`
    $S_c \leftarrow S_c \cup \{i^\star\}$;
  end
  $S \leftarrow S \cup S_c$;
end
return S;

C.2 Manual Style Check of Dataset

Following the clustering and sampling process, we conducted a rigorous manual style check to ensure the quality and suitability of selected samples for benchmark evaluation. During this review, we identified and filtered out samples that did not meet our quality standards. These problematic cases generally fall into categories such as irrelevance, poor legibility, or lack of informational content. Figure 7 illustrates six representative types of bad cases identified during manual inspection.

As seen in Figure 7(a), some images combine charts with no semantic relevance to one another, making them unsuitable as coherent infographics. Figure 7(b) shows examples where text and visual elements are pixelated and low-resolution, severely compromising legibility. Some samples, such as Figure 7(c), contain only pure chart elements without the narrative or design features that characterize infographics. Figure 7(d) depicts a screenshot of a computer application interface rather than a standalone visualization. Figure 7(e) represents template images containing layout structures but lacking actual data information. In (f), visual elements interfere with data presentation by partially obscuring labels and complicating chart readability. All samples exhibiting these issues were excluded from the final benchmark to ensure high-quality evaluation.

C.3 Infographics Taxonomy.

We aim to ensure the benchmark is structured around the major chart types commonly recognized in visualization taxonomies. Many existing efforts categorize charts into roughly ten coarse-grained families (Xu et al., 2023; Han et al., 2023; Xia et al., 2025); however, such coarse categories may mask significant stylistic variation. ChartGalaxy (Li et al., 2025) provides over seventy fine-grained real-world categories, yet several are visually similar to one another or not clearly defined in widely used visualization taxonomies such as the Data Viz Project (Ferdio, 2024). To construct a fine-grained and interpretable taxonomy, we retain only the ChartGalaxy types that appear in existing public taxonomies and merge the remaining types into the most visually similar categories. We additionally identify infographic charts that contain multi-panel (multi-chart) layouts, which are particularly difficult for both classification and subsequent generation; these are grouped separately into a *bonus* category. This process results in a fine-grained taxonomy consisting of six high-level categories with 30 types in total:

Category 1 (Composition): Pie Chart, Donut Chart, Semicircle Donut, Stacked Bar, Treemap, Voronoi Treemap, Waffle Chart, Proportional Area.

Category 2 (Categorical Comparison): Vertical Bar, Horizontal Bar, Grouped Bar, Lollipop Chart, Radar Chart, Pictorial Chart, Dot Chart.

Category 3 (Trend & Evolution): Line Graph, Stepped Line, Area Chart, Layered Area, Stacked Area, Bump Chart.

Category 4 (Deviation & Gap): Diverging Bar, Pyramid Chart, Dumbbell Plot, Slope Chart, Span Chart.

Category 5 (Correlation & Flow): Bubble Chart, Heatmap, Alluvial Diagram.

Category 6 (Bonus): Multi-panel and multi-chart layouts that combine multiple visualization types within a single infographic.

C.4 Question Taxonomy

IGENBENCH evaluates infographic generation quality through a comprehensive set of verification questions that systematically assess whether generated outputs faithfully reproduce all specified design elements and data content from the input prompts. As described in the main text, we established a taxonomy of 10 question types through expert consensus, where three visualization experts independently examined 300 randomly sampled infographic cases, iteratively refining the categorization until reaching full agreement.

Table 2 provides detailed definitions and representative examples for each question type in our taxonomy. These categories comprehensively cover both data-driven elements (e.g., *Data Marks*, *Data Encoding*) and design-oriented components (e.g., *Title & Subtitle*, *Legend & Category Mapping*, *Decorative/Non-data Elements*). Additionally, meta-level properties such as *Data Completeness* and *Ordering* ensure that generated infographics not only include correct elements but also maintain proper structural relationships. By decomposing infographic fidelity into these fine-grained categories, IGENBENCH enables precise diagnosis of model strengths and weaknesses across different aspects of infographic generation.

D Extended Experiments

D.1 Performance Breakdown on Chart Type

Figure 8 reports the Q-ACC of all models across 30 chart types. Two key observations emerge.

Model performance varies substantially across chart types, reflecting chart-specific difficulty. Across all models, Q-ACC differs markedly by chart type. Canonical visualizations with simple and widely used grammars, such as Pie Chart, Bar

Table 2: Question Types, Descriptions, and Examples

Question Type	Definition	Example
Title & Subtitle	Questions focusing on the main heading, sub-headings, or the text content of the title area.	"Does the infographic feature the title 'Years in MLS and Average Game Attendance, 2017' at the top left?"
Chart / Diagram Type	Questions identifying the specific classification, style, or overall form of the visualization.	"Is the main visual a single filled area chart showing a rising trend over time, plotted against a grid of horizontal dotted lines?'
Decorative / Non-data Elements	Questions about icons, illustrations, or artistic elements that do not directly encode data.	"Is there a cartoon robot holding money sitting on the lower data line on the right?"
Annotations & Callouts	Questions about specific numeric labels, explanatory text, or callouts not part of axes.	"Is there a separate box in the bottom right corner that presents the text 'U.S. Overall' along with the national average growth rate?"
Axes & Scales	Questions about axes, tick marks, gridlines, ranges, or scale labels.	"Is the vertical y-axis on the left labeled with the percentage values '+120%', '+80%', '+40%', '0%', and '-40%'?"
Legend & Category Mapping	Questions involving the key that explains how colors, shapes, or symbols correspond to categories.	"Is a legend indicating two series, '2011' and '2012', located above the chart area?"
Data Marks	Questions regarding the primary visual objects (e.g., bars, points, regions) that directly represent data.	"Is each of the 10 highlighted states on the map marked with a numbered circle indicating its rank from 1 to 10?'
Data Completeness	Questions verifying whether the visualization includes all expected data points, categories, or specific entities without omission or extraneous additions.	"Does the treemap chart display rectangles for exactly nine brands: Tesla, BYD, AION, SGMW, Volkswagen, BMW, Hyundai, MG, and KIA?"
Data Ordering	Questions verifying that the visual sequence of elements matches the intended sorting logic (e.g., descending, ascending) described in the design.	"Is the entire chart sorted in descending order by average attendance?"
Data Encoding	Questions about how data values are mapped to visual properties like size, color, or position.	"Are the relative areas of the polygonal cells proportional to their numerical values, such that the cells for China (814) and USA (800) are the largest and nearly equal in size, followed by the 'Other' cell (327) which appears slightly larger than the India cell (271), and the cells for Sweden (25), Spain (27), and Israel (29) are among the smallest?"

Figure 7: Representative bad cases identified during manual quality review. Examples of samples excluded from IGENBENCH: (a) semantically unrelated charts combined without coherent theme; (b) low-resolution and pixelated images compromising legibility; (c) pure chart elements lacking infographic narrative features; (d) application interface screenshots rather than standalone visualizations; (e) empty templates without actual data content; (f) visual clutter with non-essential elements obscuring data presentation.

Chart, and Line Graph, achieve higher accuracy. In contrast, charts with complex layouts or unconventional encodings, including Alluvial Diagram, Radar Chart, Voronoi Treemap, and Bump Chart, show much lower performance. For example, on Pie Chart, the average Q-ACC across models is 0.53, while on Bump Chart it is only 0.25. This variation indicates that infographic generation has different levels of difficulty across chart types, mainly due to differences in structural complexity, spatial constraints, and the precision required to correctly align data elements.

Model rankings are largely stable across chart types. To quantify ranking consistency, we compute the Spearman rank correlation between model rankings across all pairs of chart types based on Q-ACC. We observe that the average Spearman rank correlation is 0.92, indicating a high degree of consistency in relative model ordering. Nonetheless, occasional deviations are observed, with the minimum correlation dropping to approximately 0.68. These fluctuations predominantly arise in comparisons involving structurally complex or less common chart types. For example, on Pie Chart, SeedDream-4.5 and GPT-Image-1.5 rank fourth (0.79), whereas on Bump Chart, the same model drops to seventh (0.15).

D.2 Potential Data Leakage

To assess the potential impact of data leakage on our benchmark evaluation, we conducted an additional experiment using 100 recently published infographics from Visual Capitalist dated after December 2025—ensuring that all samples were created after the release dates of the evaluated models. This temporal separation guarantees that these infographics could not have been included in any model's training data. Figure 9 presents the Q-ACC comparison between our original IGENBENCH benchmark (green dots) and the recent 100 samples (orange dots). The results reveal two key findings: (1) Most models demonstrate stable performance, with the majority showing consistent Q-ACC scores across both datasets and an average change of only 0.7%. This stability validates the reliability and generalizability of IGENBENCH's evaluation framework, suggesting that our benchmark accurately reflects model capabilities rather than memorization artifacts. (2) GPT-Image-1.5 exhibits significant performance degradation, with Q-Acc dropping substantially from 0.52 to 0.29 on recent samples, indicating possible data contam-

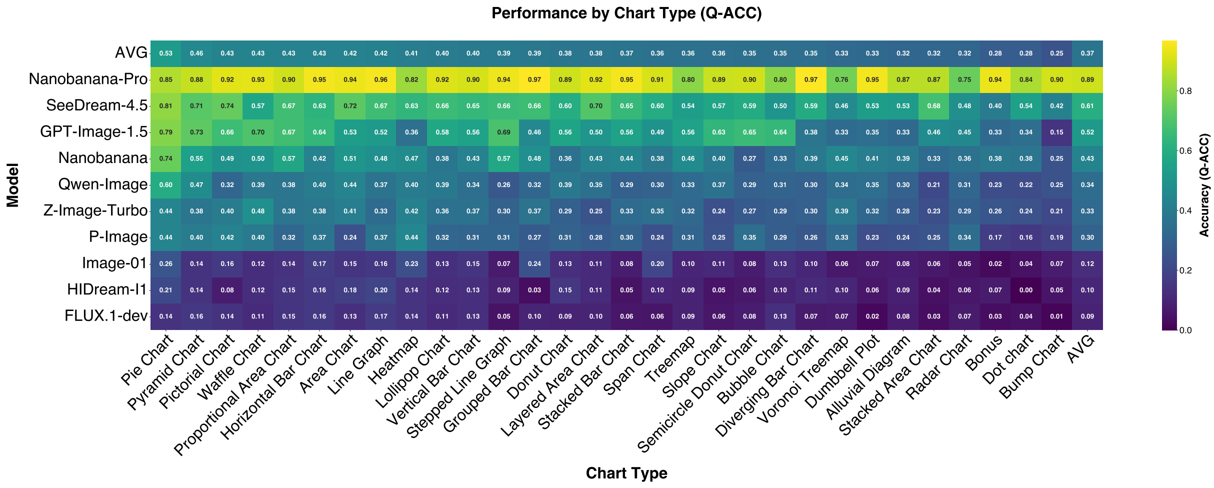


Figure 8: Performance breakdown by chart type. Q-ACC of all 10 models across 30 chart types. Common visualizations (e.g., Pie Chart, Bar Chart) achieve higher accuracy, while structurally complex types (e.g., Alluvial Diagram, Bump Chart) show much lower performance.

ination in the benchmark. We acknowledge that some degree of data leakage may exist for certain models, particularly those with more recent training cutoffs. To address this limitation, we plan to evolve IGENBENCH into a live benchmark in future work.

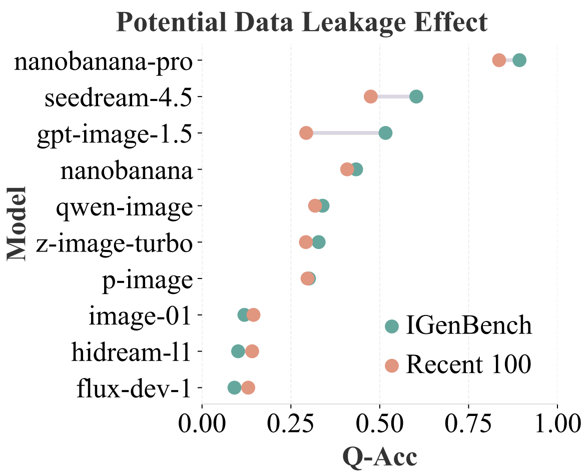


Figure 9: Potential data leakage effect. Q-ACC comparison between the original IGENBENCH benchmark (green) and 100 recently published infographics from after December 2025 (orange).

D.3 Error Analysis of Automated Evaluation

To better understand the reliability and limitations of our automated evaluation approach, we conduct a fine-grained error analysis of Gemini-2.5-Pro's judgments across different question categories. We manually examine all instances where the automated evaluator disagrees with human annotators, categorizing errors into two types: *false-positive* (where the model incorrectly marks a violation as correct) and *false-negative* (where the model incorrectly flags a correct element as wrong).

As shown in Figure 10, disagreement rates vary substantially across categories. Data Encoding exhibits the highest disagreement rate at 12.12%, primarily due to over-positive errors where the evaluator fails to detect subtle encoding violations. In contrast, categories such as Title & Subtitle (2.50%), Data Completeness (2.50%), and Ordering (0.00%) demonstrate near-perfect agreement with human judgments. These results indicate that while Gemini-2.5-Pro achieves strong overall alignment with human evaluation, certain fine-grained data encoding aspects remain challenging for automated assessment.

D.4 Evaluator Bias Investigation

Using a Google model (Gemini-2.5-Pro) as the primary evaluator could introduce bias toward Google generators. To investigate this, we selected the three evaluator MLLMs with the highest human alignment from our evaluator comparison study (Figure 6): Gemini-2.5-Pro, GLM-4.5V, and GPT-5-Mini, spanning three different providers (Google, Zhipu AI, OpenAI). We randomly sampled one item per chart type (30 items, 2,780 evaluation questions) and evaluated all 10 generation models with each evaluator. As shown in Table 3, while absolute Q-ACC values differ across evaluators, model rankings are highly stable: all pairwise Spearman $\rho \geq 0.95$. Notably, Gemini-2.5-Pro does not favor Google generators, as it assigns the lowest absolute scores for Nanobanana-Pro and Nanobanana compared to other evaluators. The high ranking consistency across evaluators from three different providers suggests that our main results are not artifacts of evaluator-specific bias.

Table 3: Q-ACC and model rankings across three evaluators from different providers. All pairwise Spearman $\rho \geq 0.95$.

Model	Gemini	GLM	GPT-5
Nanobanana-Pro	0.89 (#1)	0.91 (#1)	0.92 (#1)
Seedream-4.5	0.62 (#2)	0.73 (#2)	0.69 (#2)
GPT-Image-1.5	0.47 (#3)	0.57 (#4)	0.52 (#4)
Nanobanana	0.42 (#4)	0.63 (#3)	0.53 (#3)
Z-Image-Turbo	0.36 (#5)	0.50 (#6)	0.44 (#6)
P-Image	0.35 (#6)	0.49 (#7)	0.41 (#7)
Qwen-Image	0.33 (#7)	0.51 (#5)	0.45 (#5)
HIDream-I1	0.12 (#8)	0.19 (#8)	0.21 (#8)
FLUX.1-dev	0.12 (#9)	0.14 (#9)	0.19 (#9)
Image-01	0.11 (#10)	0.14 (#10)	0.13 (#10)

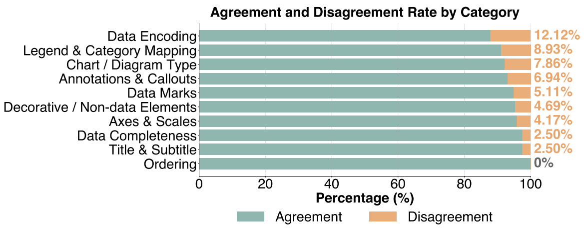


Figure 10: Agreement and disagreement rates between Gemini-2.5-Pro and human annotators across different question categories.

D.5 Reference-Generation Visual Similarity

To examine whether reference infographics carry evaluative signal, we compute four image similarity metrics between each generated infographic and its human reference across all 600 benchmark items and 10 models: CLIP Image Similarity, SSIM, PSNR, and LPIPS. As shown in Table 4, these similarity metrics show limited discriminative power: most values fall within a narrow range (e.g., CLIP Similarity spans 0.67–0.80), and models with similar scores can differ drastically in Q-ACC (e.g., Z-Image-Turbo and Nanobanana-Pro both score 0.80 in CLIP Similarity but differ by over 55 points in Q-ACC). Different metrics also produce inconsistent model rankings. We further examine whether per-item similarity predicts per-item Q-ACC by computing Spearman rank correlation across all 7,096 (item, model) pairs, as shown in Table 5. Perceptual-level metrics (CLIP, LPIPS) show moderate positive correlations with Q-ACC, while pixel-level metrics (SSIM, PSNR) are near zero. These results confirm that while reference infographics carry some evaluative signal, our QA-based evaluation captures substantially more than visual resemblance, supporting the design choice of decomposed QA evaluation over reference-based scoring.

Table 4: Model-level average similarity between generated infographics and human references. Rankings are shown in parentheses.

Model	CLIP	SSIM	PSNR	LPIPS↓
Nanobanana	0.80 (#1)	0.52 (#1)	8.45 (#2)	0.64 (#3)
Nanobanana-Pro	0.80 (#1)	0.48 (#5)	8.45 (#2)	0.61 (#1)
Z-Image-Turbo	0.80 (#1)	0.50 (#3)	7.73 (#5)	0.64 (#3)
Seedream-4.5	0.78 (#4)	0.46 (#6)	7.61 (#6)	0.66 (#5)
Qwen-Image	0.78 (#4)	0.44 (#8)	6.93 (#8)	0.67 (#6)
P-Image	0.77 (#6)	0.46 (#6)	7.16 (#7)	0.68 (#7)
GPT-Image-1.5	0.77 (#6)	0.39 (#10)	7.07 (#8)	0.67 (#6)
HIDream-I1	0.75 (#8)	0.41 (#9)	5.96 (#10)	0.70 (#10)
FLUX.1-dev	0.70 (#9)	0.49 (#4)	8.64 (#1)	0.68 (#7)
Image-01	0.67 (#10)	0.52 (#1)	8.58 (#4)	0.69 (#9)

Table 5: Spearman correlation between per-item image similarity metrics and Q-ACC across 7,096 (item, model) pairs.

Similarity Metric	Spearman ρ	p-value
CLIP Image Sim.	0.30	$< 10^{-6}$
LPIPS (↓=better)	−0.22	$< 10^{-6}$
PSNR	0.08	$< 10^{-6}$
SSIM	0.05	$< 10^{-6}$

D.6 Prompt Length Sensitivity

We analyze how prompt complexity affects model performance across all 600 benchmark items and 10 models. We measure three complexity variables for each prompt: total token count (*total_length*), token count of the embedded data portion (*data_length*), and token count of the layout/visual specification (*semantic_length*). As shown in Table 6, all three variables show significant negative correlations with Q-ACC, with *total_length* and *data_length* exhibiting the strongest effects ($\rho \approx -0.57$). This confirms that data volume dominates prompt length and is the primary source of difficulty. Models also differ substantially in their sensitivity: GPT-Image-1.5 is most sensitive ($\rho = -0.550$), while Nanobanana-Pro is most robust ($\rho = -0.198$).

Table 6: Spearman correlation between prompt complexity variables and mean Q-ACC across 600 items.

Variable	**Spearman** ρ	p**-value**
total_length	-0.573	1.4×10^{-53}
data_length	-0.557	4.3×10^{-50}
semantic_length	-0.403	7.1×10^{-25}

D.7 Question Type Independence

To verify that the 10 question types are not redundant, we compute within-model pairwise Spearman correlations on per-item accuracy across all 600 items, then average across all 10 models. Computing correlations within each model avoids inflated correlations caused by pooling models of different overall ability. As shown in Table 7, among all dimension pairs, the highest within-model correlation is only 0.21 (Data Completeness vs. Data Encoding, and Data Completeness vs. Data Marks). Most pairs fall below 0.10, and many are near zero (e.g., Decorative vs. Completeness: 0.01, Ordering vs. Title: 0.02). "–" indicates cases where some weak models have near-zero variance on that dimension, preventing correlation computation. These results confirm that the 10 question types capture largely independent aspects of infographic fidelity with minimal redundancy.

Table 7: Average within-model pairwise Spearman correlation matrix across the 10 question types. All correlations remain below 0.21. "–" indicates near-zero variance preventing computation.

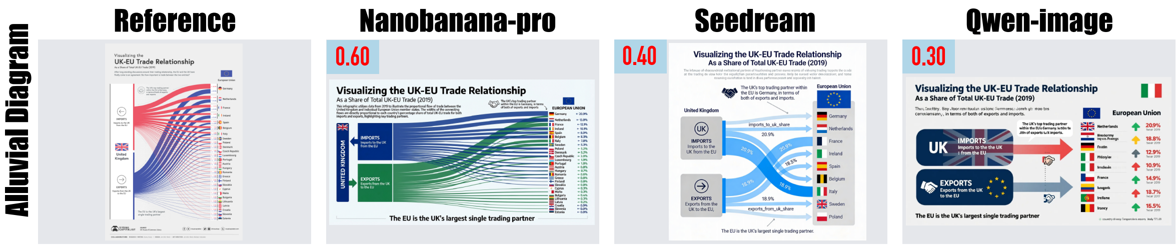
	Comp.	**Enc.**	**Order**	**Marks**	**Anno.**	**Axes**	**Leg.**	**Chart**	**Title**	**Deco.**
Comp.	1.00	.21	–	.21	.06	–	–	.15	.04	.01
Enc.	.21	1.00	.20	.16	.08	.01	.05	.10	.03	.02
Order	–	.20	1.00	–	.09	–	–	.18	.02	–
Marks	.21	.16	–	1.00	.09	.08	.06	.10	.04	.04
Anno.	.06	.08	.09	.09	1.00	.05	.06	.04	.11	.07
Axes	–	.01	–	.08	.05	1.00	–	.05	.09	.04
Leg.	–	.05	–	.06	.06	–	1.00	.07	.03	.04
Chart	.15	.10	.18	.10	.04	.05	.07	1.00	.04	.03
Title	.04	.03	.02	.04	.11	.09	.03	.04	1.00	.01
Deco.	.01	.02	–	.04	.07	.04	.04	.03	.01	1.00

E Full Case

We present a complete collection of case studies covering all chart types considered in this work. Figures 11 to 40 illustrate representative examples, spanning a total of 30 distinct chart types. Together, these cases provide a comprehensive view of the visual diversity encountered in infographics.

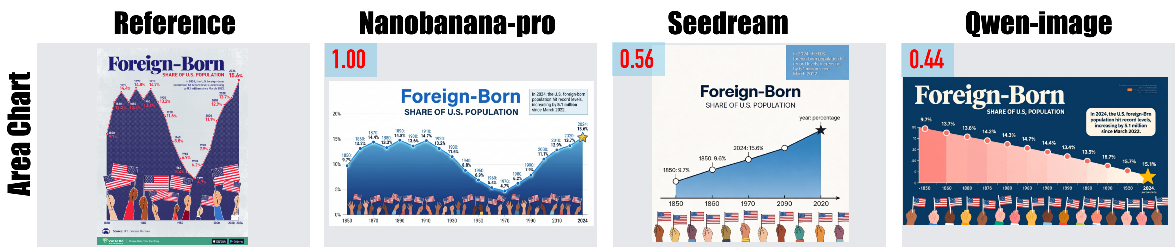
F Instruction and Prompts

In this section, we provide the instruction given to human evaluators, along with all prompts used throughout our pipeline for transparency and reproducibility. Figure 41 shows the task instruction provided to human evaluators during the evaluation phase. Figure 42 presents the prompt for chart type detection. Figure 43 shows the prompt for T2I prompt construction. Figure 44 illustrates the prompt for question generation, decomposing T2I prompts into atomic verification questions. Figure 45 displays the prompt for question type classification, categorizing generated questions into our predefined taxonomy. Figure 46 presents the prompt for question augmentation, which expands the question set based on seed questions to ensure comprehensive coverage. Finally, Figure 47 shows the prompt used for automated evaluation of generated infographics against verification questions.



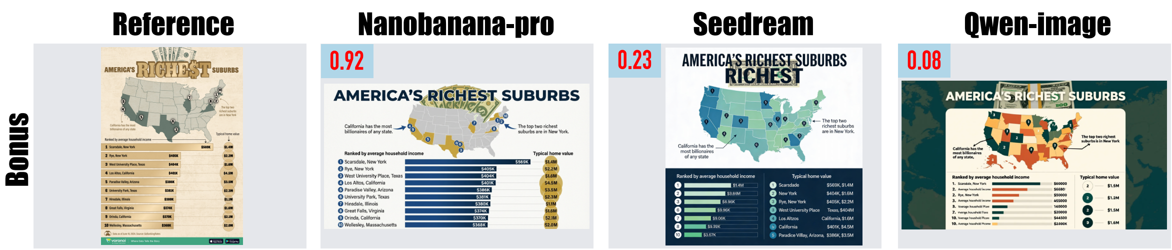
PROMPT: Create an infographic that features the title 'Visualizing the UK-EU Trade Relationship' and subtitle 'As a Share of Total UK-EU Trade (2019)' at the top, followed by a paragraph of introductory text. The central element is a flow diagram with the 'United Kingdom' on the left and the 'European Union' countries listed vertically on the right. The UK side is split into an upper section for 'IMPORTS Imports to the UK from the EU' and a lower section for 'EXPORTS Exports from the UK to the EU,' each marked with a directional arrow icon. On the right, the 'European Union' is listed at the top with its flag, followed by a list of individual countries, each with its flag and name. The width of the flows connecting the UK to each EU country represents the percentage of trade. Each country on the right has its import and export percentages listed next to it, denoted by a left-pointing arrow and a right-pointing arrow, respectively. An annotation with a handshake icon in the upper portion reads 'The UK's top trading partner within the EU is Germany, in terms of both of exports and imports,' while another at the bottom states 'The EU is the UK's largest single trading partner'.
The given data is: [{'country': 'Germany', 'imports_to_uk_share': 20.9, 'exports_from_uk_share': 18.9, 'year': 2019}, ...]

Figure 11: Case of Alluvial Diagram.



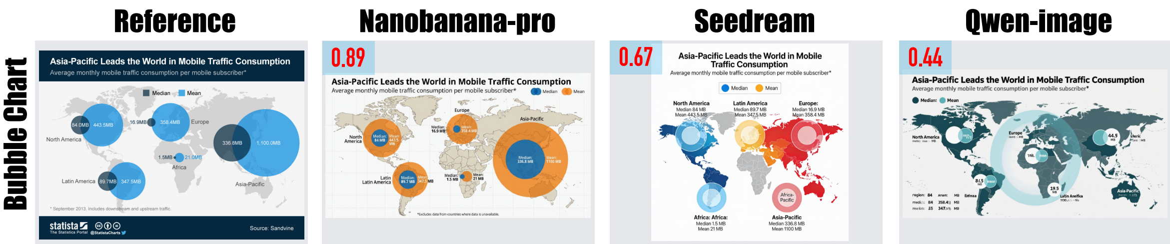
PROMPT: Create an infographic that features a large title, "Foreign-Born", at the top center, with a subtitle, "SHARE OF U.S. POPULATION", directly below it. The main visual is a large area chart that displays a single data series over time, with the x-axis running along the bottom. Data points on the chart are marked with circles, except for the final data point which is a star. Each data point is labeled with its corresponding year and percentage value placed directly above it. In the upper right corner, there is a text block that reads: "In 2024, the U.S. foreign-born population hit record levels, increasing by 5.1 million since March 2022." At the bottom of the infographic, below the x-axis, there is a row of illustrated diverse hands holding up small American flags.
The given data is: [{'year': 1850, 'percentage': 9.7}, ...]

Figure 12: Case of Area Chart.



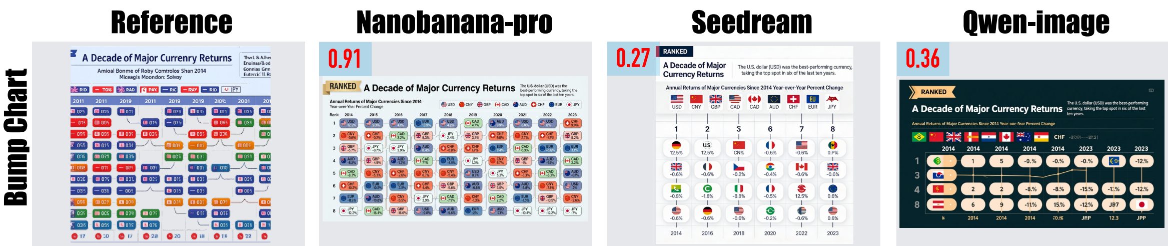
PROMPT: Create an infographic that features the title "AMERICA'S RICHEST SUBURBS" at the top, with an illustration of currency bills behind the word "RICHEST". Below the title is a map of the United States with numbered location markers on specific states, and those states are shaded differently from the rest. Two text annotations with arrows point to map regions: one says "California has the most billionaires of any state," and the other says "The top two richest suburbs are in New York." The lower half of the infographic contains a list ranked 1 through 10, preceded by the text "Ranked by average household income". Each numbered entry in the list displays the location, followed by a horizontal bar whose length represents the average household income value, which is also written on the bar. To the right of the list is a separate column of proportionally sized circles under the heading "Typical home value," with each circle aligned with a list item and displaying its corresponding value.
The given data is: [{'rank': 1, 'location': 'Scarsdale, New York', 'average_household_income': 569000, 'average_household_income_formatted': '$569K', 'typical_home_value': 1400000, 'typical_home_value_formatted': '$1.4M'}, ...]

Figure 13: Case of Bonus.



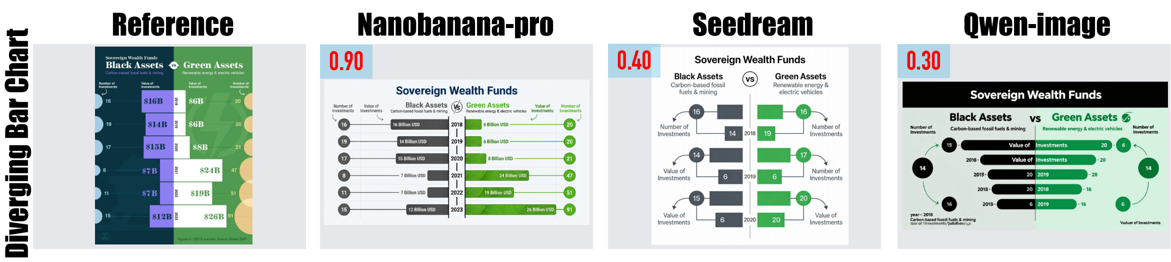
PROMPT: Create an infographic that has a title, 'Asia-Pacific Leads the World in Mobile Traffic Consumption', and a subtitle, 'Average monthly mobile traffic consumption per mobile subscriber*', positioned at the top. The main visual is a bubble chart overlaid on a world map, with five sets of data points corresponding to different regions. Each data point consists of two overlapping circles, with the size of each circle representing a value. A legend at the top indicates one circle type represents 'Median' and the other 'Mean'. Each pair of circles is located over its corresponding geographical area and is labeled with the region's name: 'North America', 'Latin America', 'Europe', 'Africa', and 'Asia-Pacific'. The numeric value and unit for each circle is displayed within or next to it.
The given data is: [{'region': 'North America', 'metric': 'Median', 'value': 84, 'unit': 'MB'}, ...]

Figure 14: Case of Bubble Chart.



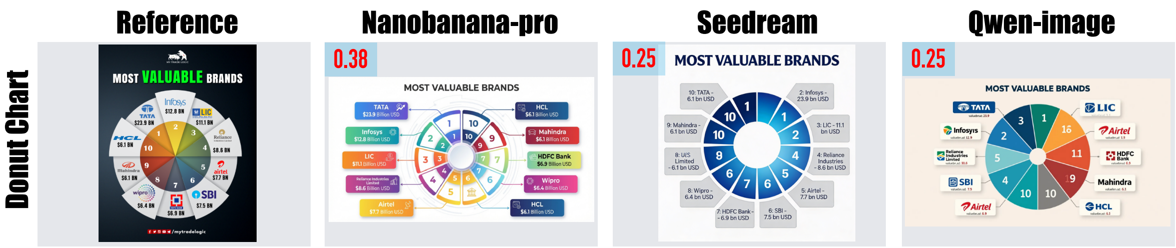
PROMPT: Create an infographic that displays a banner at the top left with the word 'RANKED' above the main title, 'A Decade of Major Currency Returns'. To the right of the main title is a text block stating, 'The U.S. dollar (USD) was the best-performing currency, taking the top spot in six of the last ten years.'. Below the title is the subtitle, 'Annual Returns of Major Currencies Since 2014 Year-over-Year Percent Change'. A horizontal legend under the subtitle lists eight currencies, each with a representative flag icon and a three-letter code: USD, CNY, GBP, CAD, AUD, CHF, EUR, and JPY. The central element is a grid chart organized into ten columns labeled by year, from 2014 to 2023, and eight rows labeled by rank, from 1 to 8. Each cell within the grid contains a rounded rectangle that shows a currency's flag icon and its numerical return percentage. Vertical lines connect the rectangles for the same currency across consecutive years, visually tracking its rank over time. The given data is: [{'year': 2014, 'rank': 1, 'currency': 'USD', 'return_percentage': 12.5},...]

Figure 15: Case of Bump Chart.



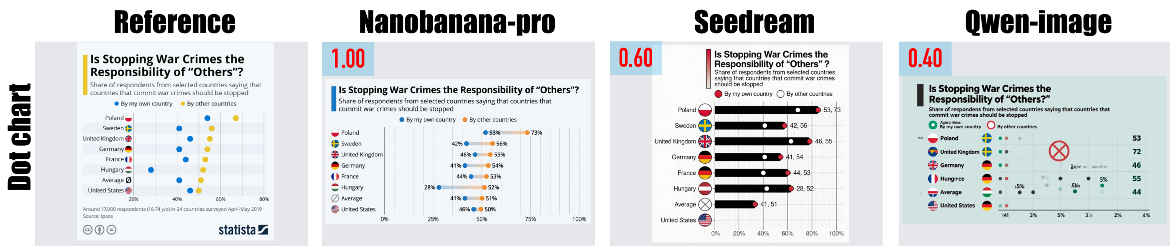
PROMPT: Create an infographic that compares two categories of investments, "Black Assets" and "Green Assets," using a mirrored horizontal bar chart layout split by a central vertical axis. The main title "Sovereign Wealth Funds" sits at the top, above the category titles "Black Assets" on the left and "Green Assets" on the right, which are separated by a "vs" icon. Subtitles "Carbon-based fossil fuels & mining" and "Renewable energy & electric vehicles" are placed under their respective category titles. Years are listed vertically along the central axis, with each year corresponding to a row of data. For each year, a bar representing "Value of Investments" extends outward from the center, with its value written inside. To the far left and far right, circles connected by a line to their corresponding bar represent the "Number of Investments," with the value placed next to the circle. Each side includes the column headers "Number of Investments" and "Value of Investments," with small curved arrows pointing from the "Number of Investments" headers to the data circles.
The given data is: [{'year': 2018, 'category': 'Black Assets', 'category_description': 'Carbon-based fossil fuels & mining', 'number_of_investments': 16, 'value_of_investments_usd_billions': 16}, ...]

Figure 16: Case of Diverging Bar Chart.



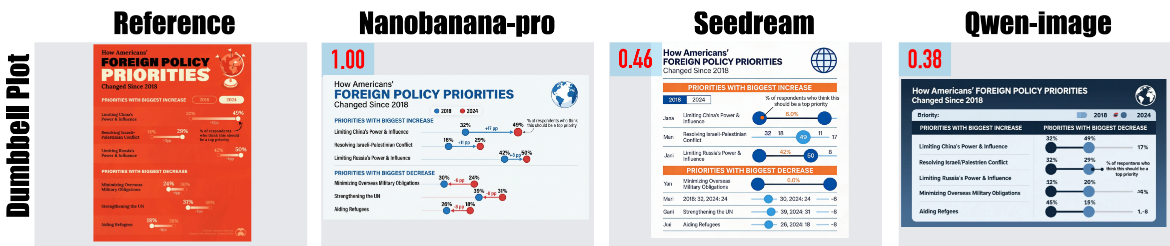
PROMPT: Create an infographic that features a title, 'MOST VALUABLE BRANDS', centered at the top. The main visual is a large circular chart divided into ten segments radiating from a central point, where the inner part of each segment contains a rank number from 1 to 10. Each segment points outwards to a corresponding label block arranged around the perimeter of the chart, with rank 1 positioned in the top-left and the rest following clockwise. These label blocks display the brand name and its associated value.the column headers "Number of Investments" and "Value of Investments," with small curved arrows pointing from the "Number of Investments" headers to the data circles.
The given data is: [{'rank': 1, 'brand': 'TATA', 'value bn usd': 23.9}, ...]

Figure 17: Case of Donut Chart.



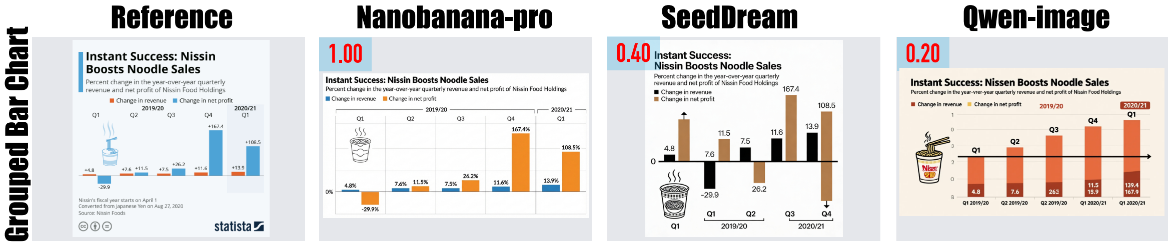
PROMPT: Create an infographic that features a title and subtitle at the top, with a vertical bar element to the left of the title. Below the subtitle, a two-item legend with circular markers indicates two data series: 'By my own country' and 'By other countries'. The main visual is a horizontal dot plot chart with a percentage-based x-axis with vertical gridlines. The y-axis lists eight categories vertically: 'Poland', 'Sweden', 'United Kingdom', 'Germany', 'France', 'Hungary', 'Average', and 'United States'. Each country name is preceded by a circular icon of its national flag, and 'Average' is preceded by a slashed circle symbol. For each category, two dots are plotted horizontally corresponding to the two data series in the legend. The title is 'Is Stopping War Crimes the Responsibility of "Others"?'. The subtitle is 'Share of respondents from selected countries saying that countries that commit war crimes
The given data is: [{'country': 'Poland', 'response': 'By my own country', 'value': 53}, ...]

Figure 18: Case of Dot Chart.



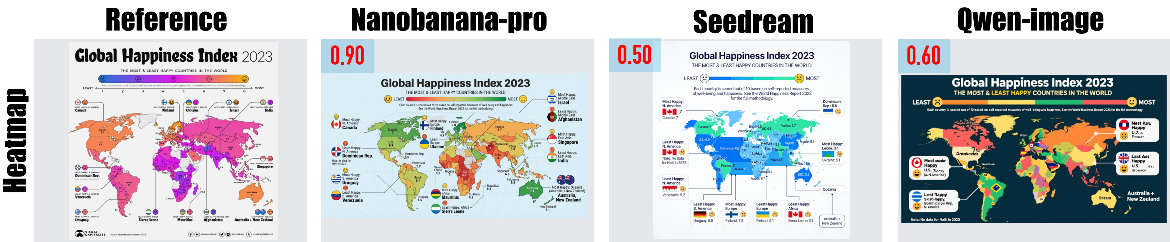
PROMPT: Create an infographic that displays the main title "How Americans' FOREIGN POLICY PRIORITIES" and a subtitle "Changed Since 2018" at the top left, with an icon of a globe located in the upper right. The layout is divided into two main horizontal sections titled "PRIORITIES WITH BIGGEST INCREASE" and "PRIORITIES WITH BIGGEST DECREASE". A legend with the labels "2018" and "2024" is positioned near the top of the content area. Each section contains three line items, with the name of the priority on the left and a horizontal dumbbell plot to the right. Each plot consists of two circles connected by a line, representing data for two years. The percentage values are placed above the circles, and the change in percentage points is labeled on the connecting line. An annotation that reads "% of respondents who think this should be a top priority" points to a data point.should be stopped'.
The given data is: [{'priority': \"Limiting China's Power & Influence\", 'trend_group': 'Priorities with biggest increase', 'percentage_2018': 32, 'percentage_2024': 49, 'change_pp': 17}, ...]

Figure 19: Case of Dumbbell Plot.



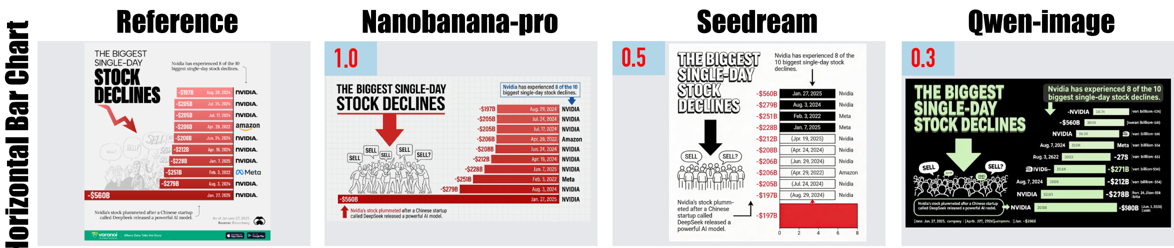
PROMPT: Create an infographic that features a title, "Instant Success: Nissin Boosts Noodle Sales", and a subtitle, "Percent change in the year-over-year quarterly revenue and net profit of Nissin Food Holdings", positioned at the top left. Below the subtitle, a legend indicates two categories: "Change in revenue" and "Change in net profit". The primary visual is a grouped vertical bar chart with a central horizontal zero axis. The chart displays data for five time periods, with labels above the bars: "Q1", "Q2", "Q3", and "Q4" are grouped under a "2019/20" heading, and a final "Q1" is under a "2020/21" heading. For each period, two vertical bars represent the categories from the legend, extending upwards for positive values and downwards for negative values. Each bar is topped or bottomed with its corresponding numerical data label. To the left, below the first "Q1" label, there is a line-drawing icon of a steaming instant noodle cup. The given data is: [{'time_period': 'Q1 2019/20', 'change_in_revenue_pct': 4.8, 'change_in_net_profit_pct': -29.9}, ...]

Figure 20: Case of Grouped Bar Chart.



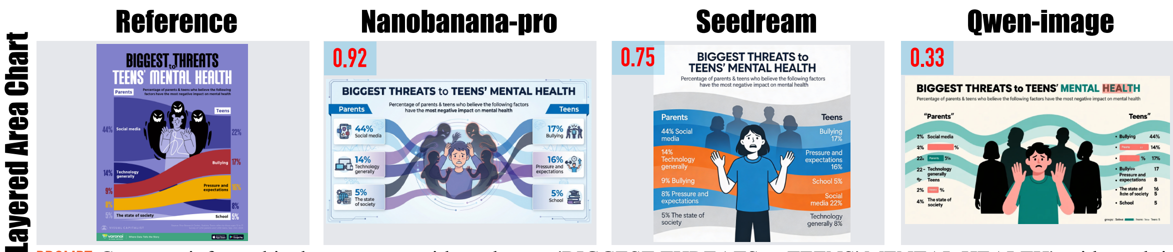
PROMPT: Create an infographic that features the title 'Global Happiness Index 2023' and subtitle 'THE MOST & LEAST HAPPY COUNTRIES IN THE WORLD' at the top. A horizontal scale below the title is labeled 'LEAST' to 'MOST' with sad and happy emoticons, accompanied by the text 'Each country is scored out of 10 based on self-reported measures of well-being and happiness. See the World Happiness Report 2023 for the full methodology.' The infographic's centerpiece is a world map where countries are displayed as a choropleth chart, with some countries labeled with their names and numeric scores. This map is framed by callouts connected by lines to specific countries, which identify the 'Most Happy' and 'Least Happy' for various regions. The regions highlighted are N. America, S. America, Europe, Africa, Middle East, East Asia, and Oceania, with the Oceania callout labeled 'Australia + New Zealand'. Each callout consists of a circular flag icon, a circular emoticon, the category label, and the country's name. A text annotation 'Note: No data for Haiti in 2023' is positioned under the 'Least Happy: N. America' callout. [{'country': 'Canada', 'score': 7, 'category': 'Most Happy: N. America'}, ...]

Figure 21: Case of Heatmap.



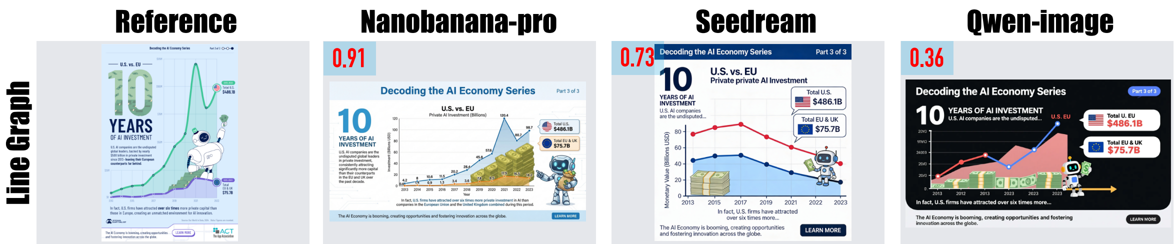
PROMPT: Create an infographic that features the title 'THE BIGGEST SINGLE-DAY STOCK DECLINES' in large, stacked text at the top left. Below the title is a large downward-pointing arrow, followed by a line drawing of a crowd with speech bubbles containing the words 'SELL' and 'SELL?'. The main visual is a horizontal bar chart showing the 10 biggest declines, with the single largest decline represented by a wide bar at the bottom and the other nine smaller bars stacked vertically above it on the right side. Each bar contains a label for the monetary value on its left side and the date in its center, with the corresponding company name positioned to the right of each bar. At the top right, there is a text block that reads 'Nvidia has experienced 8 of the 10 biggest single-day stock declines.' with an arrow pointing down towards the chart. At the bottom left, an annotation reads 'Nvidia's stock plummeted after a Chinese startup called DeepSeek released a powerful AI model.' with an arrow pointing up to the largest bar.country's name. A text annotation 'Note: No data for Haiti in 2023' is positioned under the 'Least Happy: N. America' callout. The given data is: [{'date': 'Jan. 27, 2025', 'company': 'NVIDIA', 'label': '-$560B', 'value_billions': -560}, ...]

Figure 22: Case of Horizontal Bar Chart.



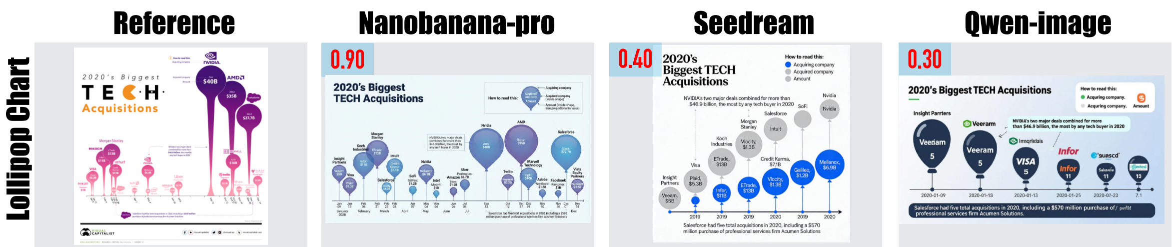
PROMPT: Create an infographic that presents a title at the top, 'BIGGEST THREATS to TEENS' MENTAL HEALTH', with a subtitle 'Percentage of parents & teens who believe the following factors have the most negative impact on mental health' below it. The infographic is divided into two sides, designated by a 'Parents' label on the upper left and a 'Teens' label on the upper right. A central illustration depicts a person with a worried expression, hands up, surrounded by four shadowy figures. Layered, wavy bands flow horizontally across the page from left to right, behind the illustration. On the 'Parents' side, a vertical list displays percentages followed by category labels: 'Social media', 'Technology generally', and 'The state of society'. On the 'Teens' side, a corresponding vertical list shows category labels followed by percentages, including 'Bullying', 'Pressure and expectations', and 'School'.
The given data is: [{'category': 'Social media', 'group': 'Parents', 'percentage': 44}, ...]

Figure 23: Case of Layered Area Chart.



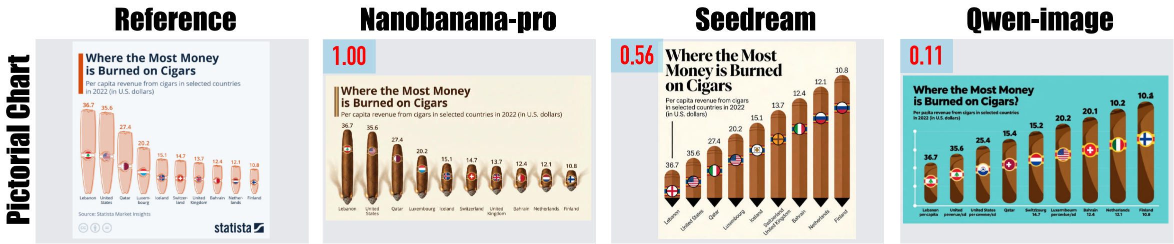
PROMPT: Create an infographic that features the title "Decoding the AI Economy Series" with "Part 3 of 3" at the top right. The main visual is a large line chart titled "U.S. vs. EU" comparing private AI investment, with a vertical axis for monetary value and a horizontal axis for years 2013 through 2023. On the left, a large number "10" is displayed above the text "YEARS OF AI INVESTMENT" and a descriptive paragraph starting "U.S. AI companies are the undisputed...". The chart displays two data lines with marked points, and the area under the lower line is illustrated as a stack of bills. To the right of the chart, two callout boxes show total investment: one labeled "Total U.S. $486.1B" with a U.S. flag icon and one below it labeled "Total EU & UK $75.7B" with an EU flag icon. A cartoon robot holding money sits on the lower data line on the right. Below the chart is a sentence starting "In fact, U.S. firms have attracted over six times more...". A footer contains the text "The AI Economy is booming, creating opportunities and fostering innovation across the globe." next to a button labeled "LEARN MORE".
The given data is: [{'year': 2013, 'region': 'U.S.', 'investment_billions': 4.2}, ...]

Figure 24: Case of Line Graph.



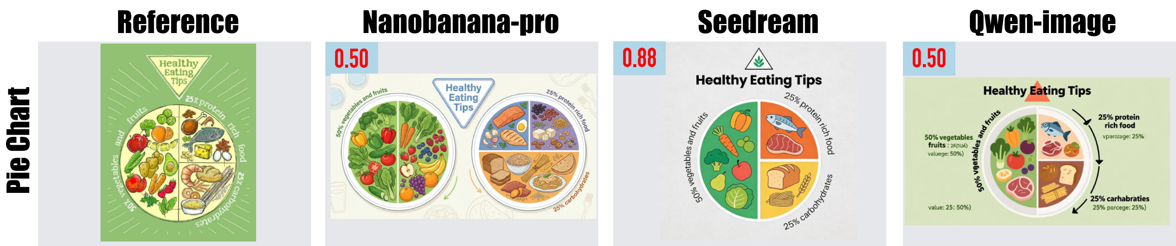
PROMPT: Create an infographic that displays information on a horizontal timeline axis positioned at the bottom. The main title, '2020's Biggest TECH Acquisitions', is located at the top left of the layout. Each data point is represented by a vertically oriented, balloon-like shape, where the size is proportional to the data value, originating from its corresponding date on the timeline. The name of the acquiring company is placed directly above its shape, while the name of the acquired company and the amount are located inside the shape. A legend titled 'How to read this:' in the upper right quadrant explains the data mapping, indicating 'Acquiring company', 'Acquired company', and 'Amount'. Two text annotations are present: one in the middle of the graphic reads 'NVIDIA's two major deals combined for more than $46.9 billion, the most by any tech buyer in 2020', and another at the bottom reads 'Salesforce had five total acquisitions in 2020, including a $570 million purchase of professional services firm Acumen Solutions'.
The given data is: [{'date': '2020-01-09', 'acquiring_company': 'Insight Partners', 'acquired_company': 'Veeam', 'amount_billions_usd': 5}, ...]

Figure 25: Case of Lollipop Chart.



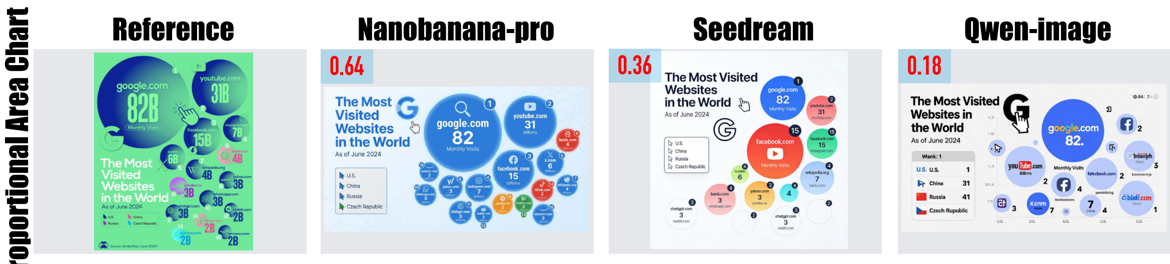
PROMPT: Create an infographic that features a large title at the top left, "Where the Most Money is Burned on Cigars," with a subtitle directly below it reading, "Per capita revenue from cigars in selected countries in 2022 (in U.S. dollars)." The main visual is a vertical bar chart composed of ten downward-pointing, cigar-shaped illustrations, arranged in descending order of height from left to right. Each cigar-shaped bar represents a country, with a circular icon of that country's flag on a band around the middle of the cigar. The numerical value is placed directly above each cigar, and the corresponding country name is placed directly below it. A vertical line element is positioned to the left of the main title. The given data is: [{'country': 'Lebanon', 'per_capita_revenue_usd': 36.7},...]

Figure 26: Case of Pictorial Chart.



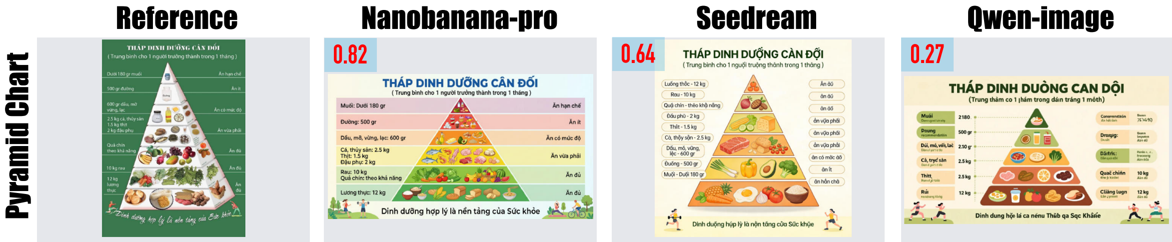
PROMPT: Create an infographic that features the title 'Healthy Eating Tips' inside a triangle at the top center. The main visual is a large circle representing a plate, which is divided into three sections. The left half of the circle is one large section, while the right half is split horizontally into two equal quarter-sections. Curved text labels each section along its outer edge: the left section is labeled '50% vegetables and fruits', the top-right section is labeled '25% protein rich food', and the bottom-right section is labeled '25% carbohydrates'. Each section is filled with illustrations of corresponding foods, such as vegetables and fruits in the largest section, fish and meat in the protein section, and bread and pasta in the carbohydrates section.
The given data is: [{'label': 'vegetables and fruits', 'value': 50, 'percentage': '50%'}, ...]

Figure 27: Case of Pie Chart.



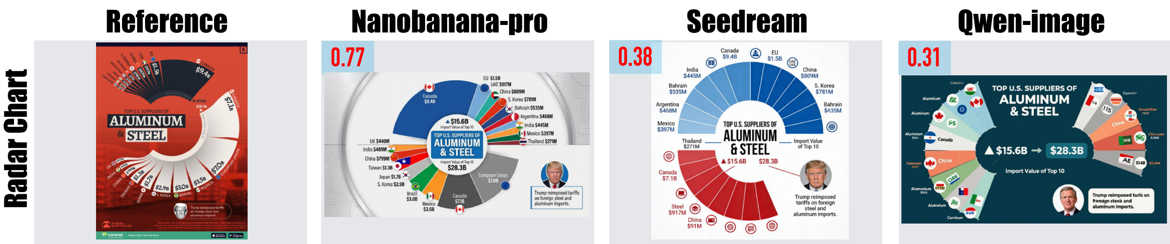
PROMPT: Create an infographic that features a title on the left reading 'The Most Visited Websites in the World' with a subtitle 'As of June 2024'. The main visualization is a bubble chart with 15 circles of varying sizes scattered across the canvas, where the size of each circle corresponds to the data value. Each bubble contains the website's domain name, a large number representing billions of visits, and has a rank number placed adjacent to it. The largest circle for the top-ranked site also includes the text 'Monthly Visits' below the number. Each circle is accompanied by a small, simplified icon representing the associated website. A legend in a rectangular box on the left side lists 'U.S.', 'China', 'Russia', and 'Czech Republic', each preceded by a cursor-shaped icon. A large, stylized letter 'G' is placed near the main title, and a hand cursor icon is positioned in the upper central area.
The given data is: [{'rank': 1, 'website': 'google.com', 'monthly_visits_billions': 82, 'country': 'U.S.'}, ...]

Figure 28: Case of Proportional Area Chart.



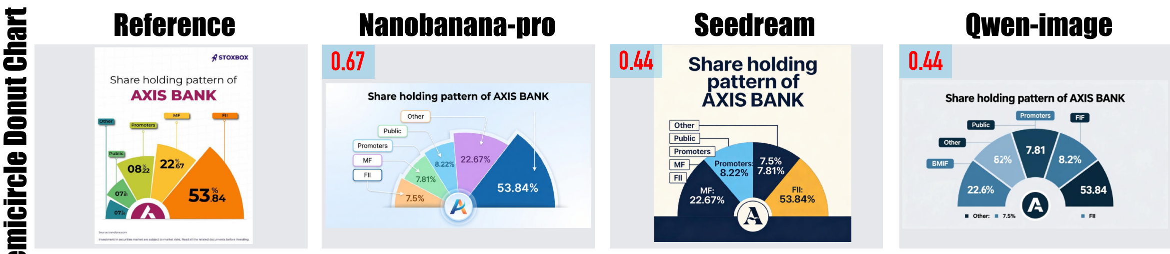
PROMPT: Create an infographic that displays a title at the top, "THÁP DINH DƯỠNG CÂN ĐỐI", with a subtitle below it, "(Trung bình cho 1 người trưởng thành trong 1 tháng)". The main visual is a large food pyramid chart positioned in the center, divided into seven horizontal tiers. Each tier contains illustrations of food items relevant to its category, with the base being the widest and the tip being the narrowest. To the left of the pyramid, a vertical list of text labels corresponds to each tier, detailing the food group and recommended monthly quantity. To the right of the pyramid, another vertical list of text labels corresponds to each tier, providing consumption frequency recommendations. At the bottom of the infographic, below the pyramid, is the text "Dinh dưỡng hợp lý là nền tảng của Sức khỏe". On either side of this bottom text are small illustrative drawings of people exercising. The given data is: [{'item': 'Muối', 'quantity': 'Dưới 180 gr', 'recommendation': 'Ăn hạn chế'}, ...]

Figure 29: Case of Pyramid Chart.



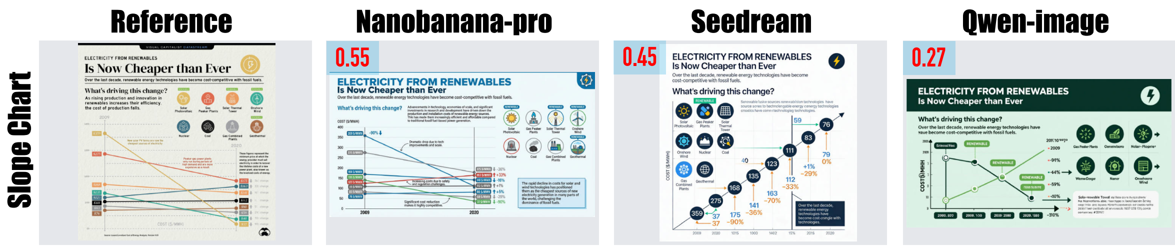
PROMPT: Create an infographic that features a central title block with the text 'TOP U.S. SUPPLIERS OF', followed by 'ALUMINUM' and '& STEEL' on separate lines below. The infographic is structured around two opposing semi-circular radial charts originating from the center. The upper chart fans out towards the top, displaying data with wedges, each labeled with a supplier name, a circular icon, and a value. The lower chart mirrors this structure, fanning out towards the bottom, also displaying data with wedges and corresponding labels containing a supplier name, a circular icon, and a value. Positioned between the two charts are central labels reading '▲ $15.6B' and '$28.3B', with the text 'Import Value of Top 10' pointing to each chart's total. In the lower right, there is a circular cutout image of a person's headshot. Adjacent to the image is a text block containing the sentence: 'Trump reimposed tariffs on foreign steel and aluminum imports.'. The given data is: [{'category': 'Aluminum', 'supplier': 'Canada', 'value_billions': 9.4, 'label': '$9.4B'}, ...]

Figure 30: Case of Radar Chart.



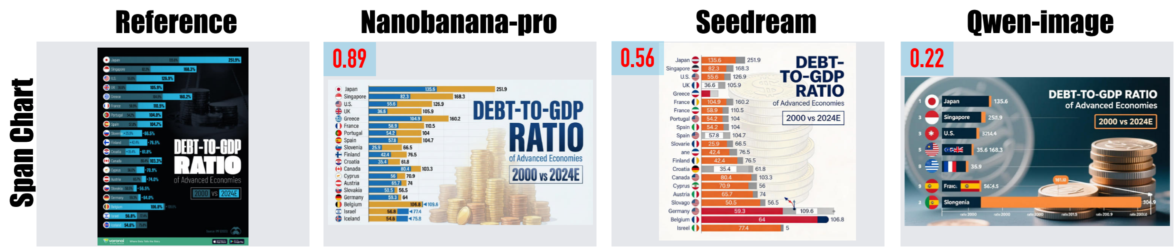
PROMPT: Create an infographic that features a large, centered title, "Share holding pattern of AXIS BANK", at the top. Below the title is a semi-circular chart composed of five wedges of varying sizes, positioned in the lower half of the frame. Each wedge contains a large percentage value. Five rectangular labels, "Other", "Public", "Promoters", "MF", and "FII", are placed above and to the left of the chart. Thin lines connect each label to its corresponding wedge. A circular element containing a stylized letter 'A' is situated at the base of the semi-circular chart.
The given data is: [{'category': 'Other', 'percentage': 7.5}, ...]

Figure 31: Case of Semicircle Donut Chart.



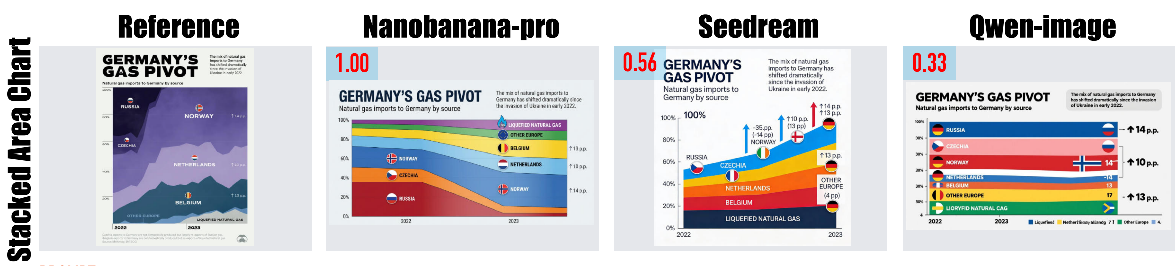
PROMPT: Create an infographic that features a main title "ELECTRICITY FROM RENEWABLES Is Now Cheaper than Ever" and a subtitle "Over the last decade, renewable energy technologies have become cost-competitive with fossil fuels." Below the title, a section header reads "What's driving this change?" followed by an explanatory paragraph, positioned next to a grid of eight circular icons representing energy sources, each with a text label below it and some with a "RENEWABLE" tag above. The main visual is a slope chart with a vertical axis labeled "COST ($/MWH)" on the left, comparing data points from a vertical line labeled "2009" to another labeled "2020". Each of the eight energy sources is represented by a line connecting its 2009 cost value to its 2020 cost value. Data points are marked with circles and accompanied by rectangular labels showing the numeric cost. To the right of the 2020 labels, the percentage change is listed alongside a directional arrow. Annotations with arrows point to specific lines within the chart, and a descriptive text box is located on the lower right. A single circular icon with a lightning bolt symbol is in the top right corner.
The given data is: [{'source': 'Solar Photovoltaic', 'type': 'Renewable', 'cost_2009': 359, 'cost_2020': 37, 'change': '-90%'}, ...]

Figure 32: Case of Slope Chart.



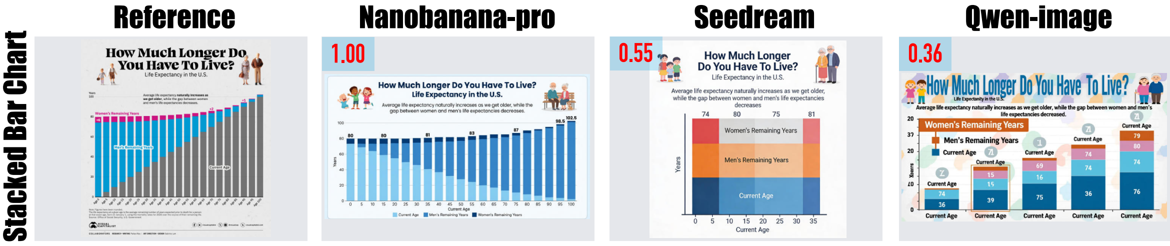
PROMPT: Create an infographic that displays a large title, "DEBT-TO-GDP RATIO", on the right side, with the text "of Advanced Economies" and a framed "2000 vs 2024E" below it. The main visual element is a horizontal bar chart on the left, listing countries vertically from top to bottom. Each country's entry begins with a circular flag icon and its name, followed by a bar representing its data. For most countries, the bar is composed of a shorter segment nested within a longer one, with a numeric value inside the shorter segment and another at the end of the longer bar. For some countries, the visualization consists of a single bar with a value inside it, and a second value indicated by a pointer next to the bar. A large, faint illustration of stacked coins is visible in the background, partially overlapping the title and the chart area.
The given data is: [{'country': 'Japan', 'ratio_2000': 135.6, 'ratio_2024': 251.9}, ...]

Figure 33: Case of Span Chart.



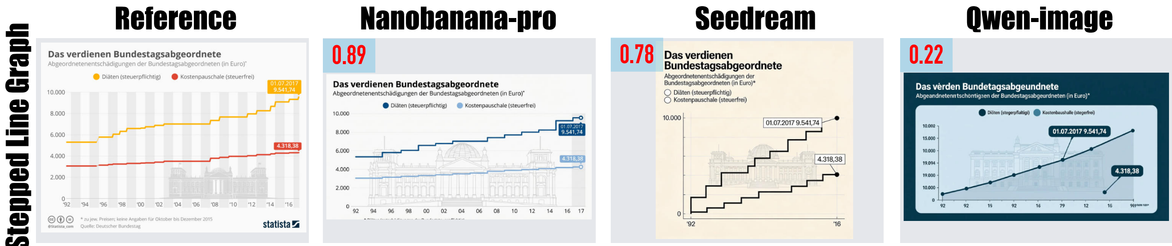
PROMPT: Create an infographic that features a large title, 'GERMANY'S GAS PIVOT', at the top left, with the subtitle 'Natural gas imports to Germany by source' positioned directly below it. To the top right, there is a text block that reads: 'The mix of natural gas imports to Germany has shifted dramatically since the invasion of Ukraine in early 2022.'. The main visual is a 100% stacked area chart with a vertical axis on the left marked with percentage increments and a horizontal axis at the bottom with labels for '2022' and '2023'. This chart is segmented into several stacked areas, each labeled with its source: 'RUSSIA', 'CZECHIA', 'NORWAY', 'NETHERLANDS', 'BELGIUM', 'OTHER EUROPE', and 'LIQUEFIED NATURAL GAS'. Each country name is accompanied by a circular icon of its flag. On the right side of the chart, annotations with an upward arrow indicate percentage point changes, including '↑ 14 p.p.', '↑ 10 p.p.', and '↑ 13 p.p.'. The given data is: [{'source': 'Russia', 'change_pp': -35}, ...]

Figure 34: Case of Stacked Area Chart.



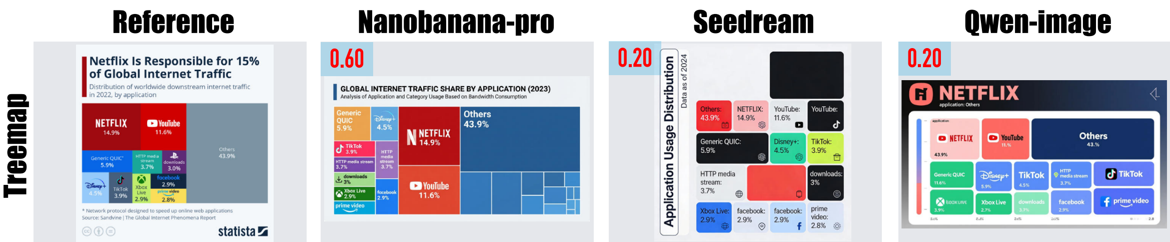
PROMPT: Create an infographic that displays a title, 'How Much Longer Do You Have To Live?', with a subtitle, 'Life Expectancy in the U.S.', positioned at the top center. Small illustrative figures of children are on the top left, and an elderly couple are on the top right. A central vertical stacked bar chart is the main feature, with its vertical axis labeled 'Years' and its horizontal axis showing labels for 'Current Age' in five-year increments. Each bar is composed of three segments stacked vertically: the bottom segment represents the current age, the middle segment represents men's remaining years, and the top segment represents the additional remaining years for women. A descriptive sentence, 'Average life expectancy naturally increases as we get older, while the gap between women and men's life expectancies decreases,' is placed above the chart. Labels within the chart identify 'Women's Remaining Years', 'Men's Remaining Years', and 'Current Age' segments. Small numeric annotations are placed above select bars. The given data is: [{'current_age': 0, 'men_remaining_years': 74, 'women_gap_years': 6, 'men_total_expectancy': 74, ' women_total_expectancy': 80}, ...]

Figure 35: Case of Stacked Bar Chart.



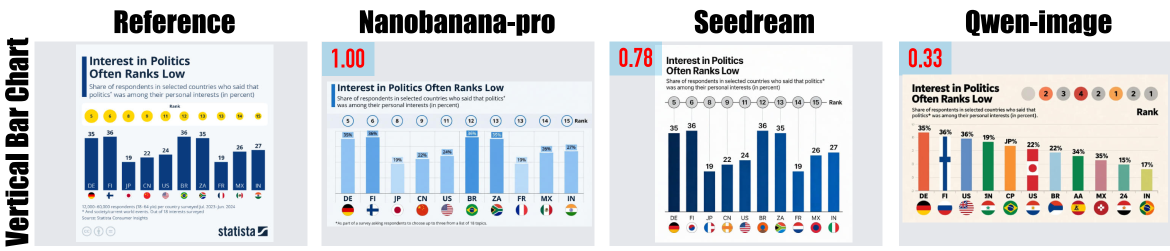
PROMPT: Create an infographic that features a title, 'Das verdienen Bundestagsabgeordnete', at the top left, with a subtitle 'Abgeordnetenentschädigungen der Bundestagsabgeordneten (in Euro)*' directly below. A centered legend below the subtitle identifies two categories with circle icons: 'Diäten (steuerpflichtig)' and 'Kostenpauschale (steuerfrei)'. The main visual is a stepped line chart with a vertical axis on the left marked with values from 0 to 10.000, and a horizontal axis at the bottom labeled with years from '92 to '16. Two stepped lines plot the data for the two categories across the years. An annotation box reading '01.07.2017 9.541,74' points to the final data point of the upper line, while another annotation box reading '4.318,38' points to the final data point of the lower line. A faint line-art illustration of a large building is centered in the background of the chart's plot area. The given data is: [{'year': 1992, 'category': 'Diäten (steuerpflichtig)', 'value': 5.3}, ...]

Figure 36: Case of Stepped Line Graph.



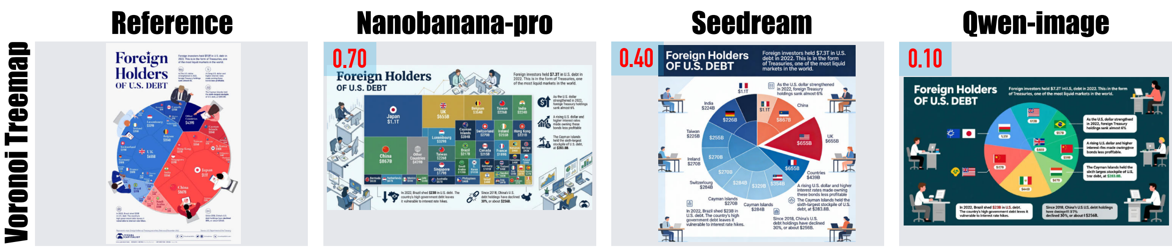
PROMPT: Create an infographic that features a large title and subtitle at the top left, with a vertical bar placed to the left of the title. The main visual element is a treemap chart that occupies the area below the header. This chart is comprised of multiple rectangular blocks, where the size of each block is proportional to its value. Inside each block, there is text specifying the name of an application or category and its corresponding percentage. Some blocks also contain a representative icon positioned near the text label. The blocks are arranged to fill the chart space, with the largest rectangles positioned in the upper and right sections. The given data is: [{'application': 'Others', 'percentage': 43.9}, ...]

Figure 37: Case of Treemap.



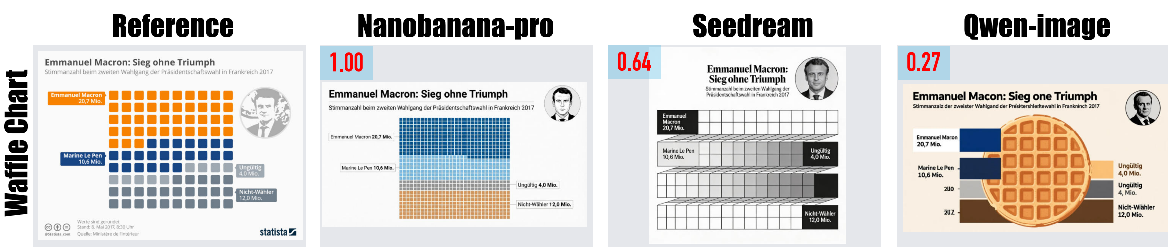
PROMPT: Create an infographic that has a main title at the top left, "Interest in Politics Often Ranks Low", with a subtitle below it reading, "Share of respondents in selected countries who said that politics* was among their personal interests (in percent)". The central element is a vertical bar chart displaying data for ten countries. Each bar has its corresponding percentage value labeled directly above it. Below each bar is a two-letter country code, and under the code is a circular icon representing the country's flag. Above the entire chart, there is a horizontal row of circular elements, each containing a rank number aligned over its respective bar. To the right of this row of circles is the label "Rank".
The given data is: [{'country_code': 'DE', 'percentage': 35, 'rank': 5}, ...]

Figure 38: Case of Vertical Bar Chart.



PROMPT: Create an infographic that displays a large, central treemap-style area chart with the title 'Foreign Holders OF U.S. DEBT' in the upper left corner. Each segment of the treemap is sized proportionally to its value and contains a small flag icon, the country's name, and a value label. A descriptive paragraph, 'Foreign investors held $7.3T in U.S. debt in 2022. This is in the form of Treasuries, one of the most liquid markets in the world.', is located at the top right. Several smaller text blocks, each accompanied by a simple icon, are placed around the chart. On the right are three blocks: 'As the U.S. dollar strengthened in 2022, foreign Treasury holdings sank almost 6%', 'A rising U.S. dollar and higher interest rates made owning these bonds less profitable', and 'The Cayman Islands held the sixth-largest stockpile of U.S. debt, at $283.8B.'. At the bottom, the text 'In 2022, Brazil shed $23B in U.S. debt. The country's high government debt leaves it vulnerable to interest rate hikes.' is on the left, and 'Since 2018, China's U.S. debt holdings have declined 30%, or about $256B.' is on the right. Stylized illustrations of people working at desks are arranged around the perimeter of the chart.
The given data is: [{'country': 'Japan', 'label_value': '$1.1T', 'value_billions': 1100}, ...]

Figure 39: Case of Voronoi Treemap.



PROMPT: Create an infographic that has a title 'Emmanuel Macron: Sieg ohne Triumph' and a subtitle 'Stimmanzahl beim zweiten Wahlgang der Präsidentschaftswahl in Frankreich 2017' at the top. The main visual is a large waffle chart composed of a grid of small squares, which is divided into four distinct horizontal blocks stacked vertically. The top block has a label to its left that reads 'Emmanuel Macron 20,7 Mio.'. The second block has a label to its left that reads 'Marine Le Pen 10,6 Mio.'. The third block has a label to its right that reads 'Ungültig 4,0 Mio.'. The bottom block has a label to its right that reads 'Nicht-Wähler 12,0 Mio.'. In the upper right corner, there is a circular, monochrome portrait illustration.
The given data is: [{'category': 'Emmanuel Macron', 'value': 20.7, 'unit': 'Mio.'}, ...]

Figure 40: Case of Waffle Chart.

Instructions Given to Human Evaluators

Task Description
You are asked to verify whether a generated infographic satisfies specific requirements. Each requirement is presented as an independent yes/no question and refers to a single visual or semantic constraint of the infographic.

Evaluation Procedure
For each question, examine the corresponding infographic and provide a binary answer:
- Yes: the requirement is fully and unambiguously satisfied.
- No: the requirement is not satisfied.

You must base your judgment only on the information visible in the infographic and the content of the question. Do not rely on external knowledge, assumptions, or prior familiarity with the topic.

Judgment Criteria
A question should be answered No if any of the following conditions apply:
- The required element is missing.
- The element is only partially satisfied.
- The element is present but incorrectly rendered.
- The element is visually ambiguous (e.g., unclear text, distorted shapes, unreadable labels, or imprecise visual encoding).

If the requirement cannot be clearly verified from the image, you should answer No.

Additional Notes
- Do not infer missing information.
- Do not attempt to correct errors or guess intended content.

Figure 41: Instructions provided to human evaluators for the annotation task.

Chart Type Detection

You are a Senior Taxonomy Engineer and Visual Data Analyst.

Your Task: Your sole task is to analyze the provided infographic image and extract its meta-information into a strict JSON format. Your analysis must be meticulous, prioritizing data rigor and adherence to the specified constraints for downstream filtering and clustering operations.

Requirements:
CHART TYPE CLASSIFICATION (`chart_type`)

Step 1: Multi-Chart Detection (High Priority)

- Examine the entire image. Does it contain two or more distinct charts or data visualizations?
- Specific Trigger: If the image is a long vertical infographic (containing multiple sections/blocks, where different sections have their own charts) or a dashboard, you MUST select "Bonus".
- Rule: If multiple charts exist, ignore the specific types of the individual charts and immediately classify as "Bonus".

Step 2: Single Chart Classification
Only if the image contains exactly one principal chart, select the best specific description from the Candidate List below.

- Candidate List (Select exactly one): ["Vertical Bar Chart", "Horizontal Bar Chart", "Radial Bar Chart", "Stacked Bar Chart", "Grouped Bar Chart", "Pictorial Chart", "Histogram", "Lollipop Chart", "Dot chart", "Diverging Bar Chart", "Dumbbell Plot", "Span Chart", "Bump Chart", "Line Graph", "Spline Graph", "Stepped Line Graph", "Slope Chart", "Area Chart", "Layered Area Chart", "Range Area Chart", "Stacked Area Chart", "Radial Area Chart", "Pie Chart", "Donut Chart", "Semicircle Donut Chart", "Radar Chart", "Scatter plot", "Proportional Area Chart", "Bubble Chart", "Heatmap", "Waffle Chart", "Alluvial Diagram", "Gauge Chart", "Funnel Chart", "Pyramid Chart", "Treemap", "Voronoi Treemap", "Bonus"]

Output (Strict JSON):
Provide the result exclusively in the following JSON format. Do not include any explanatory text, pre-amble, or conversational language.
{ "chart_type":"<selected chart type>" }

Figure 42: Prompt for chart type detection.

T2I Prompt Construction

You are an expert in infographic content extraction.

Your Task From the uploaded infographic image, extract a **generation-ready structural design description** that captures only the factual, visually verifiable elements needed for a text-to-image model to recreate the infographic.

Important Principles

- Focus strictly on **structure, layout, and visual encoding**.
- Do **not** describe aesthetic style, colors, textures, fonts, or background appearance.
- Do **not** mention logos, watermark icons, data sources, copyright icons, or footnotes.
- Do **not** infer hidden meaning or add any details not directly visible.
- Do **not** restate individual numeric values from the dataset.
- All visible **textual content in the infographic** must be included verbatim.
- The output must be **one single paragraph within 10 senetences**, beginning with: **"Create an infographic that..."** and must end with the sentence: **"The given data is: {data}."**

What the Structural Description Should Include

Describe only these factual visual components:

1. **Overall layout** (e.g., title position, chart placement, grouping)
2. **Chart type** (e.g., pie chart, horizontal bar chart)
3. **Data encoding** (e.g., bars represent values, labels placed beside elements)
4. **Text placement** (titles, subtitles, labels, annotations)
5. **Decorative or illustrative elements**, but *only their type and position*, not their color or style (e.g., "an icon of a hand with coins on the right side" — without stylistic details)

Output (JSON ONLY) Return a JSON object in the exact form: {{ "t2i_prompt": "<one-paragraph structural design description>" }}

Figure 43: Prompt used in T2I prompt construction.

Question Generation

You will receive a visual description from a T2I prompt (without the data list). Your task is to convert each sentence into **one semantic-fidelity yes/no question**.

Rules

- Split the description into separate sentences (each sentence is a ground).
- For each ground, write **one visually verifiable yes/no question**.
- Keep all specific visible text (titles, labels, annotations, etc.) exactly as written.
- **If a ground contains an ambiguous pronoun** (e.g., "this section", "this chart", "it", "they"), **rewrite the question so that it becomes fully self-contained** and does not rely on context.

Example (with mandatory self-contained rewriting)

Visual description: The right section presents a vertical bar chart with years labeled along the horizontal axis. In this chart, each bar contains its numerical value inside it.

Expected output:

[{{ "ground": "The right section presents a vertical bar chart with years labeled along the horizontal axis.", "question": "Does the right section present a vertical bar chart with years labeled along the horizontal axis?" }},
{{ "ground": "In this chart, each bar contains its numerical value inside it.", "question": "In the vertical bar chart, does each bar contain its numerical value inside it?" }}]

Now process this visual description:
{**visual_description**}

Output format (Only JSON)
{{ "ground": "...", "question": "..." }}

Figure 44: Prompt for question generation.

Question Type Classification

You are an expert infographic analyst.

Your task is to classify a single semantic verification question into exactly one primary infographic element category. You must choose the category that represents the dominant visual element being verified by the question.

Do not assign multiple categories.
Do not explain your reasoning.
Do not output anything other than the category name.

Allowed Output Categories

You must choose exactly one category from the following list: [Title & Subtitle Chart / Diagram Type， Data Marks， Data Encoding，Axes & Scales，Legend & Category Mapping, Annotations & Callouts, Decorative / Non-data Elements]

You must output the category name exactly as written. You must NOT output anything else. Any output that is not exactly one of the names above is invalid.

Category Definitions

- Title & Subtitle: the main title or subtitle of the infographic. Includes the text content and its placement as a single visual element.
- Chart / Diagram Type: the overall chart or diagram form. Examples include Sankey diagrams, bar charts, donut charts, maps, timelines, dot plots, dumbbell charts, and similar structures.
- Data Marks: primary visual objects that directly represent data. Examples include nodes, bars, circles, dots, areas, blocks, bands, or map regions. The visual object, its label, and its relative placement are treated as a single element.
- Data Encoding: how data values are mapped to visual properties. Examples include size, width, color, position, direction, magnitude, or sign (positive vs. negative).
- Axes & Scales: Axes, tick marks, gridlines, zero lines, and scale labels. Includes both horizontal and vertical axes.
- Legend & Category Mapping: legends or keys that explain how colors, shapes, or symbols correspond to categories or groups.
- Annotations & Callouts: numeric labels, explanatory text, callout boxes, or textual annotations that are not part of axes or legends.
- Decorative / Non-data Elements: icons, illustrations, background graphics, or visual elements that do not directly encode data.

Input Format

Question:
{semantic_question}

Output Format
<category name>

Figure 45: Prompt for question type classification.

Question Augmentation From Seed Question

You will receive a **data-aware design draft** from a T2I prompt (with source data).
Your task is **NOT** to invent arbitrary questions. Instead, you must **instantiate four fixed data-fidelity seed requirements** into **four chart-specific, visually verifiable yes/no questions**.
Each question should be a concrete rewriting of one seed, conditioned on the content of the infographic described in the prompt.

Data-Fidelity Seed Requirements (FIXED)
For **each chart**, you must generate **exactly four questions**, corresponding to the following four seeds:

Seed 1: Data completeness & coverage
- All required data items from the source data appear in the chart.
- No extra data items appear that are not present in the source data.

Seed 2: Annotations & Callouts (Conditional)
Generate this seed ONLY IF the design draft explicitly specifies numeric labels, percentages, counts, or axis values that are expected to appear as text in the chart.

- Verify that all numeric labels, annotations, or axis values shown in the chart are internally consistent as written, without obvious textual or numerical contradictions (e.g., inconsistent labels, impossible totals, conflicting numbers).
- If the design draft does NOT specify numeric labels or values, do NOT generate a Seed 2 question.

Seed 3: Ordering (Conditional)
Generate this seed ONLY IF the design draft explicitly specifies an intended ordering or ranking (e.g., "sorted from largest to smallest", "top-ranked categories", "in descending order").

- Verify that the visual order shown in the chart matches the intended ordering described in the design draft. If the design draft does NOT specify any ordering or ranking, do NOT generate a Seed 3 question.

Seed 4: Data Encoding (Magnitude & proportion)
- The visual encoding should preserve the **relative magnitudes and proportional relationships** implied by the data (e.g., larger differences should appear visually larger).
- Visual scaling should not introduce **severe distortion** of relative differences (e.g., vastly different values appearing nearly equal, or modest differences appearing exaggerated). - Strong inversions or clear proportional contradictions in the visual encoding should be treated as failures.

Rules
- The infographic may contain **multiple charts**. You must generate questions **per chart**.
- For **each chart**, generate **2–4 questions**:
  - Always generate **Seed 1** and **Seed 4** (so at least 2 questions per chart).
  - Do **not** introduce new requirements beyond the four seeds.
  - Do **not** write one question per data row; questions must be **chart-level**.
- Each question must be answerable by visually inspecting the chart.
- All questions must be strictly self-contained: they must be answerable using only what is visible in the infographic (visual marks, labels, legends, text, etc). Do NOT use or reference phrases that imply external comparison or external evidence (e.g., "source data", "original data", "according to the data", "from the dataset", etc).

Input data-aware design draft: {t2i_prompt}

Output format (JSON ONLY)
[{{ "ground": "Seed <n>: <one instantiated data-fidelity requirement derived from a seed>", "question": "<a concrete yes/no question verifying this requirement>" }}]

Figure 46: Prompt used for question augmentation from seed questions.

LLM Evaluation

You are a strict factual evaluator.

Your task: Inspect the infographic image (provided separately) and answer the binary factual question below.

Rules:
- Answer **1** ONLY if the requirement is clearly satisfied in the image.
- Answer **0** if the requirement is NOT satisfied, unclear, ambiguous, partially met, or cannot be confirmed.
- No partial credit. Ambiguity = 0.
- Base your judgment ONLY on visible evidence in the infographic.
- Even if the image is empty, blank, corrupted, unreadable, or clearly incorrect, you MUST still output a valid JSON object following the required format. In such cases, the answer should be 0.

FACTUAL QUESTION: {question}

Output Format (JSON ONLY):
{{ "analysis": "<your reasoning based strictly on what is visible>", "answer": "<0 or 1>" }}

The response must contain only valid JSON.

Figure 47: Prompt used in LLM evaluation.