

RAViG-BENCH: A BENCHMARK FOR RETRIEVAL-AUGMENTED VISUALLY-RICH GENERATION WITH MULTI-MODAL AUTOMATED EVALUATION

Anonymous authors

Paper under double-blind review

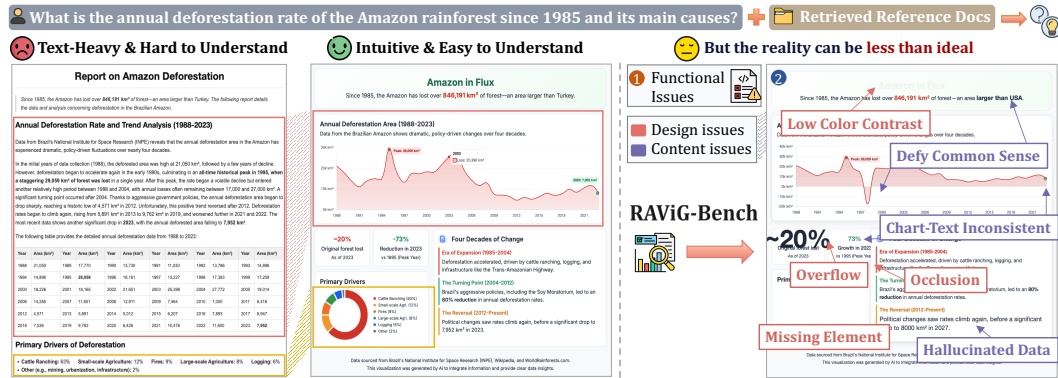


Figure 1: Compared with traditional text-heavy answers (left), RAViG responses (middle) are more intuitive and user-friendly for queries that entail presenting substantial data. However, such outputs may still suffer from various issues in practice (right), which our benchmark is designed to detect.

ABSTRACT

Retrieval-Augmented Visually-rich Generation (RAViG) extends RAG by integrating textual explanations with multiple visual elements in a well-structured layout. Despite its growing adoption, no existing benchmark offers a holistic evaluation of RAViG. Current RAG benchmarks focus on text-only generation, while natural language to visualization (NL2VIS) benchmarks focus on “show-data-as-chart” style queries and do not follow the RAG paradigm. To address this deficiency, we present RAViG-Bench, the first comprehensive benchmark specifically designed for RAViG. The benchmark features a diverse collection of authentic user queries, each paired with real-world web retrievals to simulate realistic RAViG scenarios. Besides, we introduce a novel multi-modal automated evaluation framework that holistically assesses the quality of RAViG outputs. This framework scrutinizes the generated content by evaluating the functionality, design quality, and content quality of both textual and visual components. Our extensive experiments on leading commercial and open-source LLMs provide a comprehensive analysis of their current capabilities, highlighting significant limitations and charting key directions for future research in this emergent area.

1 INTRODUCTION

Recently, Retrieval-Augmented Generation (RAG) that produces outputs comprising illustrative visual elements and explanatory text has seen widespread adoption within enterprise Large Language Models (LLMs) (Google, 2025; Shanghai Mita Network Technology Co., Ltd., 2025; Perplexity AI, Inc., 2025). This trend is driven by the increasing demand for statistical analysis in fields such as meteorology, finance, and public health. For instance, when queried for an analysis of the annual deforestation rate of the Amazon rainforest, a capable model retrieves relevant weather information and constructs a visually rich report. As depicted in the middle of Fig. 1, such a report features a line chart depicting deforestation area over time and a pie chart for sales distribution, with explanatory text embedded within the information panel and all elements organized into a user-friendly layout. Consequently, we term this paradigm, which integrates illustrative visual elements with explanatory text into the generation process, as Retrieval-Augmented Visually-rich Generation (RAViG).

Benchmark	Knowledge Source	Output Modality	Design Eval		Content Eval		GT-Free Eval	Real-World Fidelity
			Chart	Chart-Text	Text	Chart-Text		
nvBench	In-Query Data	Single Chart	✓	✗	✗	✗	✗	✗
VISEval	In-Query Data	Single Chart	✓	✗	✗	✗	✗	✗
RAGAs	Web Retrieval	Plain Text	✗	✗	✓	✗	✓	✗
CRAG	Web Retrieval	Plain Text	✗	✗	✓	✗	✗	✗
GaRAGe	Web Retrieval	Plain Text	✗	✗	✓	✗	✗	✗
RAViG-Bench	Web Retrieval	Rich Content	✓	✓	✓	✓	✓	✓

Table 1: Comparison of RAViG-Bench with existing evaluation benchmarks.

However, a significant gap exists in the literature: the lack of a dedicated benchmark for the evaluation of RAViG. As shown in Tab. 1, prevailing benchmarks are inadequate, since they either focus on text-only RAG or on the "show-data-as-chart" question-answering paradigm. Existing RAG benchmarks lack the necessary data to support visually-rich generation and the multi-modal metrics required to evaluate the interplay between visual and textual contents. Moreover, their query data, often derived from knowledge graphs or LLM generations, deviates significantly from real-world distributions (Yang et al., 2024a; ES et al., 2024; Sorodoc et al., 2025). On the other hand, benchmarks for natural language to visualization (NL2VIS) primarily evaluate the visual fidelity of generated charts, but they operate under a different premise where chart data is provided alongside the natural language query (Chen et al., 2025; Luo et al., 2025). In such settings, the data for visualization is pre-defined, reducing the task to one of direct data-to-chart translation. Furthermore, these benchmarks typically focus on single-chart outputs, failing to address the composite nature of RAViG, which involves an integrated presentation of both text and multiple visualizations.

The need for a dedicated RAViG benchmark is further underscored by several significant and recurrent challenges observed in existing LLMs. 1) Functional Failures: We choose HTML for visually-rich generation due to its widespread adoption, cross-platform compatibility, and high extensibility for integrating multi-modal content. However, LLMs sometimes produce syntactically incorrect or incomplete HTML, preventing the render of a visually-rich webpage. 2) Design Deficiencies: The generated visual elements often violate fundamental principles of user interface design, thereby impairing readability and user acquisition of information. As depicted in Fig. 1, common issues include low color contrast that renders text or icons illegible, text overflow, and empty content boxes to element rendering failure. 3) Content Hallucination: LLMs are prone to hallucination, generating content that is factually incorrect or inconsistent. As shown in Fig. 1, a generated line chart might contradict the accompanying textual summary, or data presented in tables may be fabricated.

Motivated by these observations, we posit the necessity of a benchmark designed for RAViG. Such a benchmark should feature diverse, real-world queries that genuinely necessitate visually-rich answers. Furthermore, it requires a robust multi-modal evaluation mechanism capable of assessing not only the individual quality of visual and textual components but also their interaction.

To this end, we introduce the first dataset for Retrieval-Augmented Visually-rich Generation. Our dataset is characterized by three key attributes: 1) Realism. The majority of our queries are drawn from the NQ dataset (Kwiatkowski et al., 2019), which comprises real user questions to the Google Search engine. An automated selection strategy is employed to identify queries that are genuinely suitable for visually-rich responses. 2) Diversity. The dataset spans six distinct query categories derived from a wide range of real-world scenarios. To mitigate the inherent imbalance in the distribution of real-world queries, we augment the dataset with synthetic examples, preserving the original distribution while ensuring an adequate amount for each category. 3) Curation. For each query, we collect the top-10 web search results from a real search engine, retaining the inherent noise of realistic retrieval. Subsequently, human annotators confirm the suitability of each query for visually-rich generation by jointly considering the query type and the content of its references.

Second, we propose a novel, automatic multi-modal evaluation framework that assesses the generated visually-rich answers along three critical dimensions: 1) Functionality Validation. We employ syntax checkers to verify whether the generated output is well-formed and can be rendered without errors. 2) Design Quality Assessment. Drawing from an analysis of common failure cases, we identify five prevalent design issues that hinder information acquisition. We develop a hybrid evaluation paradigm that combines an "MLLM-as-a-Judge" approach with rule-based methods to automatically detect these visual defects. 3) Content Quality Evaluation. We extend traditional RAG evaluation

metrics from text-only generation to holistically assess both textual and visual content across three dimensions: Reasonableness, Comprehensiveness, and Faithfulness. For this coordinated assessment, we employ an LLM-based judge assisted by a “claim extraction and verification” strategy, allowing fine-grained and accurate evaluation.

To summarize, our contributions are summarized as follows:

- We propose the first dataset for Retrieval-Augmented Visually-rich Generation (RAViG), featuring real-world queries, diverse categories, and web-sourced retrievals to simulate realistic scenarios.
- We present a comprehensive, automated multi-modal evaluation framework for RAViG that assesses the interplay of textual and visual elements across functionality, design, and content quality.
- We benchmark leading commercial and open-source LLMs, providing a comprehensive analysis that reveals their current limitations and outlines key directions for future research.

2 RELATED WORK

Existing RAG benchmarks mainly evaluate textual outputs, measuring aspects such as accuracy, faithfulness, and hallucination (Gan et al., 2025; Sorodoc et al., 2025; Yang et al., 2024a; Saad-Falcon et al., 2024; Liu et al., 2023; Niu et al., 2024; Katsis et al., 2025; ES et al., 2024; Chen et al., 2024), with some extending to multi-modal understanding (Zhan et al., 2025; Strich et al., 2025; Masry et al., 2022; Dong et al., 2025; Ma et al., 2024; Ding et al., 2024; Wu et al., 2025) or code generation (Liu et al., 2024b; Wang et al., 2025). Among these, MRAMG-Bench (Yu et al., 2025) is most similar to our setting, but it is essentially a text-only generation task, as its answers combine model-generated text with retrieval-sourced images. Existing benchmarks, however, rarely assess an LLM’s ability to autonomously design and generate visually rich responses that integrate retrieved knowledge. In parallel, NL2VIS benchmarks focus on converting structured queries into single chart from predefined data (Cheng et al., 2023; Vázquez, 2024; Li et al., 2024; Luo et al., 2021a; 2025; Chen et al., 2025; Podo et al., 2024; Pan et al., 2025; Song et al., 2024; Joseph et al., 2025), limiting the model to direct data-to-chart translation. Our setting is more open-ended and realistic: given a query, the model must decide when and how to combine multiple visualizations with text into a coherent answer. To our knowledge, no existing benchmark performs a comprehensive, end-to-end evaluation of this capability. See Appendix A for a detailed discussion.

3 RAVIG-BENCH DATASET

The RAViG-Bench dataset aims to provide a challenging and realistic benchmark for retrieval-augmented visually-rich generation. It is constructed through a four-stage pipeline: 1) selecting queries from a real-world dataset, 2) synthesizing data to address underrepresented categories, 3) creating reference documents from a real search engine, and 4) conducting a final human review (Sec. 3.1). We also perform detailed statistical analyses to further characterize the dataset (Sec. 3.2).

3.1 DATASET CONSTRUCTION

Automated Query Selection. To better capture real user needs, we derived tasks from the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019), which contains over 300,000 genuine Google search queries and serves as a high-quality real-world data resource. We removed non-English queries and then employed Gemini-2.5-pro (Comanici et al., 2025) with a detailed prompt to identify queries suitable for visually-rich responses. The prompt leverages query keywords, type, and semantic intent to apply explicit selection criteria, identifying queries that request for quantitative or statistical data (e.g., population, price), trend or time-series analysis, quantitative comparisons (e.g., market share), or proportions and compositions. A query is suitable only when the data contains explicit numerical attributes and its visual form would aid comprehension. For each query, the model outputs a binary suitability decision with justification.

Automated Query Synthesis. We classified all queries into six categories reflecting common data visualization intents: trend (changes over time), proportion (part-to-whole relationships), comparison (contrasting distinct items), distribution (how data spreads over a range), ranking (ordering by a metric), and details (retrieving raw or specific information). Detailed descriptions and examples are provided in Appendix E. To mitigate the highly skewed distribution of real-world queries, we augmented each type with an equal number of synthetic queries, ensuring adequate samples per category without substantially altering the original distribution. As LLMs are now widely employed for data synthesis (Yuan et al., 2021; Chen et al., 2022), we employed Gemini-2.5-pro to generate

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

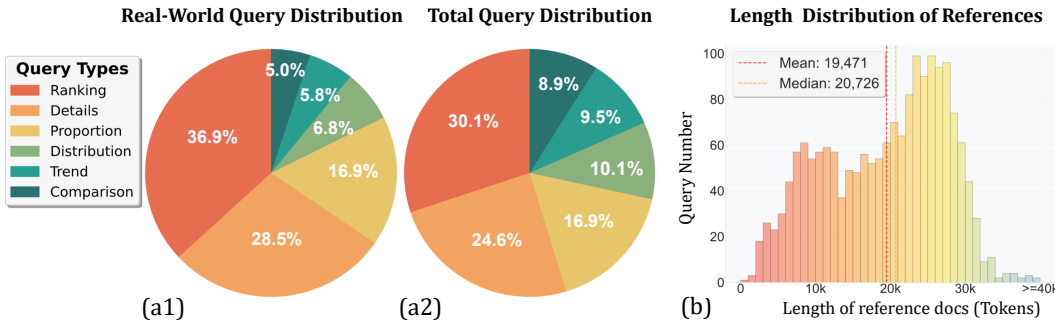


Figure 2: **Dataset statistics.** (a1&a2) Distribution of query categories. (b) Length distribution of reference documents.

queries that met the aforementioned criteria. These automated stages yielded an initial query set that captures real-world scenarios and is inherently suitable for visualization.

Reference Documents Retrieval. To simulate realistic RAG scenarios, we paired each query with the contents of the top 10 web pages returned by the Google Search API for response generation. Apart from manually removing content that is overtly biased or harmful, all other documents are retained without filtering, preserving the natural noise distribution in real search environments. This setup creates a noisy yet structured information environment that challenges models to perform cross-document synthesis and selective integration. Each document is represented in a standardized key-value schema containing fields such as page title, synopsis, content, and additional metadata, as detailed in Tab. 4 in Appendix B. This multi-granularity representation provides rich contextual information to support model reasoning.

Final Human Review. The suitability of producing a visually enriched answer depends not only on the query type but also on the content of its reference documents. For example, when asked “What are the global smartphone sales trends over the past five years?”, purely descriptive news would warrant a textual summary, whereas aggregated time-series sales data would justify a line chart. Therefore, to refine dataset quality, we first used Gemini-2.5-pro to summarize the quantitative content within each query’s references, after which human annotators validated query suitability by jointly considering the query and the content of its references. This process yielded 1,798 queries for the RAViG-Bench dataset, including 1,198 real-world queries and 600 synthetic queries.

3.2 DATASET STATISTIC

Query Category Distribution. We analyzed the distribution of queries across six categories. Real-world queries from the NQ dataset follow the original distribution shown in Fig. 2(a1). After augmentation with synthetic queries, the final distribution is presented in Fig. 2(a2). This process retains the key characteristics of the real-world data, ensuring the realistic fidelity of RAViG-Bench.

Reference Document Characteristics. As the primary information source for answer generation, the length of reference documents directly influences task difficulty by determining both the amount of information a model must process and the size of the required context window. As shown in Fig. 2(b), lengths range from a few thousand to over 40k tokens, presenting a broad spectrum of context scales. This diversity not only affects computational requirements, but also brings varying levels of information density, challenging models to adapt summarization and reasoning strategies when handling both concise and highly verbose documents.

4 RAVIG-BENCH EVALUATION FRAMEWORK

We propose an automated, multi-stage, and multi-modal framework to evaluate LLMs on RAViG. First, it validates functionality to ensure appropriate model responses and complete, well-formed HTML (Sec. 4.1). Passed responses are then assessed for design quality (Sec. 4.2), checking for visual issues that might impair readability. Finally, efficient responses that pass the first two stages are scored for content quality (Sec. 4.3). We then report the agreement between our evaluation results and human expert judgments, demonstrating the effectiveness of our framework (Sec. 4.4).

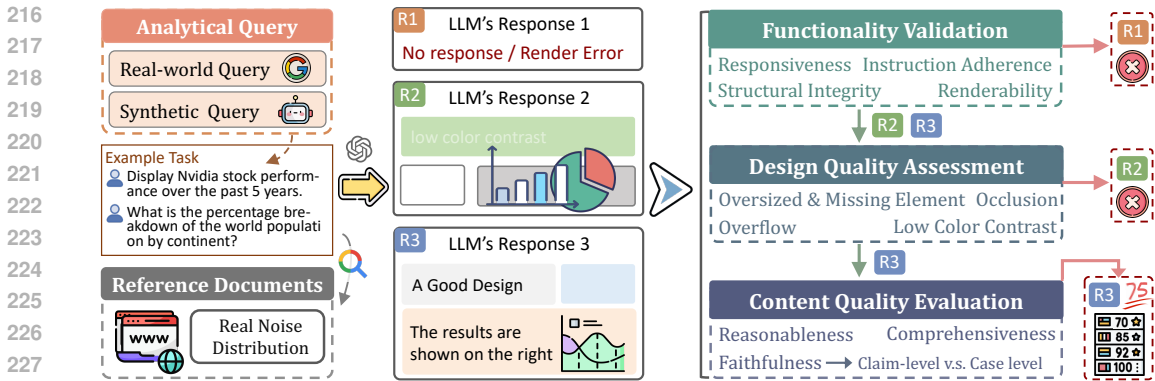


Figure 3: **RAViG-Bench Overview**. Given analytical queries and retrieved references (left), LLMs generates visually-rich responses (middle), which we then evaluate in three stages (right): functionality and design checks filter out invalid or flawed responses, while the content stage assigns a quality score to the remaining well-designed responses.

4.1 FUNCTIONALITY VALIDATION

We developed a multi-step process to assess the functionality of LLMs. For each query, we first check responsiveness, requiring the model to produce an output within at most five attempts. Next, we verify the instruction adherence by ensuring the output contains only HTML code without extraneous conversational text. We then validate the structural integrity of the HTML output using the lxml library to detect issues such as mismatched tags, truncation, or missing DOCTYPE declarations. Code that passes these checks is rendered in a sandbox environment simulating a real-world, web-based QA scenario of an LLM via the Chromium engine with an 800-pixel-wide fixed viewport. An output is deemed functionally valid only if it renders successfully; in such cases, we also capture a screenshot of the rendered result as visual evidence for subsequent design quality assessment.

4.2 DESIGN QUALITY ASSESSMENT

Based on extensive case analysis, we classify the visual defects that may impede user acquisition of information into five types (Fig. 4). Inspired by related “MLLM-as-a-Judge” work (Chen et al., 2025), we employ GPT-4o to detect missing elements and occlusion. For oversized elements, overflow, and low color contrast, where LLMs are prone to significant hallucination, we adopt rule-based evaluations. A design is deemed “passed” only if all five defects are absent.

Oversized Element occurs when components such as charts or icons are rendered at an excessively large scale (Fig. 4(a)), disrupting layout and harming usability. For example, an oversized chart may require excessive vertical scrolling, while an oversized icon can visually unbalance the page. Detection is performed in a Playwright sandbox, programmatically measuring rendered component size. A chart is flagged if its rendered height exceeds twice the viewport height; an icon is flagged if its rendered size exceeds the code-specified dimensions, or if no size is specified, which causes rendering at a large intrinsic size.

Missing Element occurs when elements defined in the HTML fail to render and remain invisible to the user (Fig. 4(b)), typically due to layout or formatting errors. Detection is performed by providing GPT-4o with both the HTML code and its rendered screenshot, prompting it to compare the two to identify elements present in code but absent from the screenshot. Furthermore, we include few-shot instances in the prompt to help the model better understand the task.

Occlusion occurs when overlapping elements hinder readability (Fig. 4(c)). When evaluating an entire page screenshot directly, the model’s attention can be diluted, causing it to miss localized or minor overlaps (Fig. 4(c), top). To mitigate this, we adopt a divide-and-conquer strategy: HTML `<h1>/<h2>` tags are used to segment the page into semantic modules. Screenshots of modules from the same response are then processed as a group, with GPT-4o examining each module individually for occlusion. A response is classified as defective if any module contains an occlusion issue. This modular approach improves detection accuracy by focusing the model on smaller, localized regions. Few-shot examples are also included in the prompt.

Overflow occurs when a container’s content exceeds its boundaries, resulting in truncation or occlusion of adjacent components (Fig. 4(d)). Unlike perceptual issues, detecting overflow requires

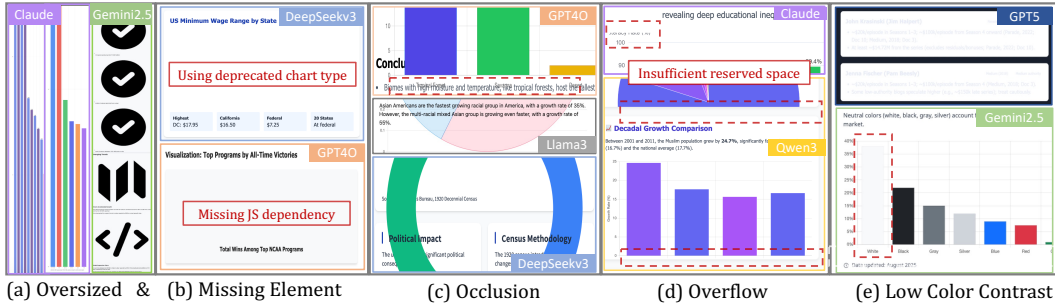


Figure 4: Examples of five primary types of **design defects** (Please zoom in for details).

precise geometric analysis, which vision models often struggle with due to limited fine-grained spatial reasoning. We therefore developed a rule-based algorithm. The page is first segmented into visual modules by combining CSS-based structural analysis (for explicitly defined components such as cards or panels) with *heuristic analysis* that identifies distinct visual modules via unique styling cues (e.g., backgrounds, borders, shadows). Within each module, the algorithm checks whether any child element’s geometry extends beyond its parent’s bounds. To exclude deliberate designs, three cases are filtered out: 1) invisible elements, 2) decorative elements (e.g., with `aria-hidden`), and 3) intentional overflows within scrollable containers. This ensures that only genuine, unhandled overflows are flagged, yielding highly accurate detection.

Low Color Contrast occurs when text or chart elements are insufficiently distinguishable from their background (Fig 4(e)). Accurate assessment demands precise, quantitative color analysis, a task prone to LLM hallucinations. We therefore adopt a rule-based method grounded in the Web Content Accessibility Guidelines (WCAG 2.1). The method first determines effective foreground and background colors. *For text*, the background color is computed via recursive alpha-blending of parent colors up the DOM tree until a fully opaque value is obtained or the root node is reached, correctly handling layered and semi-transparent containers. *For charts*, a hybrid approach is used: 1) retrieving color configurations from charting library APIs (e.g., ECharts, Chart.js) when available; 2) sampling pixels from the borders of localized chart screenshots to infer an equivalent background color, robust to non-uniform patterns. Once the colors are identified, we calculate the contrast ratio (CR) following the WCAG standard, which involves linearizing sRGB values (Eq. 1):

$$C_{\text{linear}} = \begin{cases} \frac{C_{\text{sRGB}}}{12.92} & \text{if } C_{\text{sRGB}} \leq 0.03928 \\ \left(\frac{C_{\text{sRGB}} + 0.055}{1.055}\right)^{2.4} & \text{if } C_{\text{sRGB}} > 0.03928, \end{cases} \quad (1)$$

computing relative luminance (Eq. 2):

$$L = 0.2126 \cdot R_{\text{linear}} + 0.7152 \cdot G_{\text{linear}} + 0.0722 \cdot B_{\text{linear}}, \quad (2)$$

and finding the final ratio of the lighter color (L_1) to the darker (L_2) via $CR = \frac{(L_1 + 0.05)}{(L_2 + 0.05)}$. However, standard WCAG thresholds proved too strict for our setting. To better align with human perception, we manually annotated 2,000 samples and statistically derived thresholds from issue cases: $\theta_{\text{text}} = 1.5$ and $\theta_{\text{chart}} = 1.05$. A defect is flagged when $CR < \theta$.

4.3 CONTENT QUALITY EVALUATION

We design a multi-dimensional automated framework to jointly assess the quality of textual content and the information conveyed by visualizations, as well as their alignment. This addresses the limitation of prior approaches that focused solely on text. We first parse the generated HTML with BeautifulSoup to isolate core content by removing irrelevant elements (e.g., CSS, JavaScript), thereby focusing the LLM evaluator on substantive information. Then, we employ Gemini-2.5-pro under the “LLM-as-a-Judge” paradigm to assess the parsed content across three dimensions. For *Faithfulness*, we further apply a “claim extraction and verification” strategy, in which key claims are first extracted and then evaluated individually to improve accuracy and granularity.

Reasonableness evaluates whether a response contains factual errors, logical fallacies, context inconsistency, or statements that contradict common sense. We define seven specific error types across

	Design Quality				Content Quality		
Dimensions	Missing	Occlusion	Overflow	Contrast	Reasonableness	Comprehensiveness	Faithfulness
Results	97.5%	94.0%	94.5%	95.0%	93.0%	0.912 / 0.892	94.0%

Table 2: **Agreement** between our automated multi-modal evaluation and human assessments. two categories. The first category covers *general content errors*: 1) Factual and Commonsense Inaccuracy. Statements that contradict established facts or widely accepted common sense (e.g., “humans can survive without oxygen”). 2) Context Inconsistency. Internal contradictions within the response, especially misalignment between text and charts. 3) Temporal Inconsistency. Mismatches in time-related statements, or failure to align them with a benchmark date (set to the response generation time), including confusing past events with future predictions. The second category covers *visualization-related errors*: 4) Numerical Implausibility. Chart data that violates basic numerical logic (e.g., percentages in a pie chart not summing to 100%). 5) Unit Inconsistency. Comparing values with different units on the same axis without a clear indication. 6) Missing Essential Information. Omission of key elements such as titles, legends, or axis labels, which hinders unambiguous interpretation. 7) Inappropriate Chart Type. Selecting a visualization form that does not suit the analytical purpose (e.g., using a pie chart for time-series trends). These error types are encoded into a structured prompt for accurate automated evaluation. Because any such error can misinform the user, we apply strict binary scoring: 100 if entirely reasonable, 0 if any error is found.

Comprehensiveness assesses whether a response covering all aspects explicitly requested in the query. Following WildBench (Lin et al., 2025), we first generate a checklist for each query to evaluate coverage and adherence of its corresponding response. Using WildBench’s scoring rubrics and prompt, an LLM evaluator then scores the response’s comprehensiveness based on this checklist. As WildBench adopts a 1–10 scale, we multiply the obtained score by 10 to match other dimensions.

Faithfulness evaluates whether the generated content strictly adheres to the provided references, thereby detecting hallucinations. To achieve a fine-grained evaluation, we employ a “claim extraction and verification” strategy. First, Gemini-2.5-pro extracts factual statements (i.e., claims) from the response. For text, it targets assertions relevant to the user’s query. For charts, it converts structured data into self-contained facts (e.g., transforming a data point on weight into “The iPhone 16 Pro Max weighs 225 grams” rather than simply describing that “The chart shows a weight of 225 grams.”). Each claim is then cross-checked against the references and labeled as *Entailment*, *Contradiction*, and *Neutral*. A claim is considered faithful if labeled *Entailment*. The score of a response equals the percentage of its claims judged as faithful (e.g., 95% yields a score of 95).

4.4 AGREEMENT WITH HUMAN EVALUATION

To validate the effectiveness of our automated evaluation system, we analyzed its agreement with human expert judgments. For each design quality dimension (excluding oversized element due to its unambiguous definition), three experts in data visualization and HTML coding independently annotated 200 randomly selected samples, comprising 100 labeled “correct” and 100 “incorrect” by our system. For content quality, we followed WildBench to correlate *Comprehensiveness* scores with Chatbot Arena Elo ratings. For the other three dimensions, we also randomly collected 100 samples each for human annotation. Specifically, we randomly sampled 10 responses per model for *Reasonableness*, and performed stratified sampling for *Faithfulness* with 40 entailment, 30 neutral, and 30 contradiction. These were annotated by three native speakers capable of reading HTML code. In all human annotation tasks, the majority vote was used as the gold standard. Results are shown in Tab. 2 (SRCC/Pearson for Comprehensiveness, Accuracy for others). Further details on inter-annotator agreement, Precision, and Recall can be found in Appendix M.2. An analysis of potential self-preference biases in the LLM evaluator is provided in Appendix N and Appendix O.

5 EXPERIMENTS

In this section, we evaluate ten leading commercial and open-source LLMs on RAViG-Bench. The results reveal specific weaknesses in current LLMs and provide guidance for future improvements.

5.1 SETUP

System Prompt. For a fair and comprehensive evaluation, we developed a structured system prompt to maximize model potential on this complex task. Inspired by the core principles of Chain-of-Thought (Wei et al., 2022), the task is divided into six sub tasks: 1) Task understanding: ensure the final output is complete HTML code based on the references. 2) Content parsing: extract the topic

Metric	GPT5*	Gemini2.5*	GPT4O*	Qwen3	DeepSeekv3	Doubao1.5*	Claude4*	Llama3	Llama4	Mistral
FSR \uparrow	0.997	0.994	0.973	0.951	0.946	0.968	0.997	0.997	<u>0.888</u>	0.994
DSR \uparrow	0.819	0.685	0.878	0.763	0.814	0.724	0.652	0.882	0.762	<u>0.274</u>
Comprehensiveness \uparrow	93.19	86.18	72.18	78.16	74.71	72.25	80.72	63.70	<u>61.90</u>	66.45
Reasonableness \uparrow	95.18	74.43	54.97	43.37	<u>34.43</u>	68.74	41.47	51.14	49.71	40.97
Faithfulness \uparrow	98.14	92.98	84.04	88.55	84.62	88.81	86.26	<u>78.08</u>	82.67	84.01
Faithfulness (Case) \uparrow	85.34	51.54	38.00	31.41	24.66	57.45	<u>21.25</u>	39.22	42.41	35.50
ECQ \uparrow	95.50	84.53	70.40	70.03	64.59	76.60	69.48	64.31	64.76	<u>63.81</u>
HPS \uparrow	78.24	57.92	61.82	53.44	52.59	55.47	45.29	56.72	49.34	<u>17.50</u>
$VC_{score}\uparrow$	1.482	0.503	-0.377	0.557	0.321	-0.263	0.416	<u>-1.127</u>	-0.984	-0.528
$HPS_{vcc}\uparrow$	78.24	57.92	54.82	53.44	52.59	51.10	45.29	37.55	34.78	<u>14.72</u>

Table 3: **Comprehensive performance of all models.** For each metric (row), the **best** result is highlighted in bold and the worst result is underlined. *denotes commercial closed-source LLMs.

and data from the reference document. 3) Chart selection: choose a suitable visualization format. 4) Layout design: plan the head and body structure, including styles and visualization containers. 5) HTML coding: write correct HTML code containing the chosen visualizations. 6) Validation and refinement: check that the output follows the reference documents and renders correctly, and output only the HTML code. Due to confidentiality constraints, we do not provide the full prompt. However, this omission does not affect the validity of our results, as all models were tested under the same controlled conditions. Examples of visually-rich responses are provided in Appendix F.

Complexity-based Score Correction. To mitigate bias from overly simple designs, we apply a complexity-based correction factor, γ_{vc} . This factor is derived from a visual richness score (VC_{score}), calculated as follows:

$$VC_{score} = 0.7z_{module} + 0.2z_{chart} + 0.1z_{table}. \quad (3)$$

Here, each z_m is the z-score for a specific element type m (distinct visual modules, charts, or tables), defined as $z_m = (x_m - \mu_m)/\sigma_m$. In this formula, x_m is the average element count of the model, and μ_m, σ_m are the cross-model mean and standard deviation. The weighting scheme emphasizes visual modules, which are strong indicators of design complexity. The final correction factor is then defined as:

$$\gamma_{vc} = \min(1.0, 1 + \alpha \cdot VC_{score}), \quad \text{with } \alpha = 0.3. \quad (4)$$

Outputs at or above the average complexity ($VC_{score} \geq 0$) receive no adjustment, as existing design defect detection sufficiently reflects their quality.

Metrics. We define five metrics to comprehensively and hierarchically evaluate model performance. “*Functional Success Rate (FSR)*” measures the proportion of all samples that pass the functionality validation. “*Design Success Rate (DSR)*” measures the proportion of functionality-passed samples that also pass the design quality assessment. “*Effective Content Quality (ECQ)*” measures the mean content quality score over effective samples that pass both prior checks, where the score is the average of three content dimensions. “*Holistic Performance Score (HPS)*” measures the mean content quality score over all samples, assigning 0 to those failing the first two checks. “*HPS_{vcc}*” is the HPS multiplied by the correction factor γ_{vc} . In addition, we report the per-dimension content quality scores of effective responses. For *Faithfulness*, we also report case-level scores, in which any response containing an issue claim is assigned a score of 0.

Models. We evaluate ten leading commercial and open-source LLMs, including Claude-Opus-4-20250514 (Claude4) (Anthropic, 2025), GPT-4o-2024-11-20 (GPT4O) (Hurst et al., 2024), GPT-5-2025-08-07 (GPT5) (OpenAI, 2025a), Gemini-2.5-pro (Gemini2.5) (Comanici et al., 2025), DeepSeek-V3-0324(DeepSeekv3) (Liu et al., 2024a), Qwen3-235B-A22B-Instruct-2507(Qwen3) (Team, 2025), Doubao-1.5-Vision-Pro (Doubao1.5) (Guo et al., 2025a), Llama-4-Scout-17B-16E-Instruct (Llama4) (Meta AI, 2025), Llama-3.3-70B-Instruct (Llama3) (Dubey et al., 2024), and Mistral-Small-24B-Instruct-2501 (Mistral) (Mistral AI, 2025). For brevity, we use these abbreviations in the tables, figures, and some analyses throughout this section. Specifically, we set the hyperparameter temperature to 0 (except for GPT-5, which only accepts the default temperature=1.0). For the reasoning models, we set the thinking budget to 1024. We further provide in Appendix P an analysis of model behavior under increased generation diversity (temperature=0.7) and additional evidence of the robustness of our evaluation framework.

432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

5.2 MAIN RESULTS

The overall results in Tab. 3 show that while both commercial and open-source LLMs can almost perfectly produce syntactically correct and functional outputs, design quality remains the primary bottleneck. Across models, Design Success Rates are notably lower than Functional Success Rates, with values mainly ranging from 60% to 80%. Considering the content quality, GPT5 performs best on almost all metrics, followed by Gemini2.5 and Doubao1.5, whereas Mistral records the weakest performance. The holistic performance score (HPS) reflects overall ability without adjusting for design complexity, with GPT5 leading at 78.24, well ahead of the second-best model and indicating substantial room for improvement in this task. Applying the complexity-based correction (HPS_{vcc}) provides a more realistic assessment: it reveals that some models, such as Llama3, achieve inflated design scores through overly simplistic layouts, whereas GPT5 and Gemini2.5 maintain relatively strong performance after correction. These findings suggest that future progress in RAViG will depend less on basic renderability, and more on advancing the integration of rich, accurate, reliable content with high-quality visual designs. Some content-issue cases are provided in Appendix I. Additional experimental results and analysis are provided in Appendix Q.

5.3 IN-DEPTH ANALYSIS

The root of LLM design failures lies in deficiencies in specific capabilities rather than an overall lack of design sense. Fig. 5 shows the category-level counts of design issues for the five most error-prone models. *Low Color Contrast* (Fig. 4 (e)) is particularly salient, suggesting that achieving adequate contrast requires understanding color semantics, the interaction between foreground and background, and the visual impact of color combinations, beyond syntax validity. *Oversized Elements* (Fig. 4 (a)) in Gemini2.5 are mainly caused by faulty CSS dependencies for SVG icons. Although the system prompt provides a valid dependency, the model, possibly due to over-confidence or pre-training bias, still uses a specific outdated version, leading to rendering failures. Doubao1.5 frequently produces *Overflow* (Fig. 4 (d)), indicating a weak grasp of global spatial layout.

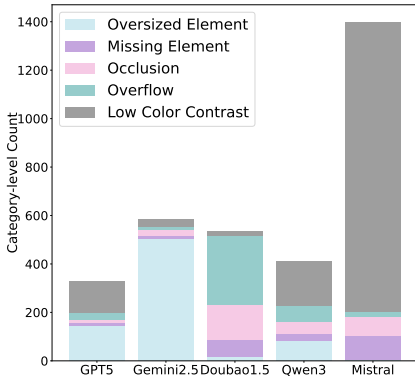


Figure 5: Distribution of design issues.

Some LLMs struggle to maintain contextual consistency in visually-rich HTML generation. An analysis of failures on the reasonableness dimension reveals that the primary issues are contextual inconsistency (Fig. 26), tense confusion (Fig. 28, left), and missing chart elements (Fig. 28, right). Contextual inconsistency is the most prevalent, and most often manifests as chart-text inconsistency (Fig. 25). These may stem from the inherent complexity of HTML generation: unlike linear plain-text generation, HTML output consists of non-contiguous segments such as text, tags, and script code, requiring the model to sustain long-range contextual memory and chart-text alignment across components, while the separated content structure further increases the difficulty of maintaining consistency.

Most LLMs show vulnerability in faithfulness, as high accuracy on individual facts does not ensure a fully trustworthy response. This gap emerges when comparing claim-level scores with the stricter case-level scores. Tab. 3 shows that most models exceed 80% faithfulness at the claim-level but drop sharply at the case-level; for instance, Claude4 falls from 86.3 to 21.3, indicating that nearly 80% of its responses contain factual errors (Fig. 30). Since a single mistake can undermine user trust, the strong case-level performance of models like GPT5 and Doubao1.5 signals a higher capacity for generating holistically reliable content, an essential step toward real-world applicability.

6 CONCLUSION

In conclusion, we introduce RAViG-Bench, the first benchmark with an automatic multi-modal evaluation framework for comprehensive assessment of LLM performance on RAViG. We adopt HTML as the representation, but the benchmark is not tied to HTML or specific visual elements. The dataset, evaluation framework, algorithms, and criteria are format agnostic; adapting to other structured formats only requires modifying the input and output processing modules. Experiments show that even state-of-the-art LLMs still have substantial room for improvement. While our work provides the first systematic capability to objectively detect readability-related defects and perform

486 basic visual complexity correction, it does not yet address subjective aesthetics or incorporate richer
487 factors such as typography and layout. In addition, aesthetic judgments related to “over-design” are
488 excluded, so visually richer yet high-quality outputs are considered evidence of model capability.
489 Addressing these limitations offers promising directions for future work to incorporate user-centered
490 aesthetic evaluation and nuanced visual refinement, extending the benchmark from assessing basic
491 usability to evaluating outputs for visual appeal, accessibility, and overall user experience.

492 493 ETHICS STATEMENT 494

495 All data used in this benchmark are either collected from the open-source Natural Questions (NQ)
496 dataset or generated by large language models, and do not contain personally identifiable informa-
497 tion or sensitive personal data. Licenses for all third-party datasets and open-source models used in
498 this work are listed in Appendix C. For any human annotations involved, annotators were informed
499 of the research purpose and consented to participate; the annotation process posed no physical, psy-
500 chological, or legal risks. To mitigate bias, during the final human review for query selection and
501 reference document annotation, we filtered out content with overt discrimination, ethical concerns,
502 or harmful elements. We explicitly prohibit the use of this benchmark or any derived models in ap-
503 plications intended to discriminate against individuals or groups, or in any contexts that could cause
504 harm. The full benchmark, along with associated code, will be released upon acceptance under a
505 license that permits research use while prohibiting unlawful or malicious purposes. No part of this
506 work violates privacy rights, confidentiality agreements, or applicable laws and regulations.

507 508 REPRODUCIBILITY STATEMENT 509

510 We have taken multiple steps to ensure the reproducibility of our work. For dataset construction, we
511 provide detailed selection criteria, categorization scheme, and the end-to-end pipeline for assembling
512 both real-world and synthetic queries in Sec. 3.1 and Appendix E. For all LLM-based evaluation
513 methods, we report in Sec. 4.2 and Sec. 4.3 the full set of dimensions and structural organization
514 contained in the prompts. For our rule-based baselines, Sec. 4.2 include the complete algorithmic
515 workflows and relevant hyperparameter settings. We commit to publicly releasing the full dataset,
516 all prompts, and code upon acceptance. These resources, together with the descriptions in the main
517 text and supplementary materials, will enable other researchers to fully reproduce our results.

518 519 REFERENCES

- 520 Anthropic. Claude opus 4. <https://www.anthropic.com/claude/opus>, May 2025.
521
- 522 Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in
523 retrieval-augmented generation. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan
524 (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Con-
525 ference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium
526 on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancou-
527 ver, Canada*, pp. 17754–17762. AAAI Press, 2024. doi: 10.1609/AAAI.V38I16.29728. URL
528 <https://doi.org/10.1609/aaai.v38i16.29728>.
- 529 Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang
530 Liu, Zhou Yu, and Dilek Hakkani-Tur. Weakly supervised data augmentation through prompting
531 for dialogue understanding. *CoRR*, abs/2210.14169, 2022. doi: 10.48550/ARXIV.2210.14169.
532 URL <https://doi.org/10.48550/arXiv.2210.14169>.
- 533 Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. Viseval: A benchmark for data
534 visualization in the era of large language models. *IEEE Trans. Vis. Comput. Graph.*, 31(1):
535 1301–1311, 2025. doi: 10.1109/TVCG.2024.3456320. URL [https://doi.org/10.1109/
536 TVCG.2024.3456320](https://doi.org/10.1109/TVCG.2024.3456320).
- 537
- 538 Liying Cheng, Xingxuan Li, and Lidong Bing. Is GPT-4 a good data analyst? In Houda Bouamor,
539 Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics:
EMNLP 2023, Singapore, December 6-10, 2023*, pp. 9496–9514. Association for Computational

- 540 Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.637. URL <https://doi.org/10.18653/v1/2023.findings-emnlp.637>.
- 541
- 542
- 543 Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li,
544 Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Sto-
545 ica. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first*
546 *International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.
547 OpenReview.net, 2024. URL <https://openreview.net/forum?id=3MW8GKNyzI>.
- 548 Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit
549 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the
550 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-
551 bilities. *arXiv preprint arXiv:2507.06261*, 2025.
- 552 Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. PDF-MVQA: A
553 dataset for multimodal information retrieval in pdf-based visual question answering. *CoRR*,
554 abs/2404.12720, 2024. doi: 10.48550/ARXIV.2404.12720. URL [https://doi.org/10.](https://doi.org/10.48550/arXiv.2404.12720)
555 [48550/arXiv.2404.12720](https://doi.org/10.48550/arXiv.2404.12720).
- 556 Kuicai Dong, Yujing Chang, Derrick-Goh-Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. Mm-
557 docir: Benchmarking multi-modal retrieval for long documents. *CoRR*, abs/2501.08828, 2025.
558 doi: 10.48550/ARXIV.2501.08828. URL [https://doi.org/10.48550/arXiv.2501.](https://doi.org/10.48550/arXiv.2501.08828)
559 [08828](https://doi.org/10.48550/arXiv.2501.08828).
- 560 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
561 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
562 *arXiv e-prints*, pp. arXiv–2407, 2024.
- 563
- 564 Shahul ES, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation
565 of retrieval augmented generation. In Nikolaos Aletras and Orphée De Clercq (eds.), *Proceedings*
566 *of the 18th Conference of the European Chapter of the Association for Computational Linguistics,*
567 *EACL 2024 - System Demonstrations, St. Julians, Malta, March 17-22, 2024*, pp. 150–158. Asso-
568 ciation for Computational Linguistics, 2024. URL [https://aclanthology.org/2024.](https://aclanthology.org/2024.eacl-demo.16)
569 [eacl-demo.16](https://aclanthology.org/2024.eacl-demo.16).
- 570 Siwei Fu, Kai Xiong, Xiaodong Ge, Siliang Tang, Wei Chen, and Yingcai Wu. Quda: Natu-
571 ral language queries for visual data analytics. *CoRR*, abs/2005.03257, 2020. URL [https:](https://arxiv.org/abs/2005.03257)
572 [//arxiv.org/abs/2005.03257](https://arxiv.org/abs/2005.03257).
- 573
- 574 Aoran Gan, Hao Yu, Kai Zhang, Qi Liu, Wenyu Yan, Zhenya Huang, Shiwei Tong, and Guoping Hu.
575 Retrieval augmented generation evaluation in the era of large language models: A comprehensive
576 survey. *CoRR*, abs/2504.14891, 2025. doi: 10.48550/ARXIV.2504.14891. URL [https://](https://doi.org/10.48550/arXiv.2504.14891)
577 doi.org/10.48550/arXiv.2504.14891.
- 578 Google. Deep research. Official Gemini Overview, September 2025. URL [https://gemini.](https://gemini.google.com/overview/deep-research/)
579 [google.com/overview/deep-research/](https://gemini.google.com/overview/deep-research/).
- 580
- 581 Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang,
582 Jiayu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*,
583 2025a.
- 584 Meng-Hao Guo, Jiajun Xu, Yi Zhang, Jiayi Song, Haoyang Peng, Yi-Xuan Deng, Xinzhi Dong,
585 Kiyohiro Nakayama, Zhengyang Geng, Chen Wang, Bolin Ni, Guo-Wei Yang, Yongming
586 Rao, Houwen Peng, Han Hu, Gordon Wetzstein, and Shi-Min Hu. R-bench: Graduate-
587 level multi-disciplinary benchmarks for LLM & MLLM complex reasoning evaluation. *CoRR*,
588 abs/2505.02018, 2025b. doi: 10.48550/ARXIV.2505.02018. URL [https://doi.org/10.](https://doi.org/10.48550/arXiv.2505.02018)
589 [48550/arXiv.2505.02018](https://doi.org/10.48550/arXiv.2505.02018).
- 590 Kevin Zeng Hu, Snehal Kumar (Neil) S. Gaikwad, Madelon Hulsebos, Michiel A. Bakker, Emanuel
591 Zraggen, César A. Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayan, and Çağatay Demiralp.
592 Viznet: Towards A large-scale visualization learning and benchmarking repository. In
593 Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (eds.), *Proceed-*
ings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow,

- 594 *Scotland, UK, May 04-09, 2019*, pp. 662. ACM, 2019. doi: 10.1145/3290605.3300892. URL
595 <https://doi.org/10.1145/3290605.3300892>.
596
- 597 Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun
598 Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. In
599 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore,*
600 *April 24-28, 2025*. OpenReview.net, 2025. URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=Usklli4gMc)
601 [Usklli4gMc](https://openreview.net/forum?id=Usklli4gMc).
- 602 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-
603 trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint*
604 *arXiv:2410.21276*, 2024.
- 605 Sebastian Antony Joseph, Syed Murtaza Husain, Stella S. R. Offner, Stéphanie Juneau, Paul Torrey,
606 Adam S. Bolton, Juan P. Farias, Niall Gaffney, Greg Durrett, and Junyi Jessy Li. Astrovisbench: A
607 code benchmark for scientific computing and visualization in astronomy. *CoRR*, abs/2505.20538,
608 2025. doi: 10.48550/ARXIV.2505.20538. URL [https://doi.org/10.48550/arXiv.](https://doi.org/10.48550/arXiv.2505.20538)
609 [2505.20538](https://doi.org/10.48550/arXiv.2505.20538).
- 610 Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, R. Chulaka Gunasekara, Young-Suk Lee, Lucian
611 Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. MTRAG: A multi-
612 turn conversational benchmark for evaluating retrieval-augmented generation systems. *CoRR*,
613 abs/2501.03468, 2025. doi: 10.48550/ARXIV.2501.03468. URL [https://doi.org/10.](https://doi.org/10.48550/arXiv.2501.03468)
614 [48550/arXiv.2501.03468](https://doi.org/10.48550/arXiv.2501.03468).
- 615 Hyung-Kwon Ko, Hyeon Jeon, Gwanmo Park, Dae Hyun Kim, Nam Wook Kim, Juho Kim, and
616 Jinwook Seo. Natural language dataset generation framework for visualizations powered by large
617 language models. In Florian ‘Floyd’ Mueller, Penny Kyburz, Julie R. Williamson, Corina Sas,
618 Max L. Wilson, Phoebe O. Toups Dugas, and Irina Shklovski (eds.), *Proceedings of the CHI*
619 *Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-*
620 *16, 2024*, pp. 843:1–843:22. ACM, 2024. doi: 10.1145/3613904.3642943. URL [https://](https://doi.org/10.1145/3613904.3642943)
621 doi.org/10.1145/3613904.3642943.
622
- 623 Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris
624 Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion
625 Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav
626 Petrov. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput.*
627 *Linguistics*, 7:452–466, 2019. doi: 10.1162/TACL_A_00276. URL [https://doi.org/](https://doi.org/10.1162/tACL_a_00276)
628 [10.1162/tACL_a_00276](https://doi.org/10.1162/tACL_a_00276).
- 629 Guozheng Li, Xinyu Wang, Gerile Aodeng, Shunyuan Zheng, Yu Zhang, Chuangxin Ou, Song
630 Wang, and Chi Harold Liu. Visualization generation with large language models: An evaluation.
631 *CoRR*, abs/2401.11255, 2024. doi: 10.48550/ARXIV.2401.11255. URL [https://doi.org/](https://doi.org/10.48550/arXiv.2401.11255)
632 [10.48550/arXiv.2401.11255](https://doi.org/10.48550/arXiv.2401.11255).
- 633 Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, CG Ishaan Gulrajani, P Liang, and
634 TB Hashimoto. AlpacaEval: an automatic evaluator of instruction-following models (2023). URL
635 https://github.com/tatsu-lab/alpaca_eval, 2023.
636
- 637 HyeonSeok Lim, Dongjae Shin, Seohyun Song, Inho Won, Minjun Kim, Junghun Yuk, Haneol
638 Jang, and KyungTae Lim. Vlr-bench: Multilingual benchmark dataset for vision-language re-
639 trieval augmented generation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, HEND
640 Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st In-*
641 *ternational Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, Jan-*
642 *uary 19-24, 2025*, pp. 6150–6168. Association for Computational Linguistics, 2025. URL
643 <https://aclanthology.org/2025.coling-main.411/>.
- 644 Bill Yuchen Lin, Yuntian Deng, Khyathi Raghavi Chandu, Abhilasha Ravichander, Valentina Py-
645 atkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with chal-
646 lenging tasks from real users in the wild. In *The Thirteenth International Conference on Learn-*
647 *ing Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL
<https://openreview.net/forum?id=MKEHCx25xp>.

- 648 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
649 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*
650 *arXiv:2412.19437*, 2024a.
- 651
- 652 Jiawei Liu, Jia Le Tian, Vijay Daita, Yuxiang Wei, Yifeng Ding, Yuhan Katherine Wang, Jun
653 Yang, and Lingming Zhang. Repoqa: Evaluating long context code understanding. *CoRR*,
654 abs/2406.06025, 2024b. doi: 10.48550/ARXIV.2406.06025. URL [https://doi.org/10.](https://doi.org/10.48550/arXiv.2406.06025)
655 [48550/arXiv.2406.06025](https://doi.org/10.48550/arXiv.2406.06025).
- 656 Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and
657 Xu Sun. RECALL: A benchmark for llms robustness against external counterfactual knowl-
658 edge. *CoRR*, abs/2311.08147, 2023. doi: 10.48550/ARXIV.2311.08147. URL [https:](https://doi.org/10.48550/arXiv.2311.08147)
659 [//doi.org/10.48550/arXiv.2311.08147](https://doi.org/10.48550/arXiv.2311.08147).
- 660 Tianqi Luo, Chuhan Huang, Leixian Shen, Boyan Li, Shuyu Shen, Wei Zeng, Nan Tang, and Yuyu
661 Luo. nvbench 2.0: A benchmark for natural language to visualization under ambiguity. *CoRR*,
662 abs/2503.12880, 2025. doi: 10.48550/ARXIV.2503.12880. URL [https://doi.org/10.](https://doi.org/10.48550/arXiv.2503.12880)
663 [48550/arXiv.2503.12880](https://doi.org/10.48550/arXiv.2503.12880).
- 664
- 665 Yuyu Luo, Jiawei Tang, and Guoliang Li. nvbench: A large-scale synthesized dataset for cross-
666 domain natural language to visualization task. *CoRR*, abs/2112.12926, 2021a. URL [https:](https://arxiv.org/abs/2112.12926)
667 [//arxiv.org/abs/2112.12926](https://arxiv.org/abs/2112.12926).
- 668 Yuyu Luo, Nan Tang, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. Synthesizing nat-
669 ural language to visualization (NL2VIS) benchmarks from NL2SQL benchmarks. In Guoliang
670 Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava (eds.), *SIGMOD '21: International Con-*
671 *ference on Management of Data, Virtual Event, China, June 20-25, 2021*, pp. 1235–1247. ACM,
672 2021b. doi: 10.1145/3448016.3457261. URL [https://doi.org/10.1145/3448016.](https://doi.org/10.1145/3448016.3457261)
673 [3457261](https://doi.org/10.1145/3448016.3457261).
- 674 Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu,
675 Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang,
676 Yixin Cao, and Aixin Sun. MMLONGBENCH-DOC: benchmarking long-context docu-
677 ment understanding with visualizations. In Amir Globersons, Lester Mackey, Danielle
678 Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances*
679 *in Neural Information Processing Systems 38: Annual Conference on Neural Informa-*
680 *tion Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 -*
681 *15, 2024*, 2024. URL [http://papers.nips.cc/paper_files/paper/2024/](http://papers.nips.cc/paper_files/paper/2024/hash/ae0e43289bffe0c1fa34633fc608e92-Abstract-Datasets_and_Benchmarks_Track.html)
682 [hash/ae0e43289bffe0c1fa34633fc608e92-Abstract-Datasets_and_](http://papers.nips.cc/paper_files/paper/2024/hash/ae0e43289bffe0c1fa34633fc608e92-Abstract-Datasets_and_Benchmarks_Track.html)
683 [Benchmarks_Track.html](http://papers.nips.cc/paper_files/paper/2024/hash/ae0e43289bffe0c1fa34633fc608e92-Abstract-Datasets_and_Benchmarks_Track.html).
- 684 Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A
685 benchmark for question answering about charts with visual and logical reasoning. In Smaranda
686 Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Com-*
687 *putational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 2263–2279. Associa-
688 tion for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.177. URL
689 <https://doi.org/10.18653/v1/2022.findings-acl.177>.
- 690
- 691 Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal ai in-
692 novation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>,
693 April 2025.
- 694
- 695 Mistral AI. Mistral small 3. <https://mistral.ai/news/mistral-small-3>, January
696 2025.
- 697
- 698 Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and
699 Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented
700 language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of*
701 *the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*
Papers), *ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 10862–10878. Association for
Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.585. URL [https://](https://doi.org/10.18653/v1/2024.acl-long.585)
doi.org/10.18653/v1/2024.acl-long.585.

- 702 OpenAI. Introducing gpt-5. [https://openai.com/zh-Hans-CN/index/
703 introducing-gpt-5/](https://openai.com/zh-Hans-CN/index/introducing-gpt-5/), August 2025a.
704
- 705 OpenAI. Introducing deep research. [https://openai.com/zh-Hant/index/
706 introducing-deep-research/](https://openai.com/zh-Hant/index/introducing-deep-research/), 2025b. Accessed: 2025-11-18.
- 707 Bo Pan, Yixiao Fu, Ke Wang, Junyu Lu, Lunke Pan, Ziyang Qian, Yuhan Chen, Guoliang Wang,
708 Yitao Zhou, Li Zheng, Yinghao Tang, Zhen Wen, Yuchen Wu, Junhua Lu, Biao Zhu, Minfeng
709 Zhu, Bo Zhang, and Wei Chen. Vis-shepherd: Constructing critic for llm-based data visualization
710 generation. *CoRR*, abs/2506.13326, 2025. doi: 10.48550/ARXIV.2506.13326. URL [https://
711 doi.org/10.48550/arXiv.2506.13326](https://doi.org/10.48550/arXiv.2506.13326).
- 712 Perplexity AI, Inc. Perplexity. Official Website, 2025. URL [https://www.perplexity.ai/.
713 Accessed: 2025-09-01](https://www.perplexity.ai/).
- 714 Luca Podo, Muhammad Ishmal, and Marco Angelini. Vi(e)va llm! A conceptual stack for evaluating
715 and interpreting generative ai-based visualizations. *CoRR*, abs/2402.02167, 2024. doi: 10.48550/
716 ARXIV.2402.02167. URL <https://doi.org/10.48550/arXiv.2402.02167>.
- 717 Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: an automated eval-
718 uation framework for retrieval-augmented generation systems. In Kevin Duh, Helena Gómez-
719 Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American
720 Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol-
721 ume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 338–354. Asso-
722 ciation for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.20. URL
723 <https://doi.org/10.18653/v1/2024.naacl-long.20>.
- 724 Gaurav Sahu, Abhay Puri, Juan Rodriguez, Amirhossein Abaskohi, Mohammad Chegini, Alexan-
725 dre Drouin, Perouz Taslakian, Valentina Zantedeschi, Alexandre Lacoste, David Vazquez, et al.
726 Insightbench: Evaluating business analytics agents through multi-step insight generation. *arXiv
727 preprint arXiv:2407.06423*, 2024.
- 728 Shanghai Mita Network Technology Co., Ltd. Metaso ai search. Official Website, 2025. URL
729 <https://metaso.cn/>. Accessed: 2025-09-01.
- 730 Yuanfeng Song, Xuefang Zhao, and Raymond Chi-Wing Wong. Marrying dialogue systems with
731 data visualization: Interactive data visualization generation from natural language conversations.
732 In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD
733 Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August
734 25-29, 2024*, pp. 2733–2744. ACM, 2024. doi: 10.1145/3637528.3671935. URL [https://
735 doi.org/10.1145/3637528.3671935](https://doi.org/10.1145/3637528.3671935).
- 736 Ionut-Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià
737 de Gispert. Garage: A benchmark with grounding annotations for RAG evaluation. In Wanx-
738 iang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Find-
739 ings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 -
740 August 1, 2025*, pp. 17030–17049. Association for Computational Linguistics, 2025. URL
741 <https://aclanthology.org/2025.findings-acl.875/>.
- 742 Arjun Srinivasan, Nikhila Nyapathy, Bongshin Lee, Steven Mark Drucker, and John T. Stasko.
743 Collecting and characterizing natural language utterances for specifying data visualizations. In
744 Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and
745 Steven Mark Drucker (eds.), *CHI '21: CHI Conference on Human Factors in Computing Sys-
746 tems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pp. 464:1–464:10. ACM, 2021. doi:
747 10.1145/3411764.3445400. URL <https://doi.org/10.1145/3411764.3445400>.
- 748 Jan Strich, Enes Kutay Isgorur, Maximilian Trescher, Chris Biemann, and Martin Semmann. T²-
749 ragbench: Text-and-table benchmark for evaluating retrieval-augmented generation. *CoRR*,
750 abs/2506.12071, 2025. doi: 10.48550/ARXIV.2506.12071. URL [https://doi.org/10.
751 48550/arXiv.2506.12071](https://doi.org/10.48550/arXiv.2506.12071).
- 752 Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

- 756 Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li,
757 Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, et al. Tongyi deepresearch technical
758 report. *arXiv preprint arXiv:2510.24701*, 2025.
759
- 760 Pere-Pau Vázquez. Are llms ready for visualization? In *17th IEEE Pacific Visualiza-*
761 *tion Conference, PacificVis 2024, Tokyo, Japan, April 23-26, 2024*, pp. 343–352. IEEE,
762 2024. doi: 10.1109/PACIFICVIS60374.2024.00049. URL [https://doi.org/10.1109/
763 PacificVis60374.2024.00049](https://doi.org/10.1109/PacificVis60374.2024.00049).
- 764 Zora Zhiruo Wang, Akari Asai, Xinyan Velocity Yu, Frank F. Xu, Yiqing Xie, Graham Neubig, and
765 Daniel Fried. Coderag-bench: Can retrieval augment code generation? In Luis Chiruzzo, Alan
766 Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL*
767 *2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 3199–3214. Association
768 for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.176. URL
769 <https://doi.org/10.18653/v1/2025.findings-naacl.176>.
- 770
- 771 Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay,
772 and Leonid Karlinsky. REAL-MM-RAG: A real-world multi-modal retrieval benchmark. In
773 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Pro-*
774 *ceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume*
775 *1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 31660–31683. Asso-
776 ciation for Computational Linguistics, 2025. URL [https://aclanthology.org/2025.
777 acl-long.1528/](https://aclanthology.org/2025.acl-long.1528/).
- 778 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi,
779 Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language
780 models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh
781 (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural*
782 *Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - De-*
783 *cember 9, 2022, 2022*. URL [http://papers.nips.cc/paper_files/paper/2022/
784 hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html).
- 785 Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. Visual-rag: Benchmarking
786 text-to-image retrieval augmented generation for visual knowledge intensive queries. *CoRR*,
787 abs/2502.16636, 2025. doi: 10.48550/ARXIV.2502.16636. URL [https://doi.org/10.
788 48550/arXiv.2502.16636](https://doi.org/10.48550/arXiv.2502.16636).
- 789
- 790 Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary,
791 Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi
792 Wang, Yifan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen,
793 Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Scott Yih, and Xin
794 Dong. CRAG - comprehensive RAG benchmark. In Amir Globersons, Lester Mackey,
795 Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.),
796 *Advances in Neural Information Processing Systems 38: Annual Conference on Neural In-*
797 *formation Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10*
798 *- 15, 2024, 2024a*. URL [http://papers.nips.cc/paper_files/paper/2024/
799 hash/1435d2d0fca85a84d83ddcb754f58c29-Abstract-Datasets_and_
800 Benchmarks_Track.html](http://papers.nips.cc/paper_files/paper/2024/hash/1435d2d0fca85a84d83ddcb754f58c29-Abstract-Datasets_and_Benchmarks_Track.html).
- 801
- 801 Xudong Yang, Yifan Wu, Yizhang Zhu, Nan Tang, and Yuyu Luo. Askchart: Universal chart under-
802 standing through textual enhancement. *CoRR*, abs/2412.19146, 2024b. doi: 10.48550/ARXIV.
803 2412.19146. URL <https://doi.org/10.48550/arXiv.2412.19146>.
- 804
- 804 Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woo-Myoung Park. Gpt3mix:
805 Leveraging large-scale language models for text augmentation. In Marie-Francine Moens, Xu-
806 anjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for*
807 *Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Repub-*
808 *lic, 16-20 November, 2021*, pp. 2225–2239. Association for Computational Linguistics, 2021.
809 doi: 10.18653/V1/2021.FINDINGS-EMNLP.192. URL [https://doi.org/10.18653/
v1/2021.findings-emnlp.192](https://doi.org/10.18653/v1/2021.findings-emnlp.192).

- 810 Qinhan Yu, Zhiyou Xiao, Binghui Li, Zhengren Wang, Chong Chen, and Wentao Zhang. Mramg-
811 bench: A comprehensive benchmark for advancing multimodal retrieval-augmented multimodal
812 generation. In Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman,
813 and Suzan Verberne (eds.), *Proceedings of the 48th International ACM SIGIR Conference on*
814 *Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*,
815 pp. 3616–3626. ACM, 2025. doi: 10.1145/3726302.3730288. URL [https://doi.org/10.](https://doi.org/10.1145/3726302.3730288)
816 [1145/3726302.3730288](https://doi.org/10.1145/3726302.3730288).
- 817 Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian
818 Gehrmann. Synthbio: A case study in human-ai collaborative curation of text datasets. *CoRR*,
819 abs/2111.06467, 2021. URL <https://arxiv.org/abs/2111.06467>.
- 820
821 Zaifu Zhan, Jun Wang, Shuang Zhou, Jiawen Deng, and Rui Zhang. MMRAG: multi-mode retrieval-
822 augmented generation with large language models for biomedical in-context learning. *CoRR*,
823 abs/2502.15954, 2025. doi: 10.48550/ARXIV.2502.15954. URL [https://doi.org/10.](https://doi.org/10.48550/arXiv.2502.15954)
824 [48550/arXiv.2502.15954](https://doi.org/10.48550/arXiv.2502.15954).
- 825 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao
826 Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez,
827 and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh,
828 Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.),
829 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural In-*
830 *formation Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10*
831 *- 16, 2023*, 2023. URL [http://papers.nips.cc/paper_files/paper/2023/](http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html)
832 [hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_](http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html)
833 [Benchmarks.html](http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html).
- 834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

A RELATED WORK

Data Construction with Real-World Use Cases and LLMs. To ensure alignment with real-world LLM usage scenarios, the majority of our benchmark data is sourced from a real-world dataset Natural Questions (Kwiatkowski et al., 2019). Works (Chiang et al., 2024; Li et al., 2023) like WildBench (Lin et al., 2025) also leverage real-world data, underscoring its importance in LLM evaluation. Additionally, several open-ended benchmarks rely on human experts or crowdsourcing workers for data creation (Srinivasan et al., 2021; Guo et al., 2025b). On the other hand, LLMs have been widely used to facilitate data synthesis, enhancing data diversity (Yuan et al., 2021; Yang et al., 2024b; Yoo et al., 2021; Chen et al., 2022; Luo et al., 2021b). For example, NL2VL (Ko et al., 2024) employs LLM-generated samples followed by human filtering, establishing a new paradigm that significantly improves data quality. Following this trend, we leverage LLMs to augment our dataset, addressing certain limitations of the filtered real-world data.

Benchmarks for Retrieval-Augmented Generation (RAG). Significant progress has been made in the evaluation of Retrieval-Augmented Generation (RAG) (Gan et al., 2025). Existing research predominantly focuses on the quality of text outputs, assessing aspects like accuracy, faithfulness, and hallucination (Sorodoc et al., 2025; Yang et al., 2024a; Saad-Falcon et al., 2024; Liu et al., 2023; Niu et al., 2024; Katsis et al., 2025). Evaluation paradigms often achieve in-depth analysis by independently assessing key components such as the retriever and generator (LLM), designing complex scenarios, or introducing novel metrics. For instance, RAGAS (ES et al., 2024) explores automated, ground-truth-free evaluation using LLMs, while RGB (Chen et al., 2024) tests model robustness by injecting noise. Another line of research extends the evaluation scope to multi-modal understanding (Zhan et al., 2025; Strich et al., 2025; Masry et al., 2022; Dong et al., 2025; Ma et al., 2024; Ding et al., 2024; Lim et al., 2025; Wasserman et al., 2025; Hu et al., 2025). Visual-RAG (Wu et al., 2025), for example, focuses on a model’s ability to retrieve and comprehend image information. However, such work focuses on models’ *understanding* of multimodal content; the capacity for LLMs to *generate* visually rich responses remains largely unexplored. Among these, MRAMG-Bench (Yu et al., 2025) is most similar to our setting, but despite involving text-image layout, it is essentially a text-only generation task, as its answers combine model-generated text with images retrieved from source material. In contrast, our benchmark requires models to autonomously design and produce visually-rich answers. Furthermore, some studies have assessed the code generation capabilities of RAG-enhanced models (Liu et al., 2024b; Wang et al., 2025). However, their emphasis is on understanding code context and generating correct outputs, with the code itself being the final object of evaluation. In summary, there is currently no benchmark that can perform a comprehensive, end-to-end evaluation of an LLM’s ability to synthesize information from multiple documents and generate appropriate visually rich answers in response to open-ended analytical queries.

Benchmarks for Natural Language to Visualization Generation. Recent years have seen rapid progress in using LLMs for natural language to visualization (NL2VIS) tasks, leading to various benchmarks and datasets to systematically measure and advance model capabilities in this area (Cheng et al., 2023; Vázquez, 2024; Li et al., 2024; Luo et al., 2021a; Hu et al., 2019; Fu et al., 2020). For example, nvBench 2.0 (Luo et al., 2025) provides support for complex and ambiguous queries, paired with disambiguation reasoning paths. VisEval (Chen et al., 2025) introduces a high-quality dataset and an automated evaluation framework, encompassing validity, legality, and readability of generated code and charts. To enable more fine-grained evaluation and feedback, EvaLLM (Podo et al., 2024) decomposes the evaluation into semantic and procedural components, and implements this through an evaluation platform that supports both automated and manual assessment. Vis-Shepherd (Pan et al., 2025) further introduces a multimodal LLM-based critic and a specialized critique dataset to enable automated feedback. In addition, dialogue-based datasets (Song et al., 2024), human-collected query corpora (Srinivasan et al., 2021), and domain-specific benchmarks (Joseph et al., 2025) further enrich the evaluation landscape. However, these benchmarks mainly focus on direct conversion from structured queries (typically including both the database and a data-specific query) into a single visual component, where the LLM’s role is limited to generating a visualization based on given inputs. In contrast, our setting is more challenging and user-centric: given open-ended questions (e.g., “What is the trend of gold prices?”), the model must decide when and how to incorporate visualizations into a rich-text response. Existing benchmarks fail to evaluate the quality of visually-rich responses generated by LLMs.

Key	Description
page id	A unique identifier for the document within a set of query results or references.
page url	The website or platform where the document was originally published.
page title	The main title of the document.
page synopsis	A brief overview or abstract of the document’s content.
page content	The complete textual content of the document.
last modified time	The date and time when the document was published.

Table 4: **Field descriptions** of the reference document schema.

Deep Research Agents and Data Insight Benchmark. Our work is also related to “deep research” agents Google (2025); Team et al. (2025); OpenAI (2025b) and data insight benchmarks like InsightBench Sahu et al. (2024). While these can also generate visual reports, RAViG-Bench is distinguished by its specific task focus and evaluation methodology. First, the RAViG task is defined for the real-time, realistic, and everyday context of conversational AI, contrasting with the time-intensive, in-depth, and analyst-style nature of typical “deep research” tasks. Second, and more importantly, our evaluation framework provides a novel, holistic assessment not found in existing benchmarks. For instance, InsightBench evaluates the correctness of data insights derived from structured data. In contrast, RAViG-Bench evaluates an LLM’s ability to synthesize information from unstructured web documents and present it in a multi-component visual layout. Our evaluation is also broader, covering not only content quality (including chart-text consistency) but also functionality and visual design quality, aspects crucial for user-facing outputs but not addressed by prior benchmarks. Notably, our framework is versatile and can be used to evaluate the HTML output of any agent, including those designed for deep research.

B DETAILED FIELDS OF RETRIEVED REFERENCE DOCUMENTS

Document Fields are listed in the Tab. 4.

C LICENSES

- Natural Questions is released under the Creative Commons Share-Alike 3.0 license.
- DeepSeek-V3 is licensed under the MIT License.
- Llama-4-Scout-17B-16E-Instruct is licensed under the Llama 4 Community License.
- Llama-3.3-70B-Instruct is licensed under the Llama 3 Community License.
- Mistral-Small-3.1-24B-Instruct-2503 and Qwen3-235B-A22B-Instruct-2507 are licensed under Apache 2.0 License.

D THE USE OF LARGE LANGUAGE MODELS (LLMs)

In the preparation of this paper, we utilized the Google Gemini Pro model (version Gemini-2.5-Pro-0613 accessed via the Google AI Platform) as a writing assistant. The use of the LLM was strictly limited to the following assistive tasks:

Language Enhancement: The model was employed to improve grammar, refine phrasing, and enhance the clarity and readability of the text. All suggestions were critically reviewed by the authors, and only those that accurately reflected our intended meaning were incorporated.

Prompt Optimization: For the prompt-based components of our methodology, we used the LLM to brainstorm and iterate on prompt structures. The final prompts used in our experiments were developed and finalized by the authors.

Crucially, the LLM did not contribute to the core research ideas, experimental design, or the analysis of results. The intellectual contributions, including the formulation of the research problem, the development of the proposed method, and the interpretation of the findings, are entirely the work of

Type	Intention	Typical Visualization
Trend	Changes over time.	Line Chart, Area Chart
Proportion	Part-to-whole relationships.	Pie Chart, Donut Chart
Comparison	Comparing distinct items.	Bar Chart, Radar Chart
Distribution	How data is spread across a range.	Histogram, Box Plot
Ranking	Ordering items by a metric.	Sorted Bar, Leaderboard
Details	Retrieving raw/detailed data.	Card List, Enhanced Table

Table 5: Six query types for visually rich answers generation.

Type	Representative Query Examples
Trend	<ul style="list-style-type: none"> - Yearly average ocean pH levels since 1980 - Number of billionaires worldwide by year since 1990 - Annual number of 3D printers sold globally since 2007 - What is the monthly average temperature in antarctica? - The trend in Netflix subscriber growth each quarter since 2013
Proportion	<ul style="list-style-type: none"> - What is the ratio of gases in air? - Composition of Earth’s crust by major elements - Of the total water on earth what percent is freshwater? - What percent of american adults have a bachelor’s degree? - What percent of the us population is employed in agriculture?
Comparison	<ul style="list-style-type: none"> - Minimum wage vs cost of living by country - Compare the speed of a cheetah versus a gazelle - Who has sold more albums: Taylor Swift or Beyonce? - How many times bigger is vy canis majoris than our sun? - How does the temperature of venus compare to that of earth?
Distribution	<ul style="list-style-type: none"> - Distribution of scores on the 2024 LSAT exam - Distribution of individual income in the United States - What is the distribution of average sleep quality scores? - Distribution of children’s book word counts published in 2023 - Describe the scope of the foodservice industry in relation to its size
Ranking	<ul style="list-style-type: none"> - Top 10 box office movies in the world - 10 most populated capital cities in the world - Three largest cities in the world by population - 10 wealthiest countries in the world by income - What are the eight most common elements in the earth’s crust?
Details	<ul style="list-style-type: none"> - GDP per capita by state in the us - What are the depths of the five great lakes? - Population of the 5 boroughs of new york city - Details of all Nobel Prize winners in Literature since 2010 - List major volcanic eruptions since 2000 with VEI and fatalities

Table 6: Examples of dataset queries

the human authors. All final content, data, and conclusions presented in this paper were personally verified and approved by the authors to ensure scientific accuracy and integrity.

E DETAILED DESCRIPTION OF QUERY CATEGORY AND EXAMPLES

As introduced in Section 3.2, the queries in RAViG-Bench are classified into six distinct categories to cover a breadth of real-world analytical tasks. These categories form the foundation for evaluating a model’s ability to generate appropriate visually-rich responses. Tab.5 summarizes these categories, detailing the core intention for each type and the typical visualization format suited to address it.

1026 Following the definitions above, Tab. 6 showcases representative queries for each category to ground
 1027 them in real-world examples.
 1028

1029 **F EXAMPLES OF VISUALLY RICH ANSWERS GENERATED BY LLMs**
 1030

1031 This section presents some examples of LLM-generated visually-rich responses. It also shows the
 1032 effects of our full webpage screenshots and module screenshots.
 1033

1034 **To Gemini-2.5pro:** Distribution Query

1035 What is the distribution of average time spent commuting by mode of transport?

1038 **The Commuting Equation**
 1039 An Analysis of Travel Time by Transportation Mode

1040

1041 **Core Insight:** While solo driving is the most common commute, active and public transport options
 1042 present a significant time investment. Biking, surprisingly, demands the longest average commute,
 1043 suggesting its use for longer-distance, dedicated journeys.

1044 **Average One-Way Commute Time by Mode**



1045

1046

1047

1048

1049

1050

1051

1052 **Key Commuting Realities**

1053

1054 **The Public Transit Time Tax**

1055 Commutes by bus, subway, and rail consistently rank among the longest. This extended
 1056 duration isn't just travel time but also includes waiting, transfers, and the "last mile"
 1057 journey from stop to destination. Data from the U.S. Census Bureau confirms bus
 1058 commutes average over 46 minutes.

1059 **The Car Conundrum**

1060 While roughly 77% of Americans
 1061 drive to work, it's a middle-of-the-
 1062 pack option for speed.
 1063 Interestingly, carpooling is slightly
 1064 faster than driving alone,
 1065 suggesting benefits from HOV
 1066 lanes or shared driving
 1067 responsibilities.

1068 **The Active Commute**

1069 Biking has the longest average
 1070 duration, implying it's often used
 1071 for longer, more deliberate
 1072 journeys rather than short hops.
 1073 Walking is a moderate-timed
 1074 option, suitable for those living
 1075 closer to their workplace.

1076

1077 **The Hidden Costs of the Commute**

1078 **Environmental**

1079 The transportation
 sector is a major source
 of greenhouse gases,
 with personal vehicles
 contributing significantly
 to smog and emissions.
 Reducing drive-alone
 trips is key to mitigating
 this impact.

1078 **Financial**

1079 Commuting is expensive.
 The costs of gas,
 insurance, vehicle
 maintenance, and transit
 fares add up, creating a
 significant financial
 burden on workers,
 especially those with
 long-distance
 commutes.

1078 **Health**

1079 Long commutes are
 linked to increased
 stress, higher risk of
 obesity and high blood
 pressure, and decreased
 job satisfaction. The
 time spent commuting
 often displaces time for
 exercise and social
 activities.

1076

Full-page

1038 **The Commuting Equation**
 1039 An Analysis of Travel Time by Transportation Mode

1040

1041 **Core Insight:** While solo driving is the most common commute, active and public transport options
 1042 present a significant time investment. Biking, surprisingly, demands the longest average commute,
 1043 suggesting its use for longer-distance, dedicated journeys.

1044 **Average One-Way Commute Time by Mode**



1045

1046

1047

1048

1049

1050

1051

1052 **Key Commuting Realities**

1053

1054 **The Public Transit Time Tax**

1055 Commutes by bus, subway, and rail consistently rank among the longest. This extended
 1056 duration isn't just travel time but also includes waiting, transfers, and the "last mile"
 1057 journey from stop to destination. Data from the U.S. Census Bureau confirms bus
 1058 commutes average over 46 minutes.

1059 **The Car Conundrum**

1060 While roughly 77% of Americans
 1061 drive to work, it's a middle-of-the-
 1062 pack option for speed.
 1063 Interestingly, carpooling is slightly
 1064 faster than driving alone,
 1065 suggesting benefits from HOV
 1066 lanes or shared driving
 1067 responsibilities.

1068 **The Active Commute**

1069 Biking has the longest average
 1070 duration, implying it's often used
 1071 for longer, more deliberate
 1072 journeys rather than short hops.
 1073 Walking is a moderate-timed
 1074 option, suitable for those living
 1075 closer to their workplace.

1076

1077 **The Hidden Costs of the Commute**

1078 **Environmental**

1079 The transportation
 sector is a major source
 of greenhouse gases,
 with personal vehicles
 contributing significantly
 to smog and emissions.
 Reducing drive-alone
 trips is key to mitigating
 this impact.

1078 **Financial**

1079 Commuting is expensive.
 The costs of gas,
 insurance, vehicle
 maintenance, and transit
 fares add up, creating a
 significant financial
 burden on workers,
 especially those with
 long-distance
 commutes.

1078 **Health**

1079 Long commutes are
 linked to increased
 stress, higher risk of
 obesity and high blood
 pressure, and decreased
 job satisfaction. The
 time spent commuting
 often displaces time for
 exercise and social
 activities.

1076

Modules

Figure 6: Visually rich answers generated by Gemini-2.5-pro.

1072
1073
1074
1075
1076
1077
1078
1079

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

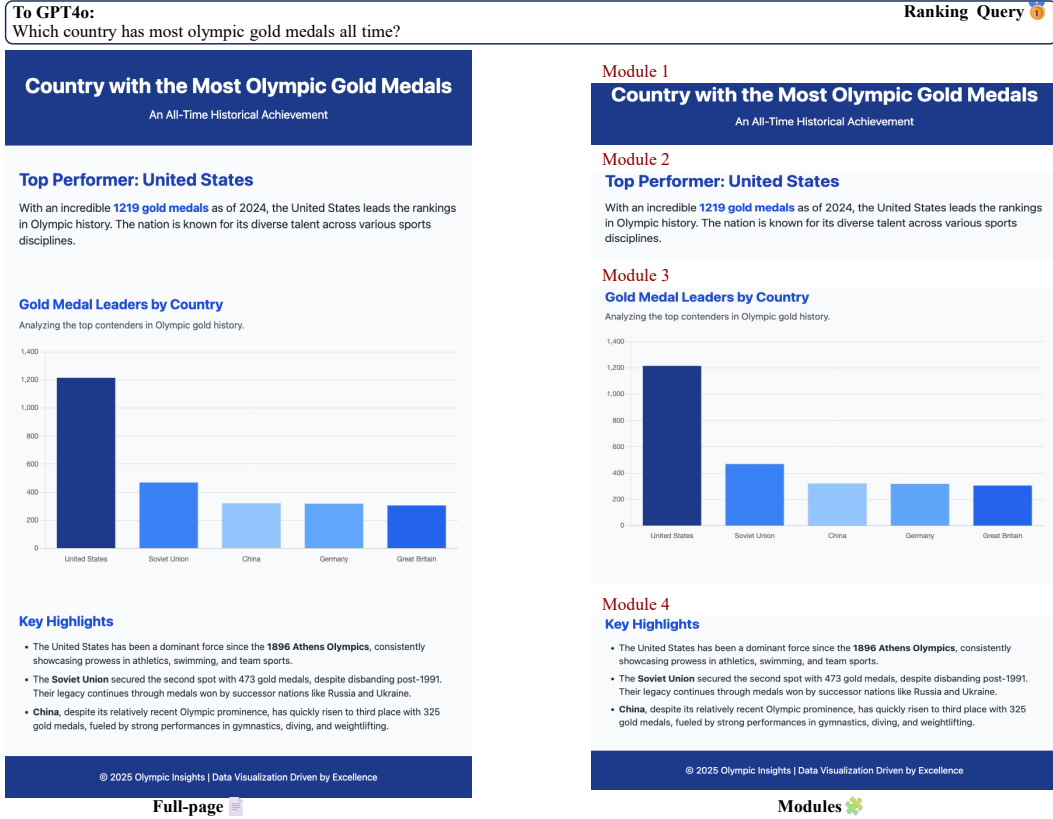


Figure 7: Visually rich answers generated by GPT-4o.

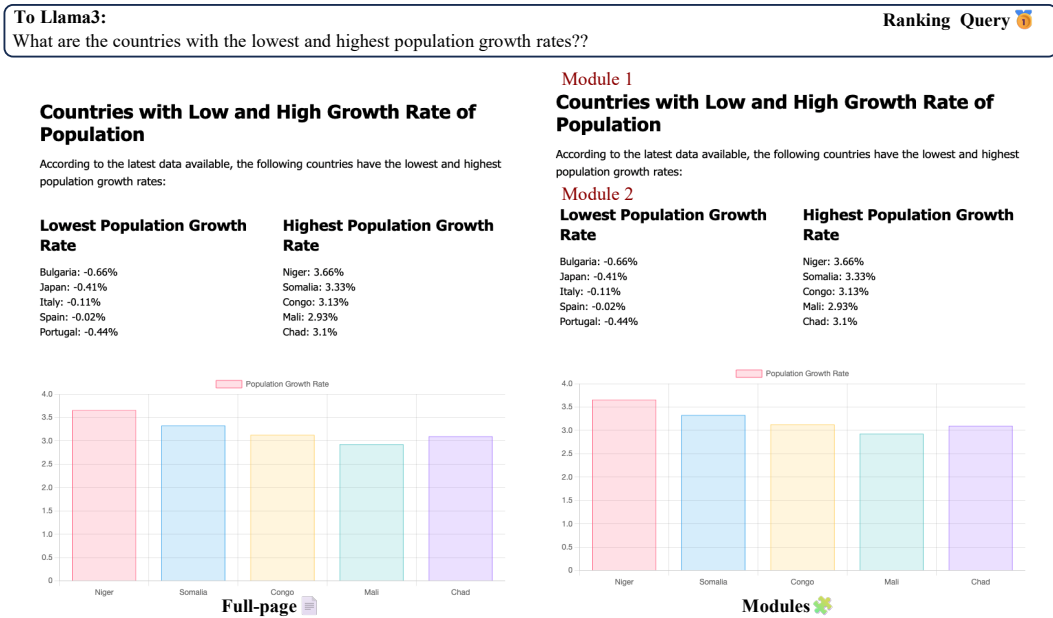


Figure 8: Visually rich answers generated by Llama3.

1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

To Claude4:
Details Query

show typical mortgage interest rates by term for July 2025

● JULY 2025 MORTGAGE RATES

Mortgage Rates by Term

Current mortgage rates across different loan terms to help you make informed decisions

Market Snapshot

As of July 2025, mortgage rates remain elevated but stable. The 30-year fixed rate hovers around 6.72%, while shorter-term loans offer slightly better rates, with 15-year mortgages averaging 5.90%.

30-Year Fixed

6.72% Average Rate

Most popular mortgage term with predictable monthly payments

Monthly Payment* **\$2,596**

15-Year Fixed

5.90% Average Rate

Lower total interest paid with higher monthly payments

Monthly Payment* **\$3,364**

5/6 ARM

7.38% Average Rate

Rate adjusts after 5 years, may fluctuate with market

Initial Payment* **\$2,779**

Rate Trends: 2024-2025

Mortgage rates have stabilized in 2025 after significant volatility in 2024

Market Context

- Federal Reserve held rates steady through five meetings in 2025
- Inflation remains above 2% target, limiting rate cuts
- Analysts expect only 1-2 modest cuts by year-end
- Rates significantly higher than 2.65% low in 2021

Key Takeaways

- Shop multiple lenders - rates vary by 0.5%+ between providers
- Credit score above 740 can secure better rates
- 20% down payment eliminates PMI costs
- Consider points to buy down rate if staying long-term

*Monthly payments calculated on \$400,000 loan amount. Rates shown are averages and may vary based on credit score, down payment, and lender. Data sources: Freddie Mac, Federal Reserve, industry reports.

Module 1

Mortgage Rates by Term

Current mortgage rates across different loan terms to help you make informed decisions

Module 2

Market Snapshot

As of July 2025, mortgage rates remain elevated but stable. The 30-year fixed rate hovers around 6.72%, while shorter-term loans offer slightly better rates, with 15-year mortgages averaging 5.90%.

30-Year Fixed

6.72% Average Rate

Most popular mortgage term with predictable monthly payments

Monthly Payment* **\$2,596**

15-Year Fixed

5.90% Average Rate

Lower total interest paid with higher monthly payments

Monthly Payment* **\$3,364**

5/6 ARM

7.38% Average Rate

Rate adjusts after 5 years, may fluctuate with market

Initial Payment* **\$2,779**

Module 3

Rate Trends: 2024-2025

Mortgage rates have stabilized in 2025 after significant volatility in 2024

Market Context

- Federal Reserve held rates steady through five meetings in 2025
- Inflation remains above 2% target, limiting rate cuts
- Analysts expect only 1-2 modest cuts by year-end
- Rates significantly higher than 2.65% low in 2021

Key Takeaways

- Shop multiple lenders - rates vary by 0.5%+ between providers
- Credit score above 740 can secure better rates
- 20% down payment eliminates PMI costs
- Consider points to buy down rate if staying long-term

*Monthly payments calculated on \$400,000 loan amount. Rates shown are averages and may vary based on credit score, down payment, and lender. Data sources: Freddie Mac, Federal Reserve, industry reports.

Full-page
Modules

Figure 9: Visually rich answers generated by Claude-Ops-4.

22

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

To DeepSeek-v3: Comparison Query

How many events in summer olympics vs winter olympics?

Summer vs Winter Olympics: Event Comparison

Understanding the scale and differences between the world's premier sporting events

Key Statistics

Winter Olympics 109 events <small>Beijing 2022 Winter Olympics</small>	Summer Olympics 339 events <small>Tokyo 2020 Summer Olympics</small>
---	---

Event Comparison

Winter Olympics feature significantly fewer events than Summer Olympics, focusing on snow and ice sports.

Winter Olympics <ul style="list-style-type: none"> ✓ 109 events across 7 sports in Beijing 2022 ✓ 116 events scheduled for Milano Cortina 2026 ✓ Featured sports include Alpine skiing, ice hockey, and figure skating 	Summer Olympics <ul style="list-style-type: none"> ✓ 339 events in Tokyo 2020 (held in 2021) ✓ 35 sports in LA 2028 Summer Olympics ✓ Major sports include athletics, swimming, and gymnastics
--	--

Historical Context

The Summer and Winter Olympic Games are organized by the International Olympic Committee every four years, staggered so there's an Olympic event every two years.

The Summer Olympics is a much bigger event with 206 countries participating in the 2020 Olympics, while the Winter Olympics had 91 countries participating in 2022.

Summer Olympics are held every leap year (2016, 2020, 2024), while Winter Olympics are held two years after the leap year (2018, 2022, 2026).

Module 1

Summer vs Winter Olympics: Event Comparison

Understanding the scale and differences between the world's premier sporting events

Module 2

Key Statistics

Winter Olympics 109 events <small>Beijing 2022 Winter Olympics</small>	Summer Olympics 339 events <small>Tokyo 2020 Summer Olympics</small>
---	---

Module 3

Event Comparison

Winter Olympics feature significantly fewer events than Summer Olympics, focusing on snow and ice sports.

Winter Olympics <ul style="list-style-type: none"> ✓ 109 events across 7 sports in Beijing 2022 ✓ 116 events scheduled for Milano Cortina 2026 ✓ Featured sports include Alpine skiing, ice hockey, and figure skating 	Summer Olympics <ul style="list-style-type: none"> ✓ 339 events in Tokyo 2020 (held in 2021) ✓ 35 sports in LA 2028 Summer Olympics ✓ Major sports include athletics, swimming, and gymnastics
--	--

Module 4

Historical Context

The Summer and Winter Olympic Games are organized by the International Olympic Committee every four years, staggered so there's an Olympic event every two years.

The Summer Olympics is a much bigger event with 206 countries participating in the 2020 Olympics, while the Winter Olympics had 91 countries participating in 2022.

Summer Olympics are held every leap year (2016, 2020, 2024), while Winter Olympics are held two years after the leap year (2018, 2022, 2026).

Full-page
Modules

Figure 10: Visually rich answers generated by DeepSeek-v3.

To Llama4: Trend Query

Show Los Angeles average PM2.5 levels by year since 1999

Los Angeles Average PM2.5 Levels by Year Since 1999

Overview

The following chart shows the average PM2.5 levels in Los Angeles by year since 1999.

Data Source

Data from [Los Angeles Almanac](#) and [NOAA](#).

Module 1

Los Angeles Average PM2.5 Levels by Year Since 1999

Module 2

Overview

The following chart shows the average PM2.5 levels in Los Angeles by year since 1999.

Module 3

Data Source

Data from [Los Angeles Almanac](#) and [NOAA](#).

Full-page
Modules

Figure 11: Visually rich answers generated by Llama4.

G EXAMPLES OF VISUALLY RICH ANSWERS FROM DIFFERENT LLMs FOR THE SAME QUERY

This section presents responses from different LLMs to the same query “How many jimmy johns are there in the us” (for lengthy responses, the output is split into two parts for display from top to bottom). For such open-ended task, traditional evaluation based on a ground truth is not applicable. Consequently, we constructed a GT-free framework to assess the quality of the generated outputs.

Query: How many jimmy johns are there in the us

Model: Claude

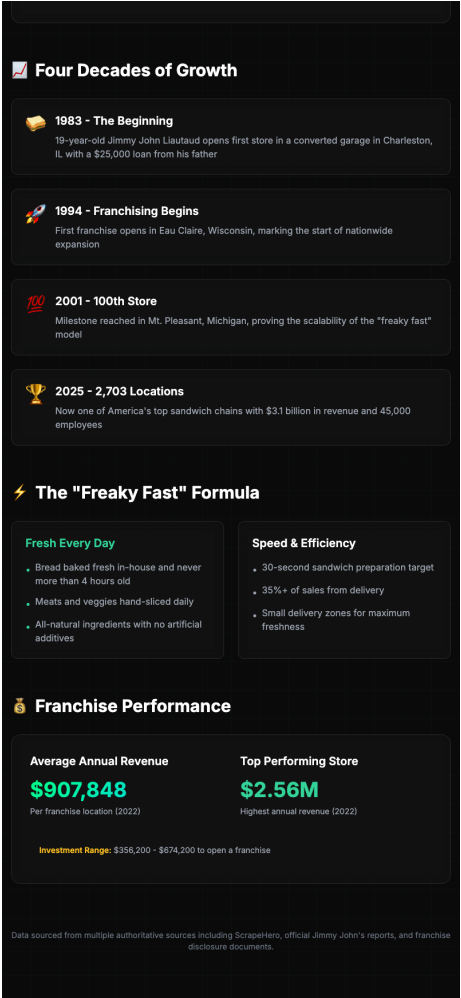
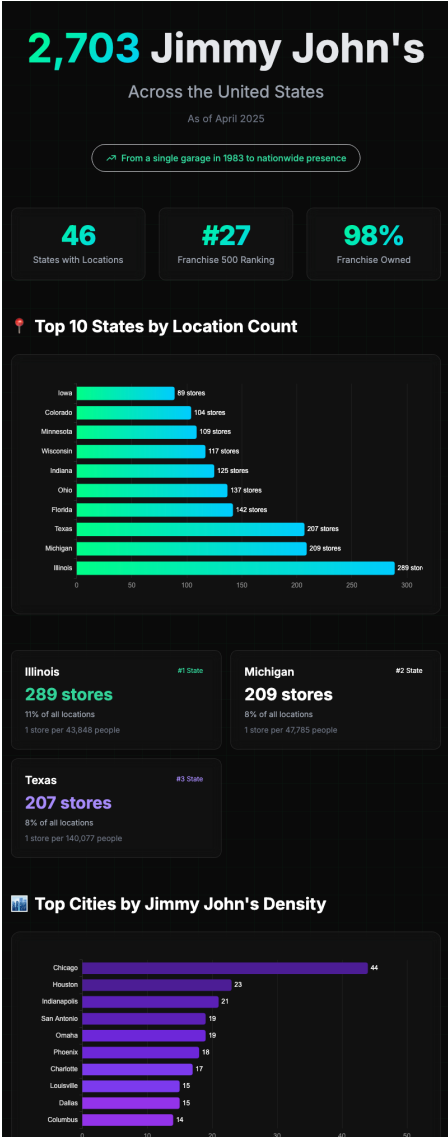


Figure 12: Query and visually rich answers generated by Claude-Opus-4.

1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349

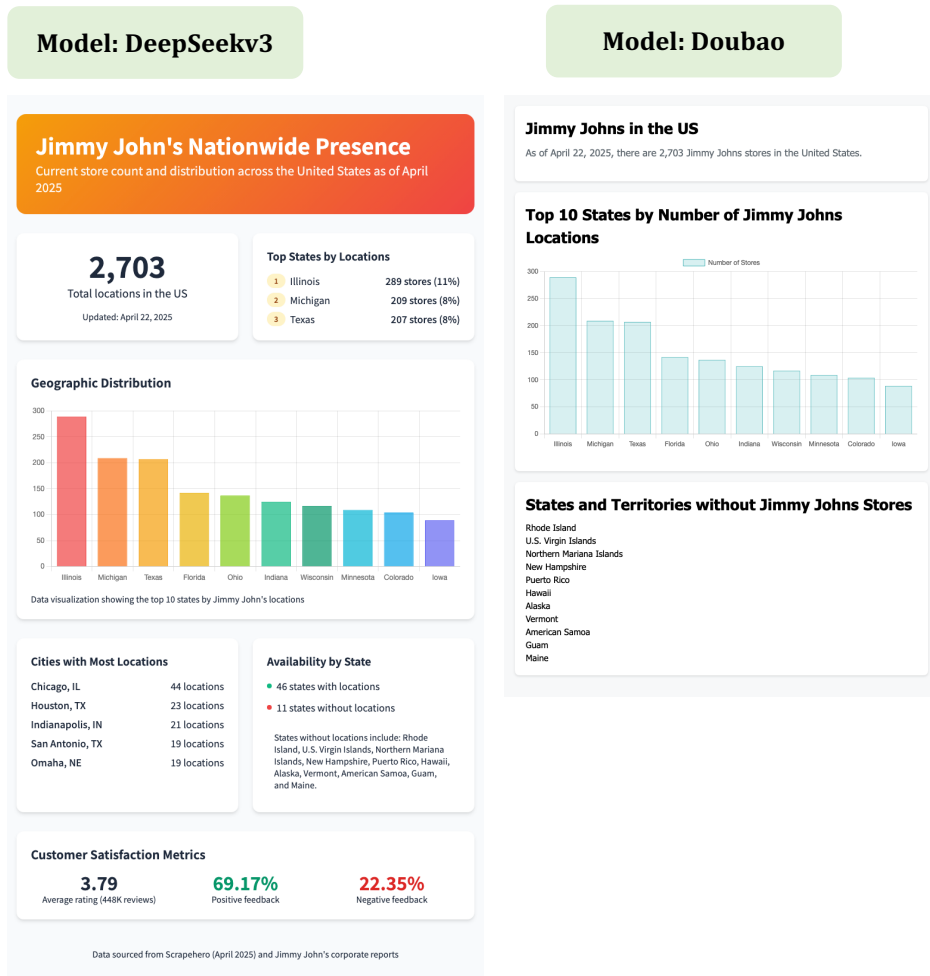


Figure 13: Query and visually rich answers generated by DeepSeek-v3 and Doubao-1.5-V-pro.

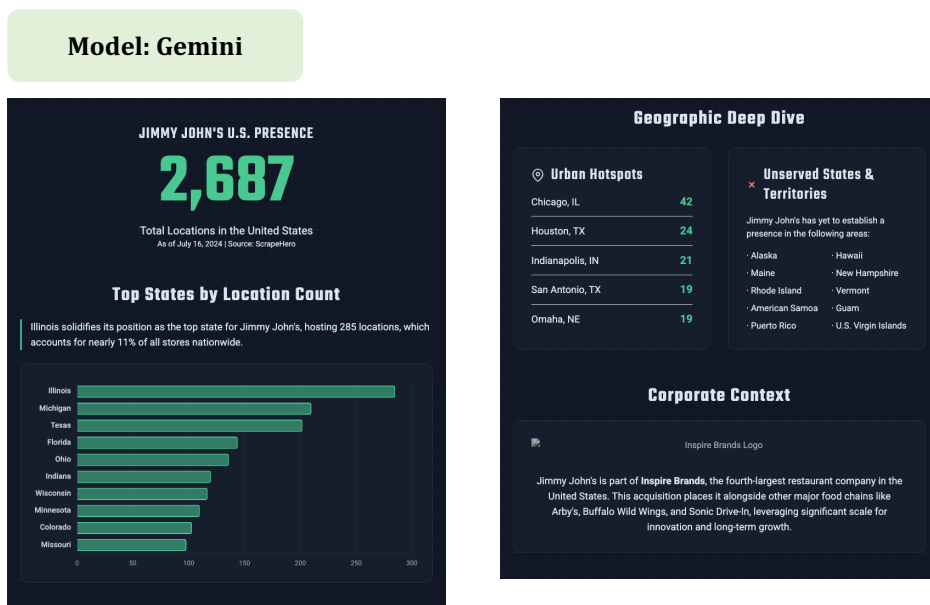


Figure 14: Query and visually rich answers generated by Gemini-2.5-pro.

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403



Figure 15: Query and visually rich answers generated by GPT-4o and GPT-5.

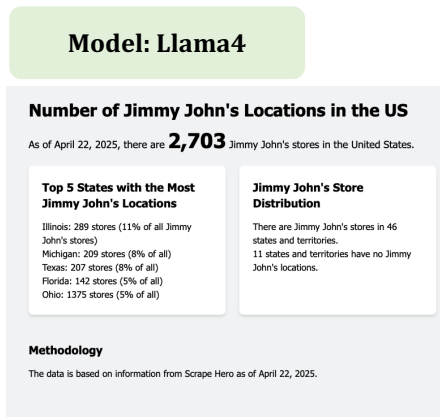
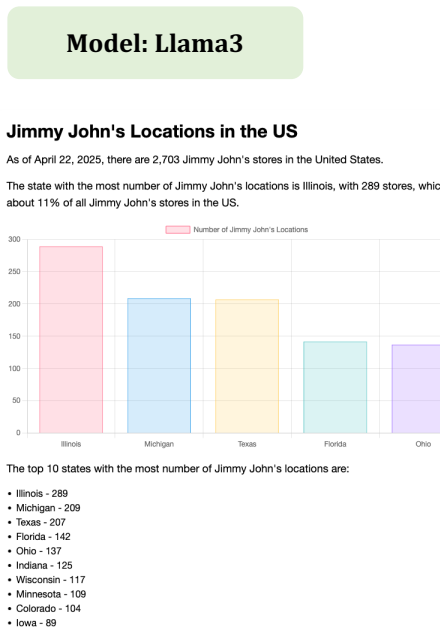


Figure 16: Query and visually rich answers generated by Llama3 and Llama4.

1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

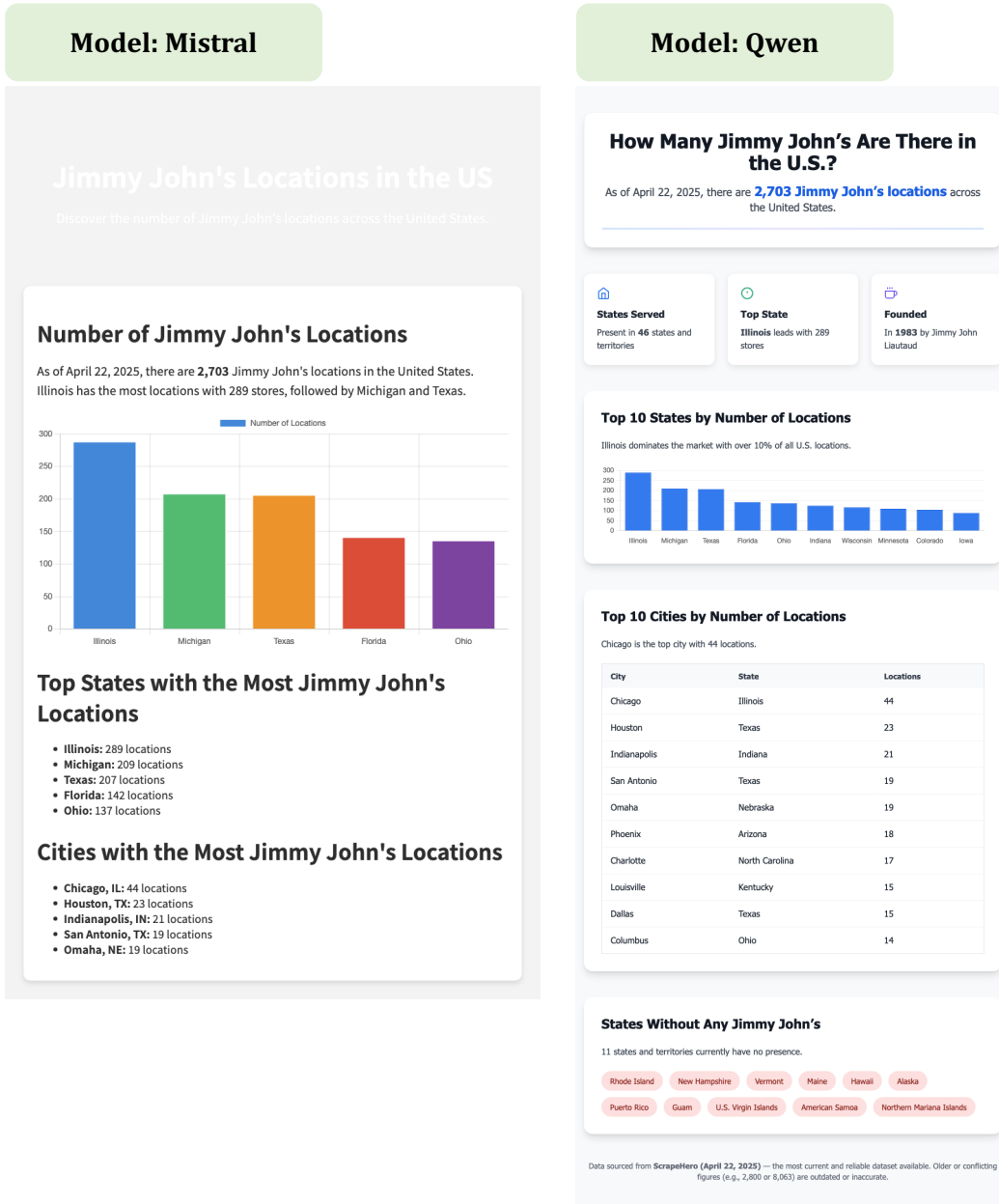


Figure 17: Query and visually rich answers generated by Mistral and Qwen3.

H EXAMPLES OF HUMAN-MACHINE INCONSISTENT CASES IN DESIGN AND CONTENT DIMENSIONS

In this section, we present an analysis of the inconsistencies between our automated judges (Gemini 2.5-Pro, GPT-4o, and GPT-5) and the majority vote of three human experts. During this analysis, we made a surprising discovery: a subset of the disagreements stemmed not from model error, but from subtle defects that human annotators had initially overlooked. These cases, often requiring meticulous cross-modal inspection (e.g., comparing fine-grained details between rendered output and source code), highlight a limitation of manual evaluation and underscore the advantage of using LLMs-as-Judges. In the following, we show several examples of human-machine inconsistencies, presenting the evaluation results from different models alongside the original human judgments.

H.1 DESIGN DIMENSIONS

Fig. 18 and 19 show representative examples of disagreement for the Missing Element dimension. Correct assessments are marked in green, and the rationale from models that made a correct judgment is included below each example. As these examples demonstrate, human evaluators can also make mistakes, particularly when identifying subtle issues such as the omission of fine-grained details or defects that require meticulous comparison with the source code.

Figures 20 and 21 present representative examples of disagreement on the Occlusion dimension. Similar to the previous examples, correct judgments are highlighted in green. For cases where a model experienced a hallucination (i.e., incorrectly identified an occlusion), we provide the model’s flawed reasoning to illustrate the nature of the error. The occlusions missed by LLMs (false negatives) are typically very subtle. And their hallucinations (false positives) are often triggered by low color contrast or small font sizes.

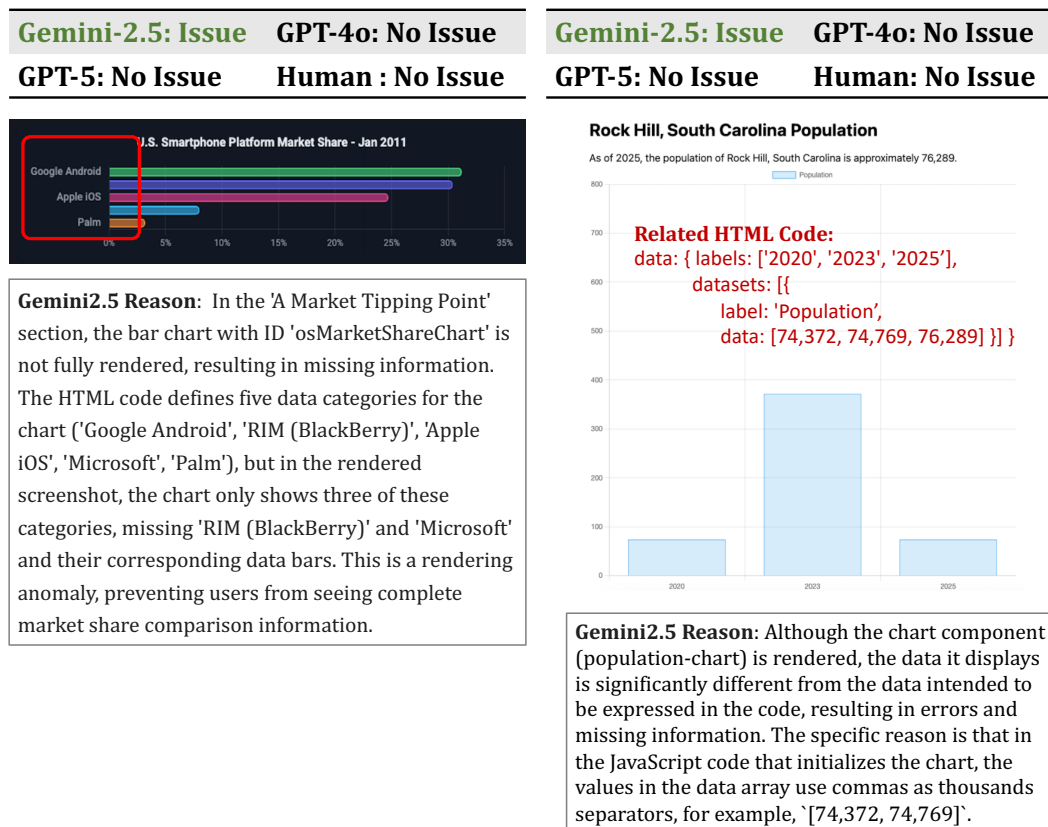


Figure 18: Examples of inconsistencies on Missing Element dimension (1), correct assessments are marked in green.

1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565

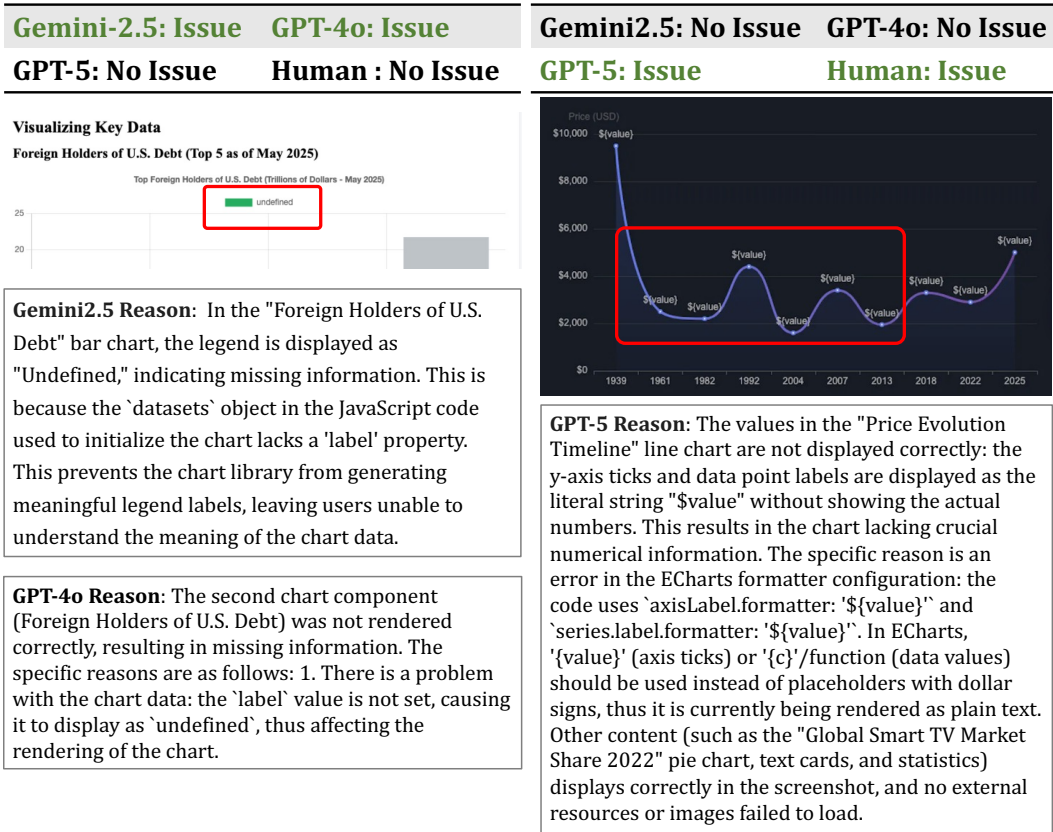


Figure 19: Examples of inconsistencies on Missing Element dimension (2), correct assessments are marked in green.



Figure 20: Examples of inconsistencies on Occlusion dimension (1), correct assessments are marked in green.

1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619

Gemini2.5: N-Issue	GPT-4o: N-Issue	Gemini2.5: N-Issue	GPT-4o: Issue								
GPT-5: Issue	Human : N-Issue	GPT-5: N-Issue	Human: N-Issue								
<p>Visualizing the Scale</p>		<p>The Dragon's Legacy</p> <p>Yusef has established itself as a top National governing authority in the Middle East since 2010, leading to a record-breaking performance in the region's most successful nation in the tournament. The year 2010 was a record for the team.</p> <table border="1"> <tr> <td>6</td> <td>4</td> <td>28</td> <td>39</td> </tr> <tr> <td>Modern Titles</td> <td>Grand Slams</td> <td>Outright Wins</td> <td>Total Titles</td> </tr> </table>		6	4	28	39	Modern Titles	Grand Slams	Outright Wins	Total Titles
6	4	28	39								
Modern Titles	Grand Slams	Outright Wins	Total Titles								
<p>GPT-5 Hallucination Output: In the image, the large pie chart overflows its container boundary, and the bottom pie chart area covers and overlaps with the area containing the caption "Employment data compared globally." This overlap and conflict between the description and the chart, interfering with readability. This is a problem of element overlap and occlusion caused by chart overflow, affecting the user's retrieval of chart information.</p>		<p>GPT-4o Hallucination Output: In the image, the statistics at the bottom (such as '6 Modern Titles', '4 Grand Slams', '28 Outright Wins', '39 Total Titles') are severely affected by the background transparency and color design, making the text almost illegible. <u>This occlusion problem is caused by insufficient contrast between the background and text colors</u> and improper transparency settings, affecting the user's ability to read and understand key information.</p>									

Figure 21: **Examples of inconsistencies on Occlusion dimension (2)**, correct assessments are marked in green.

H.2 CONTENT DIMENSIONS

Fig. 22 and 23 show representative examples of disagreement for the reasonableness dimension. Correct assessments are marked in green, and the rationale from models that made a correct judgment is included below each example. Assessing reasonableness requires LLMs to possess strong reasoning capabilities as well as a solid understanding of HTML code. Consequently, non-reasoning LLMs typically exhibit poor consistency on this dimension. Even high-performing models such as Gemini-2.5-Pro and GPT-5 can make erroneous judgments due to excessive reasoning (Fig. 23) and knowledge hallucination (Fig. 22). Meanwhile, humans also have gaps in certain areas of common sense, often struggling to identify relatively obscure factual errors or being misled by similar entities.

Fig. 24 show representative examples of disagreement for the faithfulness dimension. Similar to the previous examples, correct judgments are highlighted in green. In terms of faithfulness evaluation, LLMs demonstrate high agreement with human; Gemini-2.5-pro and GPT-5 achieve consistency rates of 94% and 95%, respectively. However, in scenarios involving long texts, noisy inputs, and required reasoning, both human experts and LLMs still exhibit some degree of hallucination.



Figure 22: Examples of inconsistencies on Reasonableness dimension (1), correct assessments are marked in green.


1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727

Gemini-2.5: Issue	GPT-4o: No Issue	Gemini-2.5: No Issue	GPT-4o: No Issue
GPT-5: No Issue	Human : No Issue	GPT-5: Issue	Human: No Issue

Key Takeaways

- ✓ Toyota continues its decade-long market leadership with a 12.5% global share in 2025
- ✓ Volkswagen ranks second with a 9.7% share, showing particular strength in European markets
- ✓ Chinese automakers like Geely and BYD are showing the fastest growth rates in 2025
- ✓ Electric vehicles continue to gain market share, with Tesla and BYD leading the segment

Arizona vs Nevada Water Usage



Overview
Arizona and Nevada have different water usage patterns. Understanding these patterns can help in managing water resources effectively.

Arizona Water Usage
On average, each Arizona resident uses about 146 gallons per day.

Nevada Water Usage
In 2023, preliminary numbers show the valley's water use at 89 gallons per person a day, compared with 104 in 2022 and 235 in 1990.

Gemini-2.5 Hallucination Output: The evaluation date is August 22, 2025. The response repeatedly presents data for the full year “2025” as a completed fact . **Since the year 2025 has not yet concluded as of the reference date, these full-year results are future events.** They should be framed as projections or estimates. Stating them as established facts in the present or past tense is a violation of temporal plausibility.

GPT-5 Hallucination Output: I evaluated internal consistency: the Nevada section explicitly states the 89 figure is for **“the valley” (i.e., a sub-state region)**, while the chart labels the category as “Nevada” and the page frames the comparison as Arizona vs Nevada. This mixes geographic scopes (statewide Arizona vs a valley in Nevada) and mislabels the chart category, creating a misleading and internally inconsistent comparison.

Figure 23: **Examples of inconsistencies on Reasonableness dimension (2), correct assessments are marked in green.**

1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781

Gemini-2.5: No Issue	GPT-4o: Issue	Gemini-2.5: Issue	GPT-4o: No Issue
GPT-5: No Issue	Human : No Issue	GPT-5: No Issue	Human: No Issue

Claim: 5% of the U.S. immigrant population is from Africa.

Docs: ...- Sub-Saharan Africa 5 - Middle East and North Africa 4...

Label: Neutral

Gemini-2.5 Reason: Doc 1 provides a 2022 regional breakdown including `Sub-Saharan Africa 5%` and `Middle East and North Africa 4%`. **We cannot sum `Sub-Saharan Africa (5%)` with `Middle East and North Africa (4%)` to get an Africa total because the 4% includes non-African Middle Eastern origins.** Therefore, the references neither confirm nor refute that 5% of the U.S. immigrant population is from Africa. [Final Conclusion] The references lack the precise, matching metric needed to verify the claim.

GPT-4o Hallucination Output: The references explicitly state that immigrants from sub-Saharan Africa make up 5% of the U.S. immigrant population (Document 1) and that 5% of new arrivals between 2021 and 2023 were from sub-Saharan Africa (Document 2). This directly supports the claim that 5% of the U.S. immigrant population is from Africa.

Strategic Insights for 2025

- Mobile-First is Non-Negotiable**
Mobile accounts for 65.3% of digital spend, with U.S. mobile ad spending reaching \$228.11 billion
- Precision Targeting Dominates**
78.1% of ad spend will be algorithm-driven by 2027, with programmatic capturing 82.4% of digital investments
- CTV Disrupts Traditional TV**
Connected TV ad spend grows 10.9% while traditional TV declines 1.8%
- Retail Media Networks Surge**
Retail media spending grows 13.9%, projected to reach \$163 billion globally

Claim: Connected TV (CTV) ad spend is growing at 10.9%, while traditional TV ad spend is declining by 1.8%.

Docs: ...Connected television ad spend is now forecast to grow by 10.9 whilst total television spend including broadcast television is expected to decline by 1.8...

Label: Contradiction

Gemini-2.5 Reason: "Total" spend is the sum of its parts. Since one part (Connected TV) is growing significantly at 10.9%, **the other part (traditional TV) must be declining by a rate greater than 1.8% for the overall total to decline by 1.8%.** Therefore, the claim incorrectly applies the decline figure for the *total* market to just the *traditional* TV market segment.

GPT-5 Hallucination Output: Doc 1 provides specific ratios (CTV +10.9%, total TV -1.8%), but lacks clear evidence supporting the -1.8% figure for traditional TV; therefore, this reference is insufficient to adequately substantiate its claim.

Figure 24: **Examples of inconsistencies on Faithfulness dimension, correct assessments are marked in green.**

I EXAMPLES OF CASES WITH CONTENT ISSUES

This section shows prevalent content quality deficiencies in leading LLMs, illustrated through specific instances of contextual inconsistency, irrelevant response and factual hallucination. Fig. 25, 26, 27, and 28 illustrate issues with Reasonableness, Fig. 29 presents failures in Comprehensiveness, and Fig. 30 provides examples of Faithfulness errors. The problematic areas and their corresponding reasons are annotated in the figure. In each example, the erroneous area is highlighted, and an annotation explains the nature of the error.



Figure 25: Examples of Reasonableness failures manifesting as chart-text inconsistencies. The problematic areas along with the rationale are annotated.

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

To Claude4 :
what percentage of the earth's surface is water

71% of Earth's Surface is Water

Our planet is a watery world, earning it the nickname "The Blue Planet"

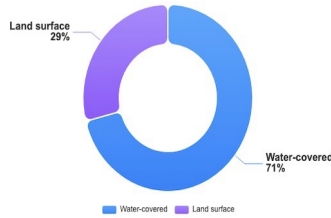
Direct Answer

According to authoritative sources including the USGS and NOAA, approximately **71%** of Earth's surface is covered by water. This includes all oceans, seas, lakes, rivers, and other surface water bodies.

Earth's Water Distribution

Surface Coverage	Water Types
Water-covered surface: 71%	Oceans (salt water): 96.5%
Land surface: 29%	Freshwater: 3.5%

Earth's water coverage visualized as a pie chart showing the dramatic dominance of water over land



Key Facts About Earth's Water

- The total volume of water on Earth is approximately **1.386 billion cubic kilometers** (333 million cubic miles)
- Of all Earth's water, **97.5% is saltwater** and only 2.5% is freshwater
- Most freshwater is locked in ice caps and glaciers, with only **0.3% available as liquid surface water**
- Despite covering 71% of the surface, oceans account for only **0.02% of Earth's total mass**

Why Earth is Called "The Blue Planet"

When viewed from space, Earth appears predominantly blue due to the vast oceans covering most of its surface. Water reflects blue light wavelengths, giving our planet its distinctive azure appearance. This extensive water coverage is unique among the rocky planets in our solar system, making Earth a true oasis in space.

Content Issue: Reasonableness
Issues Type: Internal Logical Consistency

To Deepseek-V3:
list of indian states by population growth rate

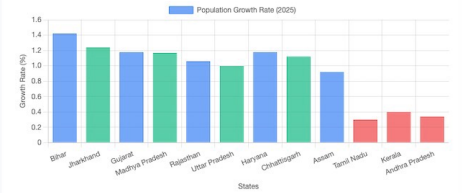
Indian States Population Growth Analysis (2025)

Data-driven insights into demographic trends across India's states and union territories

Latest Data: 2025 Projections | Source: Census India & Statistics

Fastest Growing States		Slowest Growing States	
Bihar	1.42%	Tamil Nadu	0.30%
Jharkhand	1.24%	Kerala	0.40%
Gujarat	1.18%	Andhra Pradesh	0.34%

State Population Growth Rate (2025 Projections)



Top 10 States by Population Growth Rate

RANK	STATE	POPULATION (2025)	GROWTH RATE
1	Bihar	131.0 million	1.42%
2	Jharkhand	40.6 million	1.24%
3	Gujarat	73.5 million	1.18%
4	Madhya Pradesh	89.0 million	1.17%
5	Rajasthan	83.1 million	1.06%
6	Uttar Pradesh	241.3 million	1.00%
7	Uttarakhand	11.9 million	1.00%
8	Chhattisgarh	31.0 million	1.12%
9	Haryana	31.1 million	1.18%
10	Assam	36.5 million	0.92%

Key Demographic Insights

- Fastest Growth:** Bihar leads with 1.42% growth, followed by Jharkhand (1.24%) and Gujarat (1.18%). Union territory Dadra & Nagar Haveli and Daman & Diu have the highest growth at 6.63%.
- Slowest Growth:** Southern states show slower growth: Tamil Nadu (0.30%), Kerala (0.40%), and Andhra Pradesh (0.34%). Nagaland recorded negative growth (-0.6%) in recent data.
- Population Distribution:** Five states (UP, Bihar, Maharashtra, West Bengal, and Madhya Pradesh) account for almost half of India's total population (47.90%).

Content Issue: Reasonableness
Issues Type: Internal Logical Consistency

Figure 26: Examples of Reasonableness failures manifesting as internal logical inconsistencies. The contradictory parts of the text are highlighted with red boxes.

1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943

To Qwen3-235b-a22b-instruct-2507 :
what is the population of fayetteville north carolina

Fayetteville, North Carolina

Population: 210,135 (2025 Estimate)

210k
Current Population

Population Overview

Fayetteville, North Carolina has an estimated population of **210,135** in 2025 according to the latest projections. The city is currently growing at an annual rate of 0.09%, with a 0.46% increase since the previous year.

Key Insight: The population of Fayetteville has steadily increased from 208,501 in the 2020 census to over 210,000 in 2025, reflecting consistent but modest growth.

As the 6th-most populous city in North Carolina, Fayetteville serves as the county seat of Cumberland County and is home to Fort Bragg, a major U.S. Army installation that significantly influences the city's demographics and economy.

Population Growth Trend (2020-2025)

Population has shown steady growth from 208,501 in 2020 to 210,135 projected for 2025

Key Facts

- **2025 Population:** 210,135
- **2020 Census:** 208,501
- **Annual Growth:** 0.09%
- **Metropolitan Area:** 392,336 (2023)
- **County Seat:** Cumberland County
- **Major Feature:** Home to Fort Bragg

Demographics

Based on the 2020 census, Fayetteville's population was 41.8% Black or African American and 34.5% White, with 12.6% identifying as Hispanic or Latino.

Sources

Primary Source: World Population Review (2025 estimate)
Supporting Data: Wikipedia (2020 census), Data USA (2023 data)
Last updated: August 2025

Content Issue: Reasonableness
Issues Type: Chart- Numerical Plausibility
Reason: The sum of percentages in a pie chart exceeds 100%

To Gemini-2.5-pro:
which country is experiencing the most rapid rate of population aging in the world

The New Face of Aging

Identifying the Nations on the Brink of an Unprecedented Demographic Shift

Which country is aging the fastest?

While Japan is currently the world's oldest country, authoritative data indicates that **China is experiencing one of the most rapid aging rates globally**, projected to outpace even Japan's demographic shift in the coming decades.

Source: Analysis of data from WHO and a 2019 PubMed comparative study

🕒 The Current Oldest: Japan

Japan holds the title for the highest proportion of elderly citizens, a result of decades of high life expectancy and low birth rates. Over a third of its population is already over 65.

• **28.2%** of the population is 65 or older.

- Aging process was prolonged over a longer period.
- Represents the current peak of global aging.

🚀 The Fastest Aging: China & South Korea

Nations like China and South Korea are aging at an unprecedented speed. China's demographic transition is occurring in a compressed timeframe, creating unique and urgent challenges.

- Projected to reach **28%** of population over 60 by 2040.
- South Korea is expected to have the highest old-to-young ratio among developed nations.
- Represents the future velocity of global aging.

📊 The Pace of Change: China vs. Japan (2019-2044)

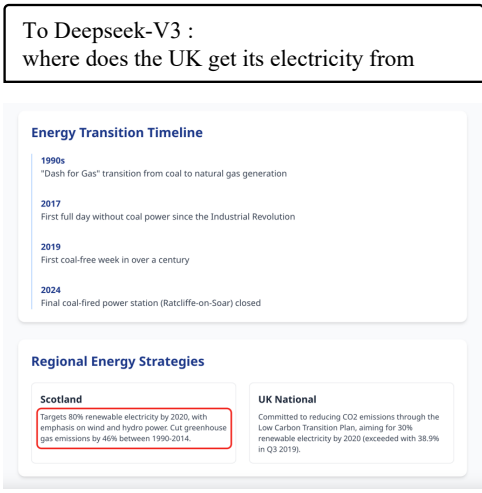
A comparative study reveals China's aging indicators are projected to increase more sharply than Japan's over a 25-year period, highlighting a more compressed and intense demographic shift.

Metric	China (2019-2044)	Japan (2019-2044)
Aging Rate Increase	~13%	~8%
Elderly Dependency Ratio Increase	~24%	~22%
Oldest-Old Coefficient Increase	~8%	~8%

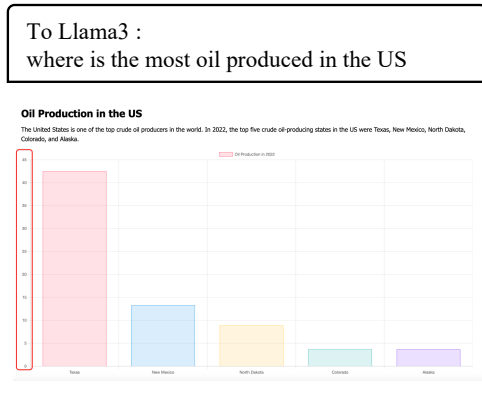
Content Issue: Reasonableness
Issues Type: Internal Logical Consistency
Reason: The response contradicts itself by stating that Japan's elderly population is both "over a third" (which is >33.3%) and specifically "28.2%"

Figure 27: Examples of Reasonableness failures manifesting as numerical errors and internal logical inconsistencies. The problematic areas along with the rationale are annotated.

1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

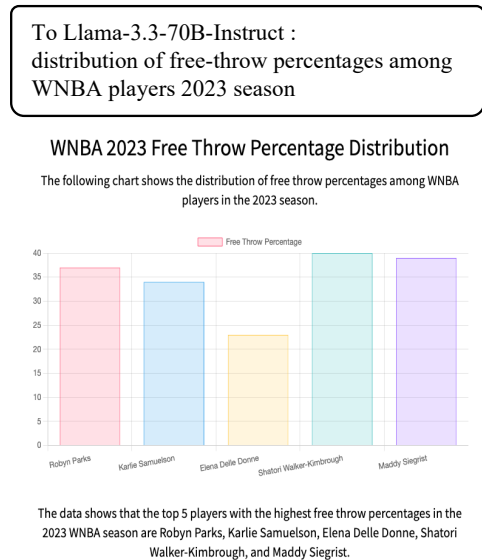


Reasonableness:
Issues Type: Tense Confusion
Reason: Using the present tense "Targets" for a goal with a deadline in the past (2020) is temporally illogical. It should be phrased in the past tense (e.g., "targeted" or "had a target of")



Reasonableness:
Issues Type: Missing Chart Elements
Reason: The y-axis is missing its unit label.

Figure 28: Examples of Reasonableness failures manifesting as tense confusion and missing chart elements. The problematic areas along with the rationale are annotated.



Comprehensiveness:
Reason: In The response fundamentally misunderstands the user's request for a "distribution" and provides wildly inaccurate data. It is misleading and unhelpful.



Comprehensiveness:
Reason: The user asked for details of all F1 races for the 2025 season, but the response only lists 3 out of the 24 scheduled races.

Figure 29: Examples of cases with Comprehensiveness Issues.

1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051

To GPT-4o :
canada's top 5 trading partners imports and exports

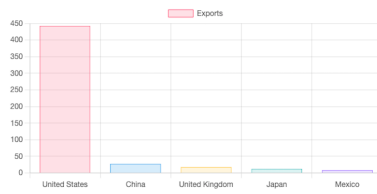
Canada's Top 5 Trading Partners

Overview

Canada's top 5 trading partners are the United States, China, United Kingdom, Japan, and Mexico. These countries account for the majority of Canada's exports and imports.

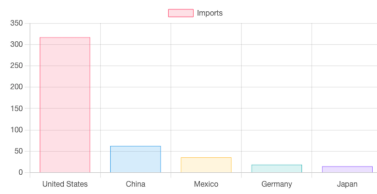
Exports

Canada's top exports include mineral fuels, vehicles, machinery, gems, and precious metals. The United States is Canada's largest export partner, accounting for 76.88% of total exports.



Imports

Canada's top imports include vehicles, machinery, electrical equipment, and plastics. The United States is also Canada's largest import partner, accounting for 49% of total imports.



Trade Balance

Canada had a trade surplus of \$6.6 billion in 2021, with exports totaling \$636 billion and imports totaling \$631 billion.

Content Issue: Faithfulness

Claim: In 2024, Canada's imports from Japan were valued at \$13.14 billion, representing 2.30% of total imports.

Label: Contradiction

Docs: <page id="3" ... 2022 Canada top 5 Export and Import partners ... **Japan 138222.32** </page>

<page id="9" ... Canada Imports By Country Value Year ... **Japan 15.54B 2024** </page>

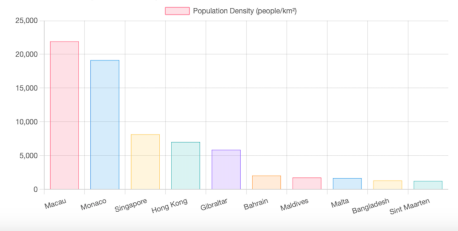
To Llama-4-Scout-17B-16E-Instruct:
list of densely populated countries in the world

List of Densely Populated Countries

The following is a list of countries and territories by population density, based on the number of people per square kilometer.

Rank	Country	Population Density (people/km ²)	Population	Area (km ²)
1	Macau	21,946	683,972	32.9
2	Monaco	19,171	39,241	2.02
3	Singapore	8,177	5,690,000	720.5
4	Hong Kong	7,044	7,519,000	1,106
5	Gibraltar	5,901	33,701	6.8
6	Bahrain	2,093	1,463,330	778
7	Maldives	1,766	436,000	300
8	Malta	1,704	475,701	315
9	Bangladesh	1,350	163,046,161	147,570
10	Sint Maarten	1,293	41,220	34

Visualization of Population Density



Content Issue: Faithfulness

Claim: Ranked 1st in population density, Macau has 21,946 people per square kilometer, a total population of 683,972.

Label: Contradiction

Docs: <page id="4" ... > ...

<synopsis>Highest population density by country 2024 **Monaco led the ranking for countries with the highest population density in 2024**, with nearly 26,000 ...</synopsis> ... </page>

Figure 30: Examples of Faithfulness issues where the generated answers contain contradicts information from the reference document.

J EXAMPLES OF RESPONSES FROM DEEP RESEARCH AGENTS AND RAVIG

In this section, we compare the responses generated by Gemini Deep Research Google (2025), Tongyi Deep Research Team et al. (2025), and our Gemini-2.5-Pro Comanici et al. (2025) with RAViG for the same query. As shown in Fig. 31 and 33, for everyday conversational queries, our end-to-end retrieval-augmented visually-rich generation (RAViG) is comparable to leading Deep Research products in both content quality and page design, while its response time ($\approx 30s$) is 10-20x faster than Deep Research ($\approx 10m$). On the other hand, the extremely long responses produced by Tongyi Deep Research (Fig. 32 and 34) are advantageous for in-depth analytical tasks, but may be less suitable for casual, chat-style queries. This contrast highlights the different application scenarios and positioning of these two types of systems.

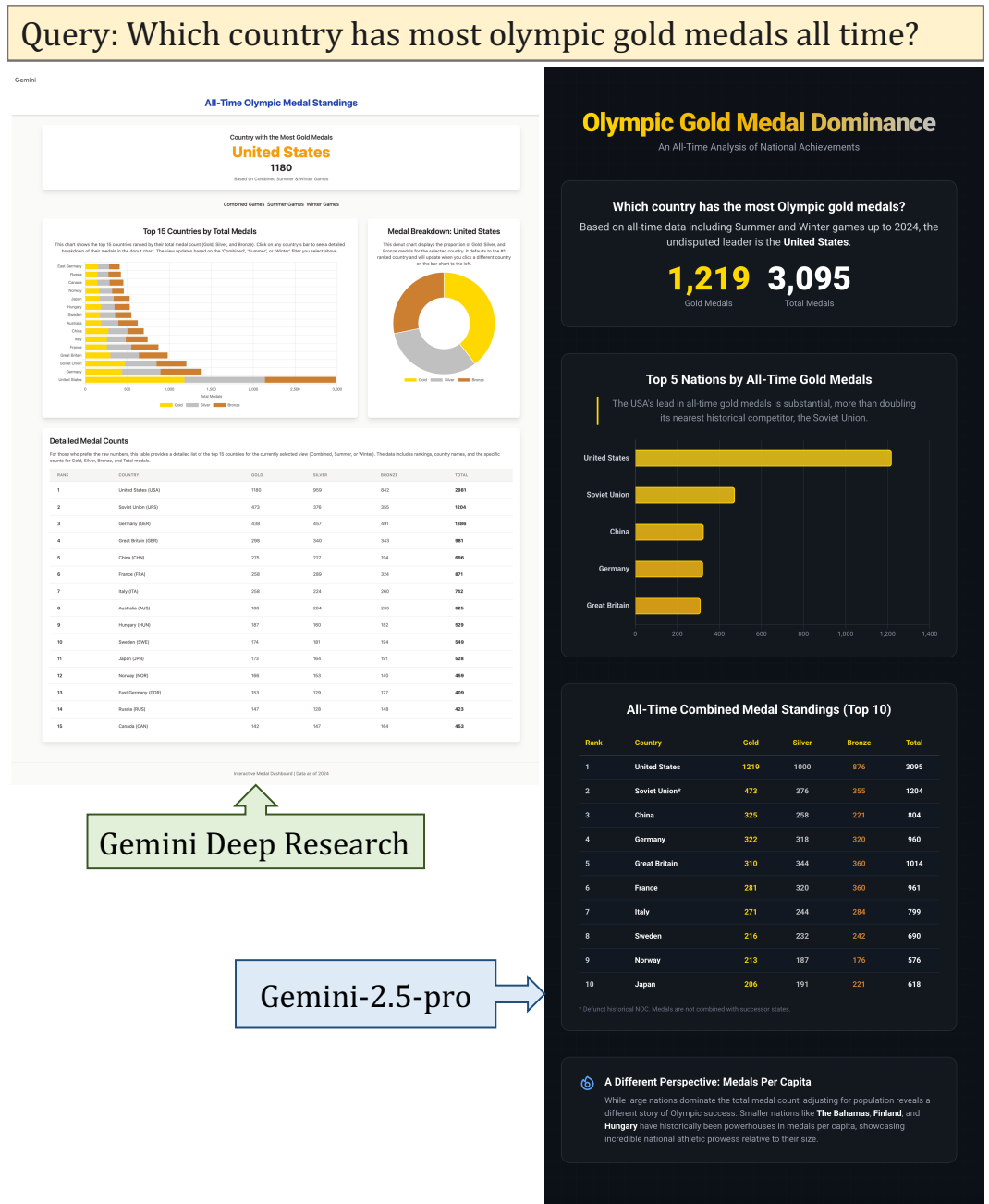
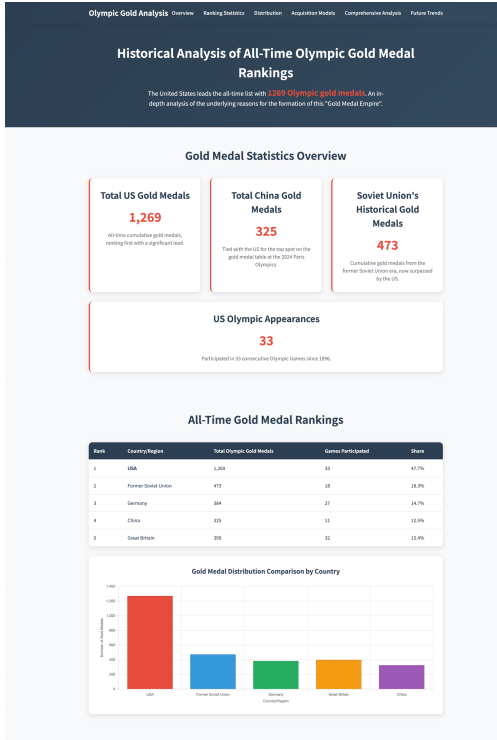


Figure 31: Responses from Gemini Deep Research and Gemini-RAViG (query 1).

2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159

Tongyi Deep Research (Qwen3 Max)

First half page



Second half page



Figure 32: Responses from Tongyi Deep Research (query 1).

2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213

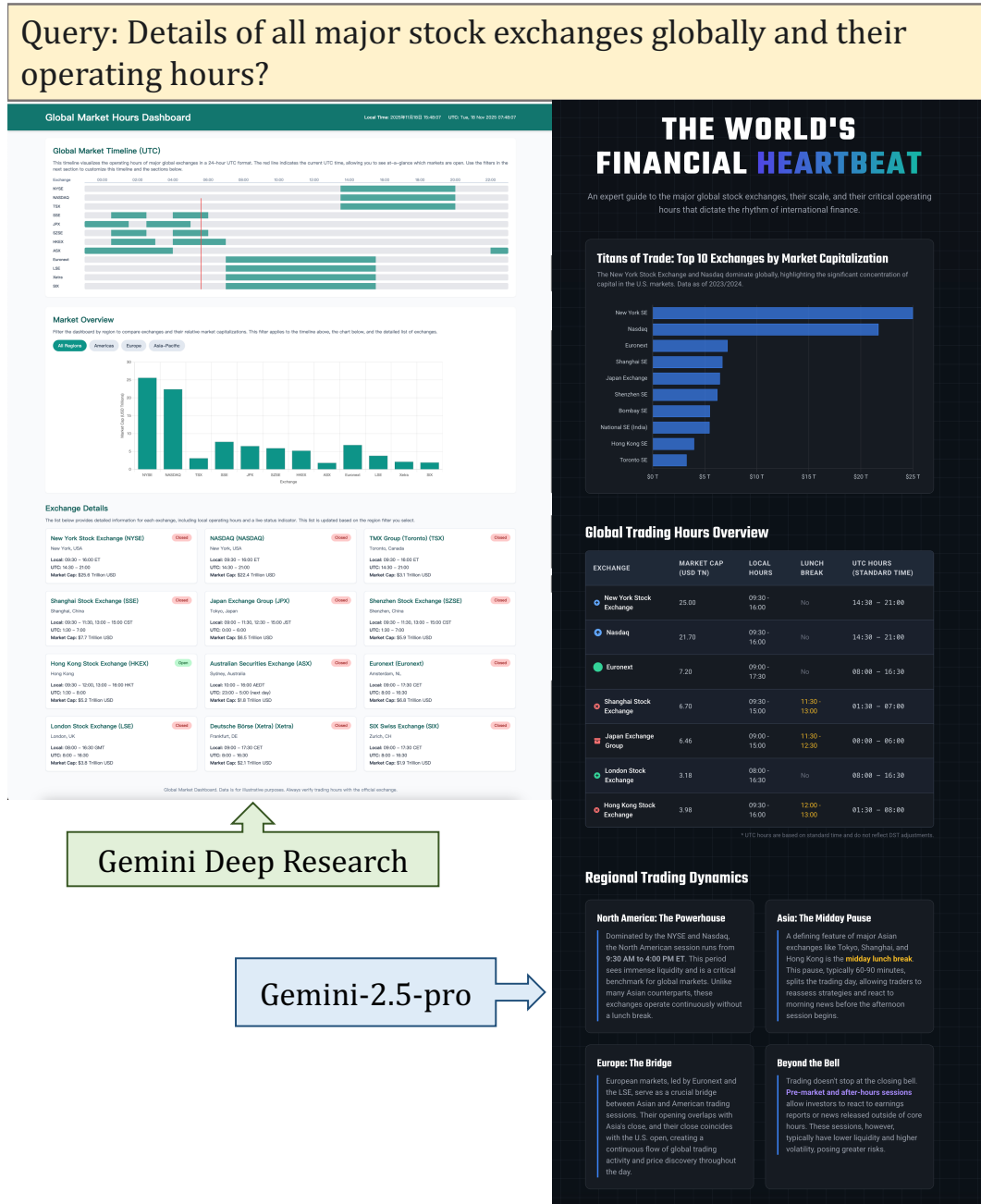


Figure 33: Responses from Gemini Deep Research and Gemini-RAViG (query 2).

2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267

Tongyi Deep Research (Qwen3 Max)

First half page

Global Exchange Trading Hours

Comprehensive Guide to Trading Hours of Major Global Stock Exchanges

The trading hours of major global stock exchanges show significant regional differences and time zone characteristics. These differences not only reflect the national development of regional financial markets but also demonstrate the mutual influence and coordination among different markets in the context of globalization. In terms of trading session distribution, European and American markets dominate daytime trading, Asia-Pacific markets retain the tradition of a midday break, while emerging markets exhibit more diverse time arrangements.

Global Exchange Overview

24 Hours of Global Trading	20+ Major Exchanges	12 Time Zones Covered	3 Major Global Indices
-------------------------------	------------------------	--------------------------	---------------------------

Trading Hours of Major Regional Exchanges

Asia-Pacific Region	European Markets	American Markets	Emerging Markets
<p>Japan Stock Exchange Monday to Friday 9:00-15:00 (Japan Time) Holiday Break: 12:00-12:30</p> <p>Hong Kong Stock Exchange Monday to Friday 10:00-12:00 & 13:00-16:00 (Hong Kong Time) Holiday Break: 12:00-12:30</p> <p>Tokyo, Seoul 9:00-11:30 & 12:30-15:00 (Japan Time) Holiday Break: 11:30-12:30</p>	<p>London Stock Exchange Monday to Friday 09:00-16:30 (London Time) Winter Time: 09:00-17:00 (London Time) No Midday Break</p> <p>Frankfurt Stock Exchange Monday to Friday 09:00-17:00 (Central European Time) No Midday Break</p> <p>Borsa Italiana Monday to Friday 09:00-17:00 (Central European Time) No Midday Break</p> <p>Amsterdam Monday to Friday 09:00-17:00 (Central European Time) No Midday Break</p>	<p>New York Stock Exchange Monday to Friday 09:30-16:00 (Eastern Time) No Midday Break</p> <p>NASDAQ Monday to Friday 09:30-16:00 (Eastern Time) No Midday Break</p> <p>Canada (TSX) Monday to Friday 09:30-16:00 (Eastern Time) No Midday Break</p> <p>Mexico (BMV) Monday to Friday 09:30-16:00 (Central City Time) No Midday Break</p> <p>Brazil (B3) Monday to Friday 10:00-17:30 (Brasilia Time) No Midday Break</p>	<p>South Africa Exchange (JSE) Monday to Thursday 09:00-15:00 (South Time) Closed Friday & Saturday</p> <p>South Korea (KOSPI) Monday to Friday 09:00-15:00 (Seoul Time) 3.5 hour overlap with London</p> <p>India (NSE/BSE) Monday to Friday 09:15-15:30 (India Time) No Midday Break</p> <p>Nigeria (NSE) Monday to Friday 10:00-16:00 (Lagos Time) No Midday Break</p>

Time Zone Differences & Conversion

Exchange Name	Local Trading Hours	UTC Conversion (Summer Time)	UTC Conversion (Winter Time)	Overlap/Session Mechanism
New York Stock Exchange (NYSE)	9:30-16:00	1:00-8:00(UTC-4)	2:00-8:00(UTC-5)	2nd Sun in Mar - 1st Sun in Nov
London Stock Exchange (LSE)	09:00-16:30	7:00-15:30(UTC+1)	8:00-16:00(UTC+0)	Last Sun in Mar - Last Sun in Oct
Hong Kong Stock Exchange (HKEX)	10:00-12:00, 13:00-16:00	1:30-4:00, 5:00-8:00(UTC+8)	1:30-4:00, 5:00-8:00(UTC+8)	No daylight saving
Tokyo Stock Exchange (TSE)	9:00-11:30, 12:30-15:00	0:00-2:30, 3:30-6:00(UTC+9)	0:00-2:30, 3:30-6:00(UTC+9)	No daylight saving
Seoul Stock Exchange (KRX)	10:00-12:00, 13:00-15:00	7:00-12:00(UTC+9)	7:00-12:00(UTC+9)	No daylight saving
Brazil Stock Exchange (B3)	10:00-17:30	5:00-12:30(UTC-3)	6:00-13:00(UTC-4)	1st Sun in Oct - 1st Sun in Apr of the next year
South Africa Stock Exchange (JSE)	9:00-17:00	7:00-15:00(UTC+2)	7:00-16:00(UTC+2)	No daylight saving

Important Reminder: European and American markets generally adopt daylight saving time, starting and ending one hour earlier from March to October each year. While most Asia-Pacific and Middle Eastern markets do not, trading sessions may be affected.

Investors need to pay special attention to these time zone differences to avoid trading errors. For example, during daylight saving time, the London Stock Exchange corresponds to 10:00-16:30 trading time, and during winter time, it is 11:00-17:00.

Global Interlinkage and Overlap Patterns of Trading Sessions

- Tokyo Time 9:00:** Asia-Pacific Market Opens. Tokyo Stock Exchange opens, corresponding to UTC 0:00. Asia-Pacific investors begin trading.
- London Time 8:00:** European Market Opens. London Stock Exchange opens, corresponding to UTC 7:00. Trading overlaps with Asia-Pacific market begins.

Second half page

New York Time 9:30

American Market Opens

New York Stock Exchange opens, corresponding to UTC 13:30. Overlap period with European market begins.

Sydney Time 16:00

Asia-Pacific Market Closes

Sydney Stock Exchange closes, corresponding to UTC 6:00. Global market enters overnight trading session.

Asia-Europe Overlap Period
Trading Time: 14:00-16:00
Overlap between NYSE's afternoon session and LSE's summer trading session. The two sessions allow Asian and European markets to trade simultaneously. An important channel for global capital flow.

US-Europe Overlap Period
Trading Time: 14:00-17:00
Overlap between NYSE's and LSE's summer trading sessions. **The period with the highest global liquidity and also the most volatile market.** Intensified trading between New York and London markets.

Trans-Pacific Window
Trading Time: 21:00-23:00
Connection between Sydney Stock Exchange's session and Chicago's overnight session. Provides opportunities for Asia-Pacific investors to participate in North American markets. A window for North American investors to understand Asia-Pacific market dynamics.

"Stagger-Style" Trading Chain
The overlap of global market trading hours forms a "stagger-style" trading chain, starting from the opening of the Asia-Pacific market (Tokyo 9:00, UTC 0:00), to the European market (London 8:00, UTC 7:00), then to the American market (New York 9:30, UTC 13:30), and finally to the closing of the Asia-Pacific market (Sydney 16:00, UTC 6:00), creating almost 24 hours a day.

Special Arrangements and Changing Trends in Trading Hours

Trend of Extending Trading Hours
In recent years, there has been a trend of extending trading hours of global exchanges, especially in emerging markets aiming to attract international investors.

- India plans to extend to 20:00.
- Frankfurt Stock Exchange extended to midday break.
- The South African Exchange extended trading hours.
- NASDAQ launched 24/7 trading model.

Technological Innovation Arrangements
Some exchanges have introduced innovative trading hour arrangements:

- Frankfurt Stock Exchange extended trading to 22:00.
- NASDAQ 24/7 trading for some tech stocks.
- Major market design weekly adjustment.
- Shortened market hours on Independence Day.

Fee Adjustments: The Abu Dhabi Securities Exchange (ADX) in the UAE cut its trading fee by 50% from 0.07% to 0.035%. This adjustment reflects the exchange fee reduction trend from this year and the first time in three years. This adjustment reflects the growing appeal of the UAE market to international investors and its progress in financial globalization.

International Cooperation: In November 2023, the Shanghai market signed a Memorandum of Understanding (MOU) with the Global Financial Market (GFM) to explore cooperation opportunities in areas such as index, ETFs, and ESG. This cooperation between China and the GFM also reflects the growing trend of collaboration among global exchanges.

Application of Global Trading Hours in Investment Strategies

Cross-Market Arbitrage
Investors can leverage time differences in global markets to capture arbitrage opportunities in international transactions and price discovery. During the Asia-Europe overlap period (14:00-16:00 Beijing Time), investors can simultaneously monitor the Hong Kong and London markets to perform arbitrage based on price discrepancies.

Time Difference Trading Strategy
Before the New York market opens, investors can monitor the London market's activity to predict the opening trend of the New York market.

- After the London market closes, investors can watch the New York market's activity to predict the opening trend of the Asian market the next day.

T+0/T+1 Settlement System Differences
Investors need to be aware of the differences in trading systems across markets:

- US Market: T+1 trading system.
- Asian Market: T+0 trading for most equities.
- European Market: T+1 trading system.

These system differences affect investor capital efficiency and cross-market trading strategies.

Special Market Opportunities

The special arrangements of global exchange trading hours also provide unique investment opportunities for investors:

- Small Stock Exchange:** Sunday/Thursday trading schedule, suitable for Middle Eastern investors.
- Hong Kong Exchange:** First extended trading hours, offering longer cross-market trading opportunities.
- Frankfurt Stock Exchange:** No midday break, enabling more continuous trading.
- NASDAQ 24/7:** 24-hour trading for some tech stocks, enhancing liquidity.

Future Development Trends of Global Exchange Trading Hours

Extension of Trading Hours
It is expected that global exchange trading hours will become more extended and continuous in the future to meet the demand for around-the-clock trading. The pilot program by the Hong Kong Exchange to extend trading hours to 20:00, if successful, may be extended to other Asia-Pacific markets such as the Singapore Exchange and the Tokyo Stock Exchange.

Deeper International Cooperation
The coordination of global exchange trading hours will also deepen. For example, the MOU signed between China and the Global Financial Market marks cooperation in areas like index, ETFs, and ESG. This reflects the growing trend of collaboration among global exchanges and provides investors with more diversified investment tools.

Technology-Driven Innovation: With the development of technology and the popularization of electronic trading, some exchanges may introduce more flexible trading time arrangements. For example, NASDAQ 24/7 trading model for certain tech stocks allows investors to trade these stocks outside of regular hours, significantly enhancing market liquidity.

Summary & Outlook: The scheduling of global stock exchange trading hours is an important reference for investors in global asset allocation. By studying the trading time patterns and characteristics of each market, investors can better plan their trading strategies, optimize fund operations, and achieve optimal global asset allocation. In the context of globalization, the coordination and extension of trading hours will become a trend, providing investors with a more flexible trading environment.

Disclaimer: Note: The content of this report is generated by Qwen AI and is for reference only. Investment involves risks, so enter the market with caution.

Figure 34: Responses from Tongyi Deep Research (query 2).

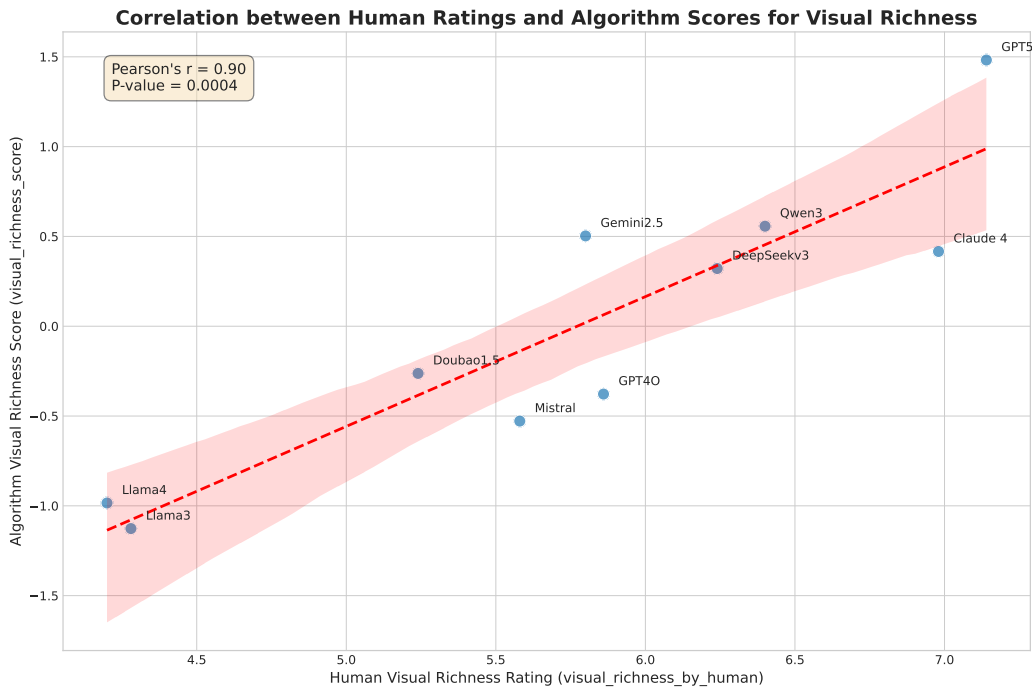


Figure 35: **Correlation bewteen human ratings and our VC_{score} .**

K HUMAN ANNOTATION OF VISUAL RICHNESS

We randomly selected 100 queries. The corresponding responses from each model were then rated by three experts on a 5-point scale (2, 4, 6, 8, 10) based on their subjective assessment of the visual richness of each answer. The final score for each response is the average of the three expert ratings. As shown in Fig. 35, A Pearson correlation analysis between our visual richness score VC_{score} and the averaged human ratings shows a strong, positive, and statistically significant correlation ($r=0.90$, $p=0.0004$). This indicates that our metric effectively captures the structural aspects of visual richness as perceived by humans.

L SENSITIVITY ANALYSIS OF THE SCORE CORRECTION MECHANISM

We analyzed how model rankings were affected by changes to key hyperparameters: α and the internal weights of VC_{score} . Firstly, we analyzed the effect of varying α from 0.1 to 1.0. As expected, model rankings showed moderate sensitivity to the penalty’s strength (Fig. 36). Crucially, this analysis revealed that our chosen value of $\alpha = 0.3$ falls within an optimal range (0.1–0.3) where the resulting rankings best align with human preferences. Secondly, we fixed the weight of table t_w and varied m_w from 0 to 0.9. We found that model rankings were highly robust to m_w , with Kendall’s $\tau > 0.95$ for the vast majority of its tested range (Fig. 37) This demonstrates that the overall signal of visual richness is strong enough that the final outcome is not dependent on the specific fixed weights.

M MORE ABOUT HUMAN-MACHINE AGREEMENT

M.1 GLOSSARY FOR HUMAN-MACHINE AGREEMENT EVALUATION

M.1.1 FLEISS’ KAPPA

Fleiss’ Kappa (κ) is a statistical measure that assesses the agreement reliability between three or more raters on categorical ratings (e.g., ”correct” vs. ”incorrect”). It quantifies the degree of agree-

2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375

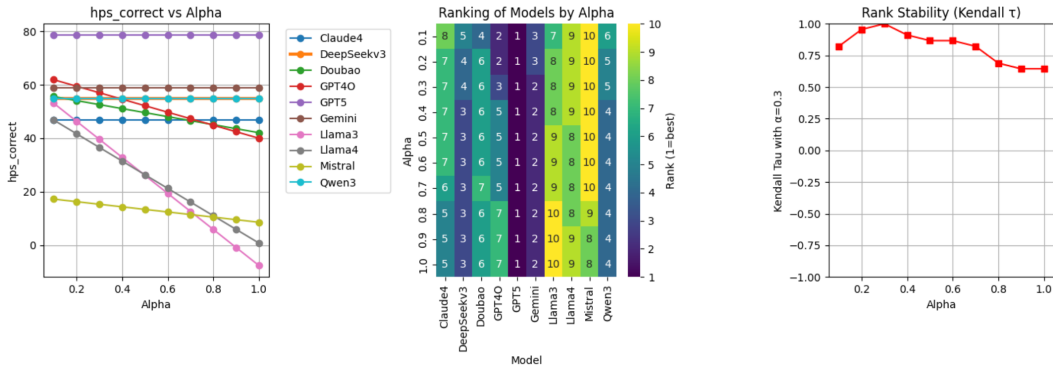


Figure 36: Changes in model scores and rankings under different α values.

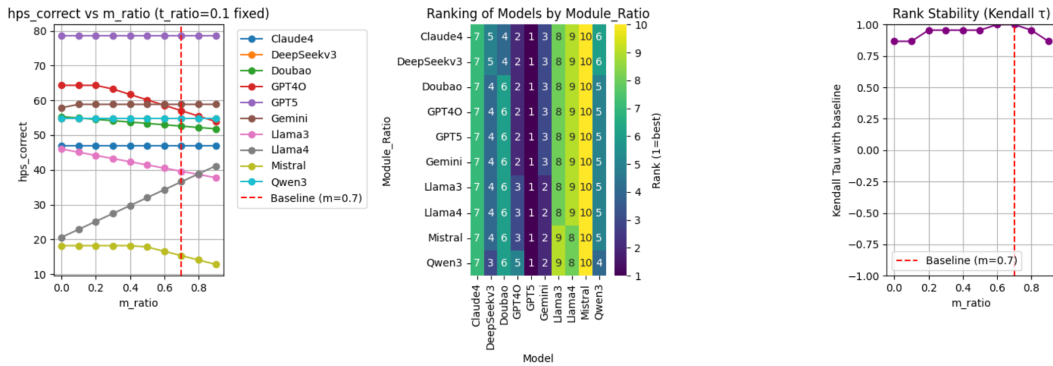


Figure 37: Changes in model scores and rankings under different m_w values.

ment that is above what would be expected by chance. The value of κ ranges from -1 to 1, where higher values indicate stronger agreement. We interpret the scores using established benchmarks:

- $\kappa \leq 0$: Poor agreement.
- $0.01 \leq \kappa \leq 0.20$: Slight agreement.
- $0.21 \leq \kappa \leq 0.40$: Fair agreement.
- $0.41 \leq \kappa \leq 0.60$: Moderate agreement.
- $0.61 \leq \kappa \leq 0.80$: Substantial agreement.
- $0.81 \leq \kappa \leq 1.00$: Almost perfect agreement.

M.1.2 INTRACLASS CORRELATION COEFFICIENT (ICC)

The Intraclass Correlation Coefficient (ICC) is a descriptive statistic used to measure the consistency of measurements made by multiple observers on the same set of subjects. It is particularly well-suited for continuous or ordinal data (e.g., scores from 1 to 10). The ICC value typically falls within the range of 0 to 1. Higher values signify greater reliability and consistency among raters.

- $ICC < 0.5$: Poor reliability.
- $0.5 \leq ICC < 0.75$: Moderate reliability.
- $0.75 \leq ICC < 0.9$: Good reliability.
- $ICC \geq 0.9$: Excellent reliability.

M.1.3 SPEARMAN’S RANK CORRELATION COEFFICIENT (SRCC)

Spearman’s Rank Correlation Coefficient (ρ) is a non-parametric measure of the monotonic relationship between two variables. It assesses how well the relationship between two variables can be

described using a monotonic function, making it ideal for ordinal data. The value of ρ is constrained to the interval $[-1, 1]$.

- $\rho = +1$: Indicates a perfect positive monotonic correlation.
- $\rho = -1$: Indicates a perfect negative monotonic correlation.
- $\rho = 0$: Indicates the absence of a monotonic correlation.
- The magnitude $|\rho|$ indicates the strength of the monotonic relationship.

M.1.4 CLASSIFICATION PERFORMANCE METRICS

We evaluate classification performance using the standard metrics of Accuracy, Precision, and Recall, which are based on the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

- **Accuracy** measures the proportion of all correct classifications.
- **Precision** measures the proportion of true positives among all instances predicted as positive.
- **Recall** measures the proportion of true positives among all actual positive instances.

M.2 ADDITIONAL RESULTS AND ANALYSIS ON HUMAN-MACHINE AGREEMENT

This part provides more detailed results of the agreement between our automated evaluation framework and human expert judgments in Tab. 7.

Metric	Design Quality				Content Quality		
	Missing	Occlusion	Color	Overflow	Reasonable	Faithfulness	Comprehensiveness
Fleiss' Kappa	0.980	0.966	0.853	0.927	0.852	0.833	-
Accuracy	0.975	0.940	0.945	0.950	0.930	0.940	-
Precision	0.990	0.980	0.990	0.960	0.925	1.000	-
Recall	0.961	0.907	0.908	0.941	0.942	0.945	-
ICC(3,k)	-	-	-	-	-	-	0.8883
SRCC	-	-	-	-	-	-	0.7770

Table 7: Detailed human-machine agreement metrics across dimensions

Analysis of the Results The high Fleiss' kappa (all ≥ 0.8) and ICC scores ($=0.8883$) demonstrate strong agreement among human experts and the reliability of the human evaluation results. The corresponding Accuracy, Precision, and Recall metrics reflect the effectiveness of our automatic evaluation framework.

N ANALYSIS OF JUDGE MODEL SELF-PREFERENCE

To validate the objectivity of our LLM-as-a-Judge evaluation framework, we conducted an analysis to investigate potential self-preference in our primary judge models on comprehensiveness scores: GPT-5 and Gemini-2.5-Pro. We chose comprehensiveness because it is the scoring dimension most susceptible to subjectivity. In this dimension, models are required to assign a score from 1 to 10 based on the perceived coverage and completeness of the answer. In contrast, the other dimensions are binary (0 or 1) and simply judge correctness, for which we have already validated accuracy. Self-preference occurs if a judge model unfairly favors outputs generated by itself or models from its own family (for example, GPT-5 favors GPT series).

Our methodology involved comparing the scores assigned by GPT-5 and Gemini-2.5-Pro to the outputs from a diverse set of ten models. A low correlation or systematic deviation in scores could indicate bias. Conversely, a strong linear correlation would suggest that both judges evaluate outputs based on similar, objective criteria, thus minimizing the impact of self-preference.

The evaluation scores assigned by the two judge models are detailed in Table 8.

Model Under Evaluation	GPT-5 Score	Gemini-2.5-Pro Score
Claude4	63.83	81.22
doubao1.5	59.73	72.20
DeepSeek-v3	61.60	74.79
Gemini2.5	69.28	86.17
GPT4O	61.98	71.97
GPT5	80.21	93.30
Llama3	56.86	63.54
Llama4	56.69	61.58
Mistral	58.55	68.44
Qwen3	62.49	75.49

Table 8: Comprehensiveness scores assigned by judge models GPT-5 and Gemini-2.5-Pro.

To assess the relationship visually and quantitatively, we plotted the average query-level scores from each judge against one another. The resulting scatter plot illustrates the degree of correlation between their evaluations, as shown in Figure 38.

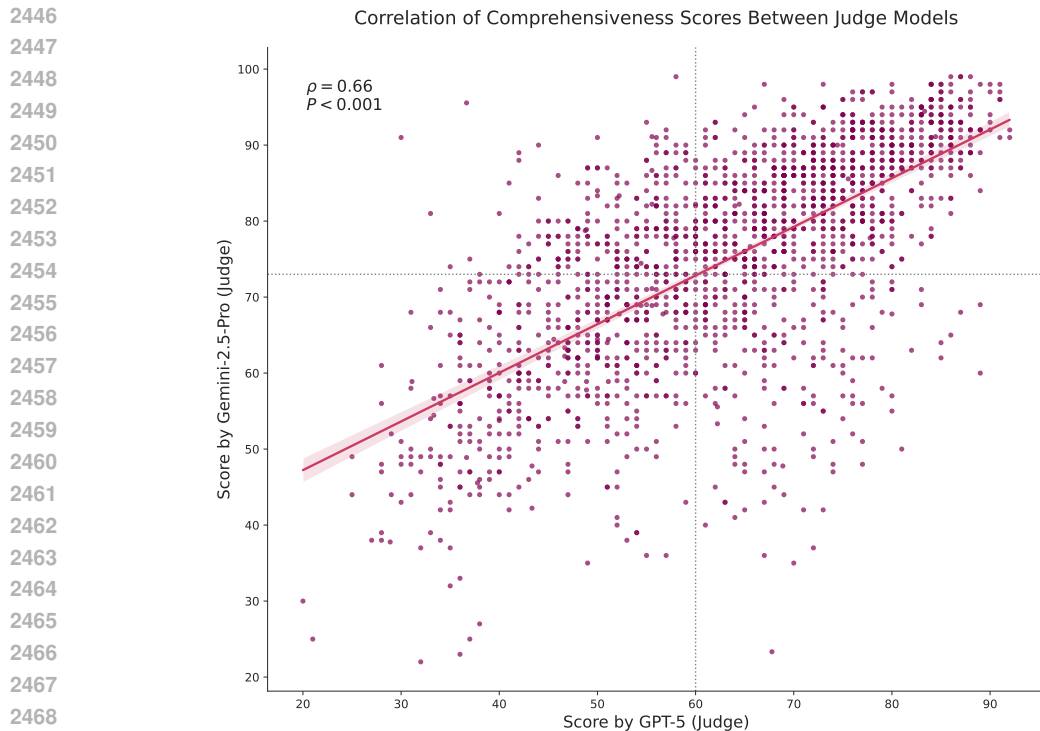


Figure 38: Correlation of comprehensiveness scores assigned by GPT-5 and Gemini-2.5-Pro. The solid line represents the linear regression fit the strong positive correlation (Pearson’s $r = 0.66$) suggests a high degree of agreement between the two judge models.

As illustrated in Figure 38, the comprehensiveness scores assigned by GPT-5 and Gemini-2.5-Pro exhibit a **strong and significant positive linear correlation**. The data points cluster tightly around the linear regression line, indicating a high degree of agreement in their relative ranking of the models. Quantitatively, the **Pearson correlation coefficient (r) is 0.66** ($p < 0.001$), which confirms the visual observation. This high level of inter-judge agreement suggests that both models are leveraging similar underlying criteria to assess the quality of the outputs.

While we cannot entirely rule out subtle biases, the strong cross-model consistency indicates that **systematic self-preference is not a dominant factor** in our evaluation. For example, GPT-5 assigns the highest score to its own output, but Gemini-2.5-Pro concurs, also giving GPT-5 the top score. This consistency across the board bolsters the credibility and objectivity of our LLM-as-a-Judge results, supporting its use as a reliable evaluation method for this task.

Judge Configuration	HPS(paper, GPT4o+Gemini2.5)	HPS(Gemini2.5+Gemini2.5)	HPS(GPT4o+GPT5)
GPT5	73.54	67.71	71.67
Gemini2.5	51.56	47.61	51.91
GPT4o	64.77	63.08	67.81
Qwen3	51.12	40.86	52.61
DeepSeekv3	51.03	48.51	57.17
Doubao1.5	56.68	51.94	55.38
Claude 4	45.86	39.71	46.01
Llama3	58.52	49.61	60.38
Mistral	16.19	15.00	17.13
Llama4	46.58	41.64	51.56

Table 9: Robustness Analysis Across Different Judge Configurations.

O ROBUSTNESS AND BIAS ANALYSIS OF THE LLM-AS-A-JUDGE PROTOCOL

Our framework is designed to minimize subjectivity. As detailed in the paper, all LLM-judged dimensions (except for Comprehensiveness) **rely on a binary "Yes/No" judgment based on strict criteria, not open-ended scoring**. This design inherently limits the potential for judge-specific bias. Appendix N reports a self-preference analysis for Comprehensiveness, showing no systematic self-preference bias. The LLM-as-a-judge paradigm itself is also well-established in prior work Zheng et al. (2023); Chen et al. (2025).

To demonstrate the robustness of our evaluation framework, we re-evaluated the responses from ten models to 100 randomly sampled queries (1,000 responses in total) *using a cross-family LLM-judge configuration*. As shown in Tab. 9 ("Model A + Model B" denotes Model A for design, Model B for content):

1) **High Ranking Stability**: The relative model rankings are highly stable across all judge configurations.

2) **No Self-Preference Bias**: When Gemini-2.5 acts as judge, it still scores its own performance significantly below GPT-5. Similarly, GPT-based judges do not artificially inflate the scores of other GPT models.

These results suggest that **our evaluation is not dominated by judge-specific biases, and the comparative ranking of models is robust**.

P ROBUSTNESS BETWEEN DETERMINISTIC AND MILDLY STOCHASTIC DECODING SETTINGS

We conducted an additional experiment to compare model performance under deterministic decoding (in our paper) versus a mildly stochastic decoding setting (temperature=0.7, top_p=0.9) on random sample of 100 queries. As shown in Tab. 10:

1) **Overall scores and rankings are stable**: The model rankings demonstrate high stability across both settings. This confirms the robustness and reliability of our evaluation framework.

2) **Larger fluctuations are driven by generation-side variability**: For instance, Gemini-2.5-pro's DSR improved notably. Our analysis traces this to a large reduction in "oversized elements" issues (from 25% to 13%), which comes from stochastic decoding helping the model avoid a deprecated CSS link that it consistently defaults to when temperature=0.

Q ADDITIONAL EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we provide the detailed result for our experiment.

2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591

	Paper Setting			Temp=0.7		
Metric	DSR	ECQ	HPS	DSR	ECQ	HPS
GPT5	0.77	95.5	73.5	0.76	94.1	71.5
Gemini2.5	0.61	84.5	51.6	0.77	81.2	62.5
GPT4o	0.92	70.4	64.8	0.94	68.6	64.5
Qwen3	0.73	70.0	51.1	0.66	73.9	48.8
DeepSeekv3	0.79	64.6	51.0	0.84	71.1	59.8
Doubao1.5	0.74	76.6	56.7	0.72	74.4	53.5
Claude 4	0.66	69.5	45.9	0.59	73.7	43.5
Llama3	0.91	64.3	58.5	0.90	63.4	57.1
Mistral	0.25	64.8	16.2	0.29	55.4	16.1
Llama4	0.73	63.8	46.6	0.73	64.0	46.7

Table 10: Comparison of Model Performance under Deterministic (Paper Setting) and Stochastic (Temperature=0.7) Decoding.

Q.1 DETAILED RESULT FOR DESIGN QUALITY TEST

Tab. 11 reports the number of issue cases for five types of design flaws across all models. The count of design quality failures is not equal to the sum of issue cases across dimensions because a single failure may involve multiple types of design flaws.

Error Type	GPT5*	Gemini2.5*	GPT4O*	Qwen3	DeepSeekv3	Doubao1.5*	Claude4*	Llama3	Llama4	Mistral
Oversized Element	145	505	60	80	58	18	233	0	1	2
Missing Element	11	13	36	33	34	68	39	151	187	103
Occlusion	13	22	21	48	59	146	25	45	21	77
Overflow	28	15	21	69	98	283	16	4	21	22
Low Color Contrast	131	30	53	182	65	20	386	12	6	1196
Design Quality Failures	320	555	171	337	237	439	620	207	227	1295

Table 11: Detailed of five design error types across all models

Q.2 DETAILED RESULT FOR CONTENT QUALITY TEST

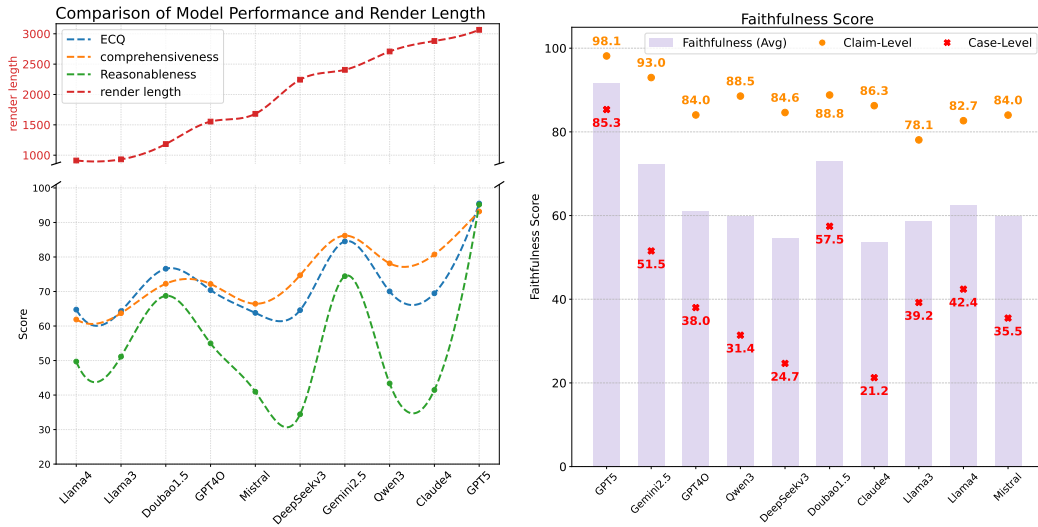
This part presents a detailed visualization of the content quality evaluation results, further exploring the relationships between different performance metrics and model characteristics like output length. The figures below provide a granular analysis of model capabilities.

Analysis of Figure 39(a) Figure 39(a) reveals that a longer output does not necessarily correlate with higher quality. While Comprehensiveness shows a weak positive correlation with render length, this trend is not absolute. For example, the increase in output length from Llama4 to DeepSeekv3 resulted in little improvement in the comprehensiveness score, and even showed a drop in the middle (Mistral). Furthermore, Reasonableness and the overall Effective Content Quality (ECQ) show little to no correlation with output length. Models such as GPT5 maintain high scores with long outputs, while others like Claude4 exhibit poor performance despite similar lengths. This demonstrates that our evaluation metrics effectively resist length bias and prioritize the quality of the content.

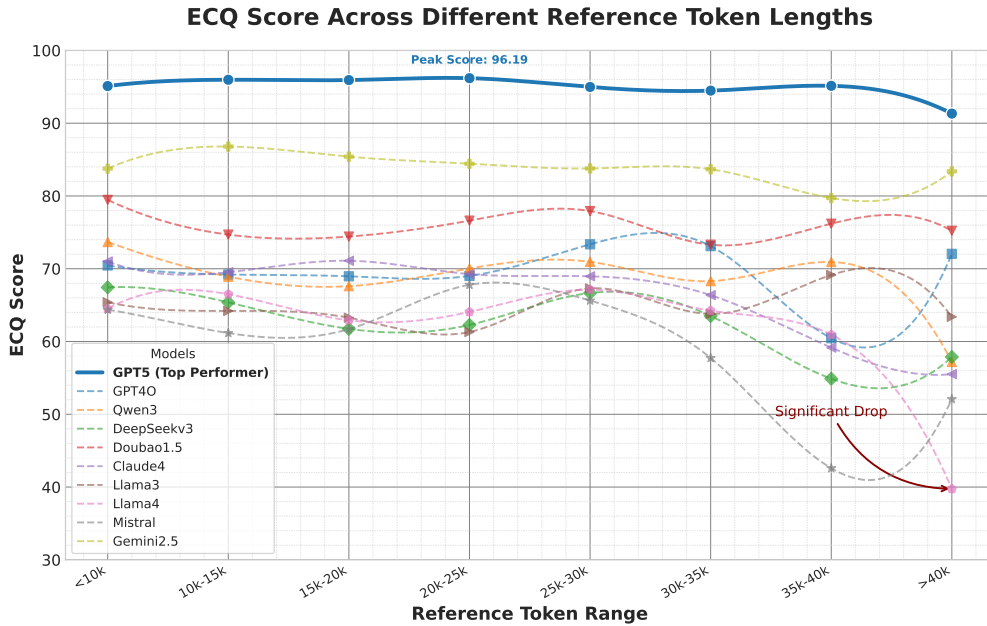
Analysis of Figure 39(b) Figure 39(b) provides a more granular view of Faithfulness scores. As highlighted in the main text, a significant drop is observed from the claim-level scores to the more stringent case-level scores for most models. This underscores a critical challenge: even a high accuracy on individual facts (claim-level) does not guarantee a fully trustworthy response (case-level), as a single error can undermine the entire output’s reliability.

Analysis of Figure 39(c) Figure 39(c) illustrates the performance of leading LLMs on the Effective Content Quality (ECQ) score, across a spectrum of reference text lengths, ranging from under 10k to over 40k tokens. The results reveal a clear performance hierarchy, with GPT5 establishing a commanding lead. It consistently achieves an ECQ score above 90 across all token ranges, significantly outperforming its competitors. In contrast, some models exhibit two distinct patterns: a lower performance ceiling and significant performance volatility. For instance, Llama4 and Qwen3 experience a dramatic drop-off, with Llama4’s score plummeting to 39.74 in the longest context window. This steep degradation suggests a failure to manage information effectively over extended

2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645



(a) Model performance and render length (b) Claim-Level vs. case-level faithfulness scores



(c) Model performance on ECQ score And Reference Tokens Range

Figure 39: Detailed visualization of content quality evaluation.

sequences, a common challenge known as "long context loss". Interestingly, this failure is not universal: models like Gemini2.5 and Doubao1.5 exhibit a performance rebound in this same range, pointing to diverse and variably effective strategies for handling extreme context lengths.