VISEXPRESS: BEYOND RENDERING– BENCHMARKING HOW LLM-GENERATED HTML ANSWERS CONVEY INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As the code-generation capabilities of Large Language Models (LLMs) advance, applications powered by LLMs increasingly present answers as HTML pages integrating text, charts, media, and interactive components. When appropriately designed, such responses can make information easier to understand and navigate. However, existing benchmarks primarily focus on code correctness, component quality, or design fidelity, paying limited attention to how effectively a complete HTML answer communicates information. To address this issue, we introduce VisExpress, a benchmark for evaluating the visual expression quality of complete LLM-generated HTML answers. VisExpress comprises 786 queries across 16 domains, combining real-user queries with curated queries that broaden the coverage of presentation tasks. It includes component-necessity annotations and covers 12 component forms organized into four categories: structured text, charts, media, and interactive components. We further develop a multimodal evaluation framework consisting of 13 metrics organized into two complementary dimensions. Design Normativity assesses the basic visual quality of page elements and layouts, while Presentation Effectiveness assesses whether component selection and information organization help users find and understand information. Using multimodal evidence—including screenshots, individual images, and recordings of animations and interactions—the framework evaluates both missing necessary representations and redundant components. Experiments across 12 models show that the model with the highest Design Normativity score trails the best model in Presentation Effectiveness by 9.6 points. While most models produce orderly layouts, component selection and information organization remain key challenges. The framework shows positive alignment with human judgments and matches the aggregate model ranking observed in our human study.

1 INTRODUCTION

LLM assistants increasingly move beyond linear text responses toward task-specific interfaces and interactive HTML answers (Chen et al., 2025; Zhang et al., 2026). Progress in code and front-end generation (Lu et al., 2025; Xiao et al., 2025) allows a single response to combine text with charts, images, animations, maps, and interactive tools. Recent user studies show that generative interfaces can be preferred over conventional conversational responses across diverse tasks (Chen et al., 2025), suggesting that answer quality also depends on how information is organized and presented.

However, visually polished HTML answers can still obscure key information, omit necessary representations, or introduce redundant components. Effective presentation requires component choices and information organization that match the user's needs. This aligns with multimedia learning principles that emphasize excluding extraneous material, highlighting essential information, and integrating related text and graphics (Fiorella & Mayer, 2021). Because open-ended queries admit multiple suitable presentations, reference-page similarity alone is insufficient: evaluation must consider both individual components and whether the complete answer is easy to follow and understand.

Existing benchmarks evaluate generated interfaces through functional testing, reference-based visual comparison, and multimodal assessment of executable artifacts (Lu et al., 2025; Xiao et al., 2025; Lin et al., 2025; Zhang et al., 2025). Recent work further examines whether interactive HTML ap-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with related HTML and web-generation benchmarks. ✓ denotes explicit coverage, ○ denotes partial or related coverage, and – denotes no explicit evaluation.

Benchmark	Evaluation target	Design Normativity		Presentation Effectiveness	
		Element Quality	Layout Regularity	Expression Coherence	Visual Expressiveness
WebGen-Bench	Generated websites	✓	✓	–	–
DesignBench	Front-end implementations	✓	✓	–	–
WebUIBench	WebUIs and HTML	✓	✓	–	–
StructEval	Structured outputs	○	○	–	–
WebGen-V	Generated webpages	✓	✓	–	○
ArtifactsBench	Executable visual artifacts	✓	✓	○	○
MiniAppBench	Interactive HTML applications	✓	✓	○	○
VisExpress(Ours)	**Multimodal HTML**	✓	✓	✓	✓

plications fulfill user intent and exhibit appropriate dynamic behavior (Zhang et al., 2026). These efforts provide a foundation for evaluating implementation and presentation quality, but pay less attention to how component selection and answer organization help users find and understand information. Table 1 compares these evaluation emphases with those of VisExpress.

To address this gap, we introduce VisExpress, a benchmark for evaluating complete HTML answers as information carriers. VisExpress comprises 786 queries across 16 domains and covers 12 component forms organized into four categories: structured text, charts, media, and interactive components. Queries are annotated with component necessity and reasons for each label, grounding evaluation in users' information needs. We further develop a multimodal evaluation framework that combines page-level screenshots, component images, and dynamic evidence to assess both individual components and their contribution to the complete answer. The framework includes 13 metrics across two dimensions. Design Normativity assesses basic visual quality through Element Quality and Layout Regularity. Presentation Effectiveness, comprising Expression Coherence and Visual Expressiveness, assesses whether component selection and information organization support efficient information access and comprehension. Together, these assessments account for omitted necessary representations and redundant components, while examining how visual hierarchy and reading progression help convey key information.

Our main contributions are summarized as follows:

- We introduce **VisExpress**, comprising 786 queries across 16 domains and 12 component forms, with component-necessity annotations that ground presentation evaluation in users' information needs.
- We develop a multimodal evaluation framework with 13 metrics across **Design Normativity** and **Presentation Effectiveness**. Combining page-level, component-level, and dynamic evidence, it assesses presentation quality, component selection, and information organization, including necessary representations that are missing and components that are redundant.
- Experiments across 12 models reveal a gap between visual quality and effective information presentation. Human evaluation yields the same aggregate ranking of four models and positive query-level agreement with VisExpress.

2 Related Work

HTML and Interface Generation Recent work enables LLMs to generate task-specific interfaces instead of presenting answers solely as text. Generative Interfaces translates user queries into adaptive UIs through structured representations and iterative refinement (Chen et al., 2025), while Generative UI combines prompting, tools, and post-processing to produce custom web pages (Leviathan

et al., 2026). Task-driven data models enable interface generation and interactive refinement (Cao et al., 2025). AlignUI further incorporates user preferences to guide interface generation (Liu et al., 2026). These methods motivate evaluating whether generated interfaces and multimodal HTML pages serve users' needs. These advances motivate evaluating how the design and presentation of generated interfaces support users' information needs.

Evaluation of Visual and Interactive Outputs DesignBench and WebUIBench use visual comparisons to evaluate interfaces (Xiao et al., 2025; Lin et al., 2025), while WebGen-V measures section-level textual, layout, and visual alignment (Wang et al., 2025). StructEval assesses format validity and content fidelity in structured outputs (Yang et al., 2025). ArtifactsBench combines temporal screenshots, source code, and task-specific checklists (Zhang et al., 2025); MiniAppEval uses browser exploration to assess intent fulfillment and static and dynamic behavior (Zhang et al., 2026). WebDevJudge examines automated judges against human preferences (Li et al., 2025). These approaches address aspects of visual quality, usability, and interaction. VisExpress jointly assesses foundational design quality and how component choices and information organization support comprehension, including missing necessary representations and redundant components (Table 1).

3 VISEXPRESS DATASET

3.1 DATASET COMPOSITION

Dataset Composition and Coverage. VisExpress contains 786 queries, combining real-user questions with supplementary queries curated to broaden the coverage of information needs and presentation tasks. We organize the dataset into 16 domains using the taxonomy established in prior work Wang et al. (2026), which was derived from an analysis of 60,000 real-user queries. Across these domains, the queries span fact lookup, comparison, process understanding, and task exploration. Different component types support these information needs: structured-text components organize precise, format-sensitive, or itemized information; chart components support numerical comparison and pattern identification; media components support visual or auditory recognition and explanations of dynamic processes; and interactive components support parameter exploration, task completion, and spatial reasoning. These four top-level component types encompass 12 component forms, detailed in Table 2 alongside the information needs they support.

Component Annotations and Coverage. For each query, we annotate the applicability of component types. Tables, code blocks, and lists are not separately annotated, while SVG animations and videos share the same animations label, yielding eight annotation types across the 12 component forms. Each component annotation records its type, necessity for the query, and reason for that necessity label. Necessity is labeled **Required** (necessary for an explicit requirement or core need), **Optional** (appropriate but not essential), or **Inappropriate** (unsuitable for the request). A query may have annotations for multiple component types that complement one another or offer al-

Table 2: Four top-level component types and 12 component forms, with corresponding presentation roles.

Top-level type	Component forms	Presentation role
Structured text	Tables, code blocks, and lists	Support precise comparison, formatted content, and itemized organization.
Charts	ECharts charts, SVG infographics, and candlestick charts	Convey numerical, structural, or financial time-series relations.
Media	Images, SVG animations, videos, and audio cards	Support visual recognition, process understanding, or audio presentation.
Interactive	Tool-based interaction and maps	Support parameter exploration, lightweight tasks, and access to spatial information.

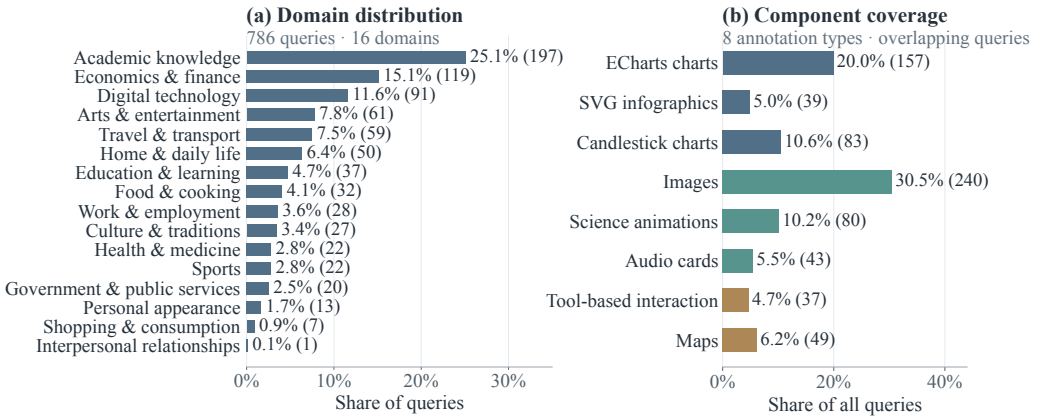


Figure 1: Domain distribution and component coverage.

ternative ways to meet the same information need; these annotations do not prescribe using all such components together.

Figure 1 shows domain distribution and component coverage. Coverage is the proportion of all 786 queries labeled Required or Optional for a component type; a query may contribute to multiple types, so these percentages cannot be summed. Appendix A provides complete annotation definitions and statistics.

3.2 Dataset Construction

We construct VisExpress in three stages: candidate collection, query curation, and multi-round annotation. Figure 2 illustrates this process using the black-hole query Q1.

Candidate Collection. We pool 2,463 presentation-oriented candidate queries assembled from diverse domain-specific sources and approximately 10,000 real-user queries. We identify substantive information needs, add context so each query can be understood independently, resolve ambiguity, and remove exact duplicates and queries with near-identical meanings.

Candidate Query Curation. We combine model-assisted screening with human review to select queries based on scenario representativeness, authenticity of information needs, and suitability for evaluation, while maintaining diversity in domains and evaluation points. Q1 asks what a black hole's surroundings would look like, expressing a need for visual information without specifying a component form.

Multi-round Annotation. For each query, we record its domain and component annotations. In each round, we randomly sample at least 5% of queries within each component subtype used during construction for human review. A query passes only if all its labels agree with human judgment; multi-round annotation stops when the overall pass rate among sampled queries exceeds 90%. For Q1, the domain is academic knowledge, the component annotation type is science animations, and the necessity is Optional because animation can help explain the accretion-disk structure and gravitational lensing, while other presentation forms can also meet the information need.

4 Evaluation Framework

Our evaluation framework evaluates the visual expression quality of HTML answers to open-ended queries, following three principles: **holism**, treating the complete answer as the unit of evaluation; **hierarchy**, distinguishing basic presentation quality, component strategy, and information organization; and **gain orientation**, judging components by their contribution to comprehension and information access. These principles guide two complementary dimensions, Design Normativity and Presentation Effectiveness, combining component-level assessments such as Chart quality with whole-answer assessments such as Content expressivity.

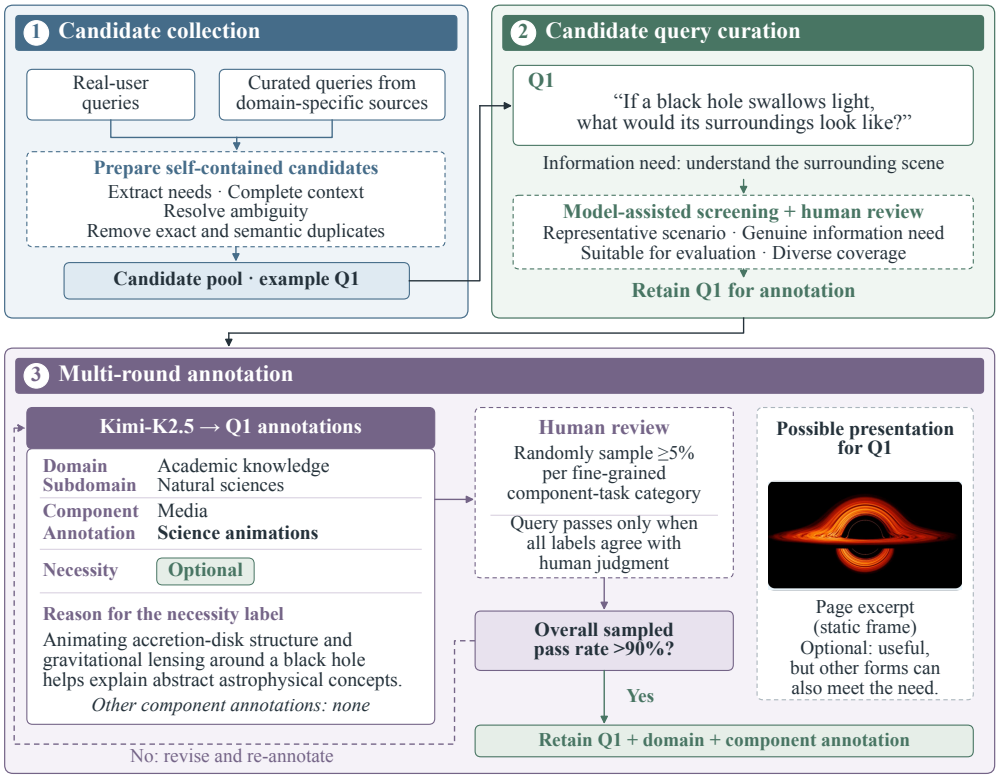


Figure 2: Three-stage construction pipeline for VisExpress. The black-hole query Q1 illustrates collection, curation, and multi-round annotation with human review.

4.1 Evaluation Dimensions

Design Normativity and Presentation Effectiveness comprise four subdimensions and 13 leaf-level metrics, defined in Appendix B. Figure 3 illustrates the hierarchy and scoring process.

Design Normativity. This dimension assesses the foundational quality of rendered elements and page layout through two subdimensions. **Element quality** concerns whether text, color, and individual components present information clearly and function as intended. It covers text rendering and typographic hierarchy, color readability and harmony, image clarity, chart readability and visual encoding, animation playback, and interaction behavior. **Layout regularity** concerns balanced spacing, consistent alignment, and effective use of space across page elements and within components. It examines whether related items form orderly groups and separate modules remain distinguishable, identifying crowding, misalignment, and excessive empty space that impede reading.

Presentation Effectiveness. This dimension assesses whether presentation helps users access key points, understand information, and follow a coherent reading progression. **Expression Coherence** concerns whether component selection and combination match the query, information structure, and surrounding content. It considers both missing necessary representations and redundant or unsuitable components, judging their information gain over plain text or other suitable forms. **Visual Expressiveness** concerns how the chosen content is organized into information priorities and a reading path. Among its leaf-level metrics, First-screen efficiency evaluates the initial presentation of useful content, while Content expressivity considers organization across the complete answer: whether core information has a clear visual focus, related content is grouped, and local emphasis guides readers from overviews to details. Thus, orderly spacing alone does not establish an effective information hierarchy, and richer component use receives credit only when it supports the reading task. Information gain is not scored as a separate item; it is a core judging principle of the leaf-level metrics in this dimension.

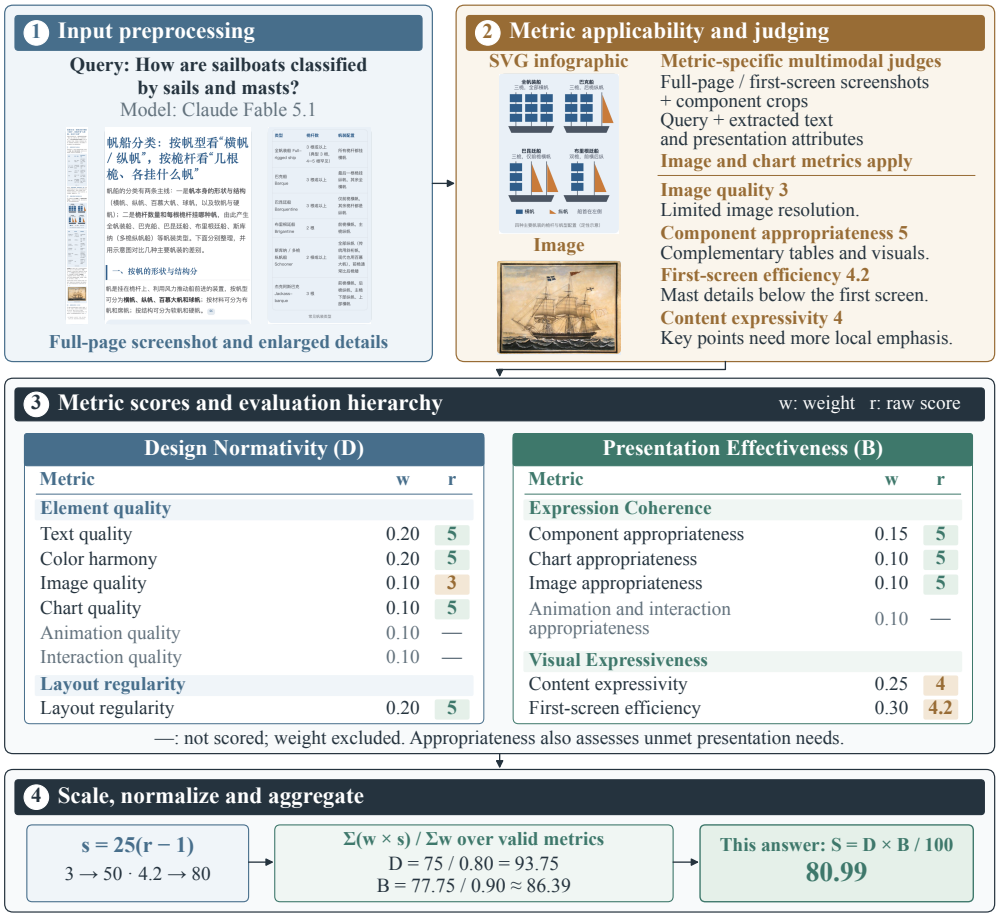


Figure 3: Evaluation framework illustrated with an HTML answer to "How are sailboats classified by sails and masts?" Metric scores are aggregated into Design Normativity, Presentation Effectiveness, and the overall score. Dashes indicate inapplicable metrics excluded from aggregation.

4.2 Evaluation Methodology

VisExpress uses metric-specific multimodal judges to assess page-level presentation and component performance.

Input Preprocessing. We receive rendered full-page screenshots, recordings of SVG animations and component interactions, and HTML source code. Screenshots are cropped into first-screen views and individual image/chart crops. From HTML, we extract answer text, component types, descriptions, planning information, animation and tool-interaction code, and presentation attributes such as text styles and colors for metric-specific evaluation.

Dynamic Metric Selection. Metrics are selected based on the answer's components and the query's needs; evaluation scopes are defined in Appendix B. Page-level metrics assess the answer as a whole, as in Text quality. First-screen efficiency covers only the first screen after loading and before scrolling. Component-level metrics assess each generated component of the relevant type separately, as in Chart quality. Appropriateness metrics also assess missing necessary representations. Type-specific appropriateness metrics are inapplicable only when both the component and a need for it are absent.

Scoring Criteria. Each metric uses a dedicated rubric. Most directly scored metrics use three levels: 1 (severe problems), 3 (basic fulfillment with remaining shortcomings), and 5 (full fulfillment). Severity reflects the extent to which a defect impairs readability, comprehension, or component use.

Text quality, Content expressivity, and Animation quality additionally use levels 2 and 4: a score of 2 indicates defects that noticeably impair reading or comprehension, while a score of 4 indicates clear and effective information presentation with minor remaining shortcomings. Appendix E illustrates these criteria through the core Chart appropriateness rubric and a worked example contrasting useful visual comparison with insufficient gain.

Metric Scores. For each metric, only valid scores within its applicable scope are included in aggregation. Page-level metrics assign scores within their evaluation scopes according to their scoring criteria. Metric scores may fall between discrete levels through averaging, as in Component appropriateness, or combining internal scores, as in First-screen efficiency. Component-level quality metrics average valid component scores, rounded to one decimal place. Appropriateness metrics also assess missing necessary components and combine judgments according to their respective rules; for example, a missing necessary image receives a score of 1 in Image appropriateness. Appendix B.3 summarizes the score-formation rules.

For answer i, internal scores z_{ijk} have weights α_{jk}. For $k = 1, 2, 3$, Layout regularity ($j = \mathrm{lr}$) uses spacing, alignment, and space utilization ($1/3$ each); First-screen efficiency ($j = \mathrm{fs}$) uses effective-content completeness ($z_{i,\mathrm{fs},1}$), focus, and component gain ($0.4, 0.3, 0.3$). We compute raw metric score r_{ij} below, rounding to one decimal place:

$$r_{ij} = \begin{cases} 1, & j = \mathrm{fs} \text{ and } z_{i,\mathrm{fs},1} = 1, \\ \sum_{k=1}^{3} \alpha_{jk} z_{ijk}, & \text{otherwise.} \end{cases} \tag{1}$$

The lowest first-screen completeness level cannot be offset by other strengths. Completeness is assessed relative to the query: a short factual question calls for a direct answer, while a complex explanation can begin with an overview or substantive points.

4.3 Overall Score Calculation

Rationale. Drawing on the Kano model's distinction between must-be and one-dimensional quality attributes (Kano et al., 1984), we treat basic presentation quality as a prerequisite for effective expression and compute the overall score from the product of the Design Normativity and Presentation Effectiveness scores. Leaf-level metric weights are manually configured according to their importance to user experience; details are provided in Appendix B.

Dimension scores. We linearly map the valid raw score r_{ij} obtained in Section 4.2 for leaf-level metric j on answer i to a 0–100 score as $s_{ij} = 25(r_{ij} - 1)$. Only valid raw scores in [1, 5] participate; raw 0 and missing or invalid scores are excluded. A raw score of 1 maps to a valid 0. Let w_j denote the metric's configured weight within its top-level dimension. For a top-level dimension or subdimension G, let $\mathcal{J}_i(G)$ denote the set of its leaf-level metrics with valid scores. Its total valid weight $W_i(G)$ and aggregate score $Q_i(G)$ are, respectively,

$$W_i(G) = \sum_{j \in \mathcal{J}_i(G)} w_j, \qquad Q_i(G) = \frac{\sum_{j \in \mathcal{J}_i(G)} w_j s_{ij}}{W_i(G)}, \quad W_i(G) > 0. \tag{2}$$

We aggregate subdimensions into a top-level dimension using their total valid weights, excluding those without valid scores. For example, if only Text quality and Color harmony have valid scores within Element Quality, its aggregation weight is 0.40 rather than the full configured weight of 0.80.

Overall score. For each evaluated model, let D_i and B_i denote the Design Normativity and Presentation Effectiveness scores, respectively, of its i-th answer generated on the VisExpress dataset, and let $\mathcal{I}$ denote the set of answers whose final evaluations are valid and whose two top-level dimensions both have valid scores. Answers affected by generation or evaluation errors are excluded if they remain invalid after retries; valid answers with an overall score of 0 are retained. For $i \in \mathcal{I}$, the answer-level overall score S_i and the evaluated model's final score on VisExpress $\overline{S}$ are

$$S_i = \frac{D_i B_i}{100}, \qquad \overline{S} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} S_i, \quad |\mathcal{I}| > 0. \tag{3}$$

Table 3: Main results on VisExpress. Overall, dimension, and subdimension scores are computed over each model's own valid-answer set following the aggregation procedure in §4.3. Bold denotes the highest score in each column. ↑ indicates higher is better.

Model	Valid # (rate)	VisExpress ↑	95% CI	Design Normativity ↑			Presentation Effectiveness ↑		
				Score	EQ	LR	Score	EC	VE
Claude Fable 5.1	786 (100.0%)	**75.87**	[74.90, 76.88]	87.91	86.61	91.67	**86.24**	**88.14**	**85.48**
GPT-6 Astra	783 (99.6%)	69.59	[68.40, 70.76]	**90.31**	**88.96**	93.57	76.63	78.70	76.47
Kimi-k3	786 (100.0%)	68.42	[67.29, 69.57]	88.70	87.37	92.01	77.01	76.94	77.61
DeepSeek V4.1 Flash	786 (100.0%)	68.08	[66.94, 69.22]	84.90	81.60	**94.17**	80.12	79.02	81.05
Hy4	786 (100.0%)	66.78	[65.58, 67.94]	81.09	77.00	92.47	82.07	81.05	82.87
Gemini 3.8 Flash	785 (99.9%)	65.76	[64.51, 67.00]	83.50	81.45	88.93	78.70	79.67	78.78
DeepSeek V4 Pro	786 (100.0%)	65.53	[64.31, 66.76]	87.69	85.89	92.14	74.54	72.92	76.02
GLM-5.3 Flash	775 (98.6%)	65.37	[64.05, 66.69]	85.39	83.36	90.65	76.08	75.57	76.72
GLM-5.3	780 (99.2%)	61.83	[60.63, 62.99]	85.26	82.66	91.64	72.40	71.30	73.40
Qwen 3.8-2.4T-A95B	781 (99.4%)	60.73	[59.52, 61.97]	74.97	68.05	93.45	81.06	81.62	81.16
Qwen 3.8 Flash	724 (92.1%)	60.54	[58.97, 62.10]	85.11	82.53	91.67	70.87	66.52	73.69
Gemini 3.1 Pro	786 (100.0%)	60.41	[59.13, 61.69]	81.95	80.25	86.47	73.50	74.38	73.24

EQ: Element Quality; LR: Layout Regularity; EC: Expression Coherence; VE: Visual Expressiveness.

Here, $D_i/100$ specifies the proportion of the Presentation Effectiveness score retained after adjustment by Design Normativity. For example, when $D_i = 80$, 80% of the Presentation Effectiveness score is retained.

5 Experiments

5.1 Experimental Setup

We evaluate 12 representative LLMs on all 786 VisExpress queries. Temperature is set to 0.5 where supported, and the maximum generation token limit is set to 24,000 for all models. Each model generates one answer per query, except for failure retries, without best-of-N sampling or post-hoc selection. Detailed model configurations are provided in Appendix C.1.

All models use a shared generation framework with two stages: generation and rendering. During generation, each model receives the user query, retrieved web references when available, and a common prompt specifying the available component interfaces and output format. The model determines the answer content, component selection and configuration, information hierarchy, layout, and color scheme, and outputs an HTML body containing answer text and component declarations. During rendering, the framework converts model-generated component declarations into rendered components and assembles a complete HTML page. To assess how HTML answers are presented in a mobile viewport, we use Playwright to render pages in Chromium with a viewport matching the client's content display area and capture full-page screenshots. The same viewport configuration is used for all evaluated models. For answers containing SVG animations or interactive components, we additionally record the animations or interaction sequences, respectively. Further details of the generation framework are provided in Appendix C.2.

5.2 Main Results

Table 3 summarizes the main results on VisExpress. Claude Fable 5.1 achieves the highest VisExpress score (75.87), followed by GPT-6 Astra (69.59). All models except Qwen 3.8 Flash have valid-answer rates of at least 98.6%, yet their visual expression quality varies substantially. The confidence intervals (CIs) describe sampling uncertainty within each model's valid-answer set, not repeated-run variation or pairwise significance. Detailed scores for all 13 leaf-level metrics are provided in Appendix C.6.

Design Normativity and Presentation Effectiveness do not necessarily align. GPT-6 Astra leads in Design Normativity, but Claude Fable 5.1 combines strong basic presentation with more effective component selection and information organization, leading GPT-6 Astra by 9.61 points in Presentation Effectiveness. VisExpress thus reveals distinct model strengths in basic presentation quality, component selection, and information organization, making the differences behind their overall scores explicit.

Orderly layouts coexist with uneven component quality. Layout Regularity is consistently high across models. GPT-6 Astra's charts are mostly complete and readable, whereas Gemini 3.8 Flash receives more deductions for obscured labels and missing or unreadable chart content, reflected in Chart Quality scores of 98.08 and 65.19, respectively. Animation and interaction present a broader challenge: all 12 models score lower on these quality metrics than on Layout Regularity (Table 16 in Appendix C.6), with observed problems including incomplete dynamic demonstrations and missing interaction feedback. The component results distinguish model-specific differences in static readability from shared limitations in dynamic and interactive presentation.

Component selection and whole-answer organization remain challenging. DeepSeek V4 Pro achieves high Image Quality (92.11) but low Image Appropriateness (30.50): its generated images are generally well presented, while necessary photographs for visual identification or comparison are sometimes omitted. At the whole-answer level, Claude Fable 5.1 shows stronger integration of text and visual components, a more coherent reading progression, and clearer emphasis on key conclusions. Kimi-k3 is more frequently penalized for weak emphasis on key information, consistent with the Content Expressivity gap (77.93 versus 63.26). Assessments of component selection and information organization capture omissions and weak visual priorities that quality assessments of generated components alone would miss.

5.3 Agreement with Human Judgments

We assess the alignment between VisExpress scores and human judgments through a ranking study of four models: Claude Fable 5.1, Kimi-k3, DeepSeek V4 Pro, and GLM-5.3. For each of the 786 queries, five distinct annotators from a pool of 60 independently rank complete HTML answers, allowing ties, using human-facing instructions derived from the VisExpress evaluation requirements. The instructions ask annotators for a holistic comparison using aligned criteria, rather than separate ratings of the 13 metrics. Model identities are hidden, and answer order is randomized. Mean assigned ranks define the human consensus ranking.

Following MT-Bench's S1/S2 agreement definitions (Zheng et al., 2023), we compare preferences induced by VisExpress scores with those derived from each annotator's ranking. On the 770 queries with valid scores for all four models and identical answers in both evaluations, agreement is 56.95% (95% CI: [55.80%, 58.10%]) when ties are included (S1), compared with 62.15% between annotators, a gap of 5.20 percentage points. Excluding comparisons involving a tie on either side gives 58.68% agreement (S2). Top-1 Agreement (the fraction of VisExpress top-ranked models also ranked best by humans) is 42.99%, above the 26.69% random reference accounting for human ties. At the aggregate model level, the human and VisExpress rankings are identical: Claude Fable 5.1 ranks first, followed by Kimi-k3, DeepSeek V4 Pro, and GLM-5.3. The aggregation analysis highlights the contribution of Presentation Effectiveness to agreement with human judgments (Appendix D.4). Appendix D details the protocols and statistics.

6 Conclusion

We introduced VisExpress, a benchmark for evaluating how effectively complete LLM-generated HTML answers convey information. It comprises 786 queries across 16 domains and a multimodal evaluation framework with 13 metrics spanning Design Normativity and Presentation Effectiveness. Experiments on 12 models show that strong visual quality does not ensure effective information communication: the leader in Design Normativity trails the best model in Presentation Effectiveness by 9.6 points. Component selection and whole-answer information organization remain key challenges. In the four-model human study, VisExpress recovers the aggregate human ranking and shows above-chance agreement at the query level. VisExpress provides a basis for evaluating and improving HTML answers that align presentation choices with users' information needs.

AI USE STATEMENT

In this work, we used generative AI tools to assist with and polish manuscript writing; support retrieval and discovery of relevant literature and resources; and generate synthetic candidate queries, support candidate screening, and annotate queries during dataset construction. All AI-assisted writing was reviewed and revised by the authors. Retrieved sources were checked by the authors before use, while synthetic candidate queries and model-assisted annotations underwent the curation and human-review procedures described in Section 3.2. We did not treat AI-generated text, search results, synthetic queries, or model-generated annotations as authoritative evidence. No AI system is listed as an author. The authors take full responsibility for the final text, claims, citations, data, and other artifacts in this work.

ETHICS STATEMENT

VisExpress evaluates the presentation quality of model-generated HTML answers; it does not evaluate factual correctness, safety, accessibility, or legal compliance. A high VisExpress score should therefore not be interpreted as evidence that an answer is trustworthy or suitable for high-stakes use. The candidate pool includes real-user queries, but the benchmark reports curated, self-contained queries rather than user-level analyses or profiles. Retrieved web materials and generated pages may still inherit factual errors, social biases, harmful content, or third-party rights restrictions from their sources and models. The multimodal models used for automatic evaluation (i.e., judge models) may likewise reproduce biases from their underlying models, and coverage of 16 domains does not imply universal coverage across populations, languages, or accessibility needs. In the human evaluation, model identities are hidden and answer order is randomized to reduce presentation and brand bias. These limitations should be considered when releasing, extending, or applying the benchmark.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the benchmark and experiments across the main paper, appendices, and supplementary material. Section 3 documents dataset composition and construction; Section 4 and Appendix B specify evaluation dimensions, metric applicability, scoring criteria, inputs, weights, and aggregation. Section 5 and Appendix C report the evaluated models, generation and judge configurations, rendering and capture pipeline, retry policies, validity rules, bootstrap procedure, and fine-grained results. Appendix D describes the human-evaluation protocol and agreement statistics. The supplementary material includes the generation and rendering implementation, the shared generation prompt, and the metric-specific evaluation prompts. The dataset accompanies the submission. Together, these materials identify the model versions and settings, processing steps, random seed, and reporting rules needed to reproduce the reported results.

REFERENCES

Yining Cao, Peiling Jiang, and Haijun Xia. Generative and malleable user interfaces with generative and evolving task-driven data model. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025. doi: 10.1145/3706598.3713285. URL https://doi.org/10.1145/3706598.3713285.

Jiaqi Chen, Yanzhe Zhang, Yutong Zhang, Yijia Shao, and Diyi Yang. Generative Interfaces for Language Models. *arXiv e-prints*, art. arXiv:2508.19227, August 2025. doi: 10.48550/arXiv.2508.19227.

Logan Fiorella and Richard E. Mayer. Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In Richard E. Mayer and Logan Fiorella (eds.), *The Cambridge Handbook of Multimedia Learning*, pp. 185–198. Cambridge University Press, 3 edition, 2021. doi: 10.1017/9781108894333.019. URL https://doi.org/10.1017/9781108894333.019.

Noriaki Kano, Nobuhiko Seraku, Fumio Takahashi, and Shin-ichi Tsuji. Attractive quality and must-be quality. *Journal of The Japanese Society for Quality Control*, 14(2):147–156, 1984. doi: 10.20684/quality.14.2_147. In Japanese.

Yaniv Leviathan, Dani Valevski, Matan Kalman, Danny Lumen, Eyal Segalis, Eyal Molad, Shlomi Pasternak, Vishnu Natchu, Valerie Nygaard, Srinivasan Venkatachary, James Manyika, and Yossi Matias. Generative UI: LLMs are effective UI generators. *arXiv preprint arXiv:2604.09577*, 2026. doi: 10.48550/arXiv.2604.09577. URL https://arxiv.org/abs/2604.09577.

Chunyang Li, Yilun Zheng, Xinting Huang, Tianqing Fang, Jiahao Xu, Lihui Chen, Yangqiu Song, and Han Hu. WebDevJudge: Evaluating (M)LLMs as Critiques for Web Development Quality. *arXiv e-prints*, art. arXiv:2510.18560, October 2025. doi: 10.48550/arXiv.2510.18560.

Zhiyu Lin, Zhengda Zhou, Zhiyuan Zhao, Tianrui Wan, Yilun Ma, Junyu Gao, and Xuelong Li. WebUIBench: A Comprehensive Benchmark for Evaluating Multimodal Large Language Models in WebUI-to-Code. *arXiv e-prints*, art. arXiv:2506.07818, June 2025. doi: 10.48550/arXiv.2506.07818.

Yimeng Liu, Misha Sra, and Chang Xiao. AlignUI: A method for designing LLM-generated UIs aligned with user preferences. *arXiv preprint arXiv:2601.17614*, 2026. doi: 10.48550/arXiv.2601.17614. URL https://arxiv.org/abs/2601.17614.

Zimu Lu, Yunqiao Yang, Houxing Ren, Haotian Hou, Han Xiao, Ke Wang, Weikang Shi, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. WebGen-Bench: Evaluating LLMs on Generating Interactive and Functional Websites from Scratch. *arXiv e-prints*, art. arXiv:2505.03733, May 2025. doi: 10.48550/arXiv.2505.03733.

Kuang-Da Wang, Zhao Wang, Yotaro Shimose, Wei-Yao Wang, and Shingo Takamatsu. WebGen-V Bench: Structured Representation for Enhancing Visual Design in LLM-based Web Generation and Evaluation. *arXiv e-prints*, art. arXiv:2510.15306, October 2025. doi: 10.48550/arXiv.2510.15306.

Wenjie Wang, Yufeng Jiang, Ge Sun, Chenghang Dong, Zheng Jun, Li Mengjie, Lixin Chen, Huan Wang, Haoyu Wang, and Bin Chen. Omnibench: A comprehensive benchmark integrating real-world, time-sensitive, and multi-hop questions with a multi-dimensional hybrid evaluation framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):33657–33665, 2026. doi: 10.1609/aaai.v40i40.40655.

Jingyu Xiao, Ming Wang, Man Ho Lam, Yuxuan Wan, Junliang Liu, Yintong Huo, and Michael R. Lyu. DesignBench: A Comprehensive Benchmark for MLLM-based Front-end Code Generation. *arXiv e-prints*, art. arXiv:2506.06251, June 2025. doi: 10.48550/arXiv.2506.06251.

Jialin Yang, Dongfu Jiang, Lipeng He, Sherman Siu, Yuxuan Zhang, Disen Liao, Zhuofeng Li, Huaye Zeng, Yiming Jia, Haozhe Wang, Benjamin Schneider, Chi Ruan, Wentao Ma, Zhiheng Lyu, Yifei Wang, Yi Lu, Quy Duc Do, Ziyan Jiang, Ping Nie, and Wenhu Chen. StructEval: Benchmarking LLMs' Capabilities to Generate Structural Outputs. *arXiv e-prints*, art. arXiv:2505.20139, May 2025. doi: 10.48550/arXiv.2505.20139.

Chenchen Zhang, Yuhang Li, Can Xu, Jiaheng Liu, Ao Liu, Changzhi Zhou, Ken Deng, Dengpeng Wu, Guanhua Huang, Kejiao Li, Qi Yi, Ruibin Xiong, Shihui Hu, Yue Zhang, Yuhao Jiang, Zenan Xu, Yuanxing Zhang, Wiggin Zhou, Chayse Zhou, and Fengzong Lian. ArtifactsBench: Bridging the Visual-Interactive Gap in LLM Code Generation Evaluation. *arXiv e-prints*, art. arXiv:2507.04952, July 2025. doi: 10.48550/arXiv.2507.04952.

Zuhao Zhang, Chengyue Yu, Yuante Li, Chenyi Zhuang, Linjian Mo, and Shuai Li. MiniApp-Bench: Evaluating the Shift from Text to Interactive HTML Responses in LLM-Powered Assistants. *arXiv e-prints*, art. arXiv:2603.09652, March 2026. doi: 10.48550/arXiv.2603.09652.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Label Definitions and Dataset Statistics

This appendix describes the data fields, component taxonomy, and annotation definitions of VisExpress, and reports domain distribution, component coverage, and component necessity statistics. The dataset contains 786 queries and 731 component annotations.

A.1 Data Fields and Annotation Definitions

Each record contains a query, its domain, and 0–3 component annotations. Each component annotation specifies a component type, its necessity for the query, and the reason for that necessity label.

Table 4: Data fields and annotation levels.

Level	Field	Meaning
Query	Query ID	A unique query identifier for linking fields and counting distinct queries
	Query	The user's information need and presentation constraints
	Primary domain	The main knowledge or application domain, drawn from 16 categories
	Secondary domain	A more specific topic or application scenario within the primary domain
Component	Top-level component type	Chart, media, or interactive components; structured-text components are not separately annotated
	Component annotation type	One of the eight component annotation types defined in Table 5
	Component necessity	The necessity of the particular component for the query (Table 6)
	Reason for necessity label	A textual explanation of why the component is labeled Required, Optional, or Inappropriate for the query

Component taxonomy. Of the 12 component forms listed in Section 3.1, tables, code blocks, and lists are not separately annotated, while SVG animations and videos share the science animations label. This yields the eight component annotation types in Table 5.

Table 5: Component annotation types and their information presentation scope.

Top-level type	Annotation type	Presentation tasks and example information needs
Charts	ECharts charts	Composition, comparison, ranking, distribution, and trends in numerical data, including statistical charts such as pie, bar, and line charts
	SVG infographics	Processes, hierarchies, composition, timelines, intervals, and non-numerical relationships, organized through graphics, spatial grouping, and textual annotations
	Candlestick charts	Financial price time series and related data relationships, including open, close, high, and low prices and trading volume

Continued on next page

Table 5 (continued)

Top-level type	Annotation type	Presentation tasks and example information needs
Media	Images	Depictions of objects, scenes, and spatial forms, with related information organized through text–image correspondence and structured arrangements
	Science animations	Demonstrations of dynamic processes and scientific mechanisms using SVG animations or videos; examples include state changes, process order, and dynamic causal relationships
	Audio cards	The need for audio cards in sound-related scenarios such as language learning, pronunciation, and listening, and their integration with page content
Interactive	Tool-based interaction	Interaction entry points and state feedback needed for parameter entry, calculations, conversions, and lightweight interactive tasks
	Maps	Location lookup, route planning, itinerary organization, and spatial decisions; examples include locations, distances, and relationships among places

Component necessity. Table 6 defines the three necessity labels. Applicable components may be complementary or alternative; not all must appear in an answer.

Table 6: Component necessity labels.

Label	Definition
Required	The component is necessary to satisfy an explicit presentation requirement or a core information need
Optional	The component is a reasonable presentation option that can help convey information, but other forms can also satisfy the need
Inappropriate	The component conflicts with the presentation requirements or adds redundancy and presentation overhead without providing task-relevant information gain

A.2 Dataset Statistics

Domain distribution. Each query has one primary-domain label, and the percentages in Table 7 use all 786 queries as the denominator.

Table 7: Primary-domain distribution. N: query count.

Domain	N	Share	Domain	N	Share
Academic knowledge	197	25.06%	Work & employment	28	3.56%
Economics & finance	119	15.14%	Culture & traditions	27	3.44%
Digital technology	91	11.58%	Health & medicine	22	2.80%
Arts & entertainment	61	7.76%	Sports	22	2.80%
Travel & transport	59	7.51%	Government & public services	20	2.54%
Home & daily life	50	6.36%	Personal appearance	13	1.65%
Education & learning	37	4.71%	Shopping & consumption	7	0.89%
Food & cooking	32	4.07%	Interpersonal relationships	1	0.13%
Total				**786**	**100.00%**

Component coverage and necessity. Component coverage is the share of 786 queries annotated Required or Optional for that component. Top-level coverage uses the union of applicable queries across subtypes; overlapping query sets mean coverage rates are not additive.

Table 8: Component necessity and query coverage. Req., Opt., and Inapp. denote Required, Optional, and Inappropriate, respectively.

Top-level type / annotation type	Component annotations				Query coverage	
	Req.	Opt.	Inapp.	Total	Count	Share
Charts (overall)					**274**	**34.86%**
ECharts charts	98	59	3	160	157	19.97%
SVG infographics	2	37	0	39	39	4.96%
Candlestick charts	15	68	0	83	83	10.56%
Media (overall)					**339**	**43.13%**
Images	57	183	0	240	240	30.53%
Science animations	3	77	0	80	80	10.18%
Audio cards	11	32	0	43	43	5.47%
Interactive (overall)					**86**	**10.94%**
Tool-based interaction	5	32	0	37	37	4.71%
Maps	28	21	0	49	49	6.23%
Total	**219**	**509**	**3**	**731**	—	—

Of the 731 explicit component annotations in Table 8, 29.96% are Required, 69.63% Optional, and 0.41% Inappropriate. Independently rounded percentages may not sum to 100%.

B Evaluation Details

B.1 Metric Definitions and Weights

Table 9: Definitions, evaluation scopes, and weights of the VisExpress metrics, organized into two top-level dimensions and four subdimensions, with 13 leaf-level metrics.

Metric (weight)	Definition and evaluation scope
Design Normativity	
Element quality (total weight: 0.80)	
Text quality (0.20)	Assesses the rendering correctness and structural completeness of page text (excluding text within images and charts), and whether font sizes, typefaces, and typographic hierarchy support clear reading. *Scope:* Page-level (text).
Color harmony (0.20)	Assesses whether the overall page color scheme follows general design conventions to keep content readable, ensure visual comfort, and maintain a clean, harmonious interface; color contrast and semantic use of color should support information hierarchy and emphasis. *Scope:* Page-level.

Continued on next page

Table 9 (continued)

Metric (weight)	Definition and evaluation scope
Image quality (0.10)	Assesses image clarity and visual quality, including interference from watermarks or promotional overlays and the presence of inappropriate content. *Scope:* Component-level (images).
Chart quality (0.10)	Assesses whether charts follow basic data visualization principles, with complete structures, clear data and annotations, and appropriate proportions, and whether they clearly convey data relationships or logical connections in a process without misleading visual representations. *Scope:* Component-level (ECharts charts, SVG infographics, candlestick charts).
Animation quality (0.10)	Assesses the readability, completeness, and stability of animated demonstrations during playback, including technical defects and design problems that affect information access or viewing quality. *Scope:* Component-level (SVG animations, videos).
Interaction quality (0.10)	Assesses whether operations in tool-based interaction effectively support understanding principles or completing tasks, focusing on interaction logic, immediacy of feedback, and smoothness and continuity of operation. *Scope:* Component-level (tool-based interaction).
Layout regularity (total weight: 0.20)	
Layout regularity (0.20)	Assesses whether page elements and modules have balanced spacing, consistent alignment, and appropriate use of space, so that information blocks are clearly separated and orderly, without crowding or excessive spacing. *Scope:* Page-level.
Presentation Effectiveness	
Expression Coherence (total weight: 0.45)	
Component appropriateness (0.15)	Assesses whether the complete answer selects suitable forms for expressing specific information, considering component necessity, how well candidate forms match information complexity, and how well their placement fits the surrounding context, to identify redundant use and missing necessary representations. *Scope:* Page-level (all component types).
Chart appropriateness (0.10)	Assesses whether chart forms match the data types, data characteristics, and core information needs, and whether they clearly and efficiently convey structured information, data relationships, or complex concepts to support understanding, insight, or decision-making. *Scope:* Type-specific (ECharts charts, SVG infographics, candlestick charts).
Image appropriateness (0.10)	Assesses how well images match the text and user needs in theme, emotion, style, and imagery, and their incremental value for factual confirmation, supplementary information, atmosphere, or creative expression. *Scope:* Type-specific (images).

Continued on next page

Table 9 (continued)

Metric (weight)	Definition and evaluation scope
Animation and interaction appropriateness (0.10)	Assesses the necessity of animated demonstrations and interactive components, their fit with the text and user intent, and their incremental value for conveying information, judging whether dynamic demonstrations or interaction aid understanding, exploration, or task completion, and identifying misuse, redundancy, or missing necessary representations. *Scope:* Type-specific (SVG animations, videos, tool-based interaction).
Visual Expressiveness (total weight: 0.55)	
Content expressivity (0.25)	Assesses whether visual elements, layout, and information hierarchy help the complete answer communicate efficiently and clearly with appropriate pacing, creating a coherent reading experience by highlighting key points and connecting overviews with details, while preventing unhelpful visual richness from interfering with information access. *Scope:* Page-level.
First-screen efficiency (0.30)	Assesses content presentation on the first screen after the answer has loaded and before scrolling: whether it provides effective content suited to the current task, highlights key information, and uses components that provide incremental value for conveying information, thereby supporting rapid information access and a clear direction for further reading. *Scope:* Page-level (first screen).

Metric weights sum to 1 within each top-level dimension; each subdimension weight is the sum of its constituent metric weights. Applicability, scoring, and aggregation follow Sections 4.2–4.3. Audio content and downstream image retrieval strategies are outside the evaluation scope.

B.2 Evaluation Inputs

Table 10: Evaluation inputs for each metric.

Metric	Query	Page	1st	Comp.	Text	Code	Prior
Text quality	✓	✓	–	–	✓	–	–
Color harmony	✓	✓	–	–	✓	–	–
Image quality	–	–	–	✓	✓	–	–
Chart quality	–	–	–	✓	✓	○	–
Animation quality	–	–	–	✓	–	○	–
Interaction quality	–	–	–	✓	✓	✓	✓
Layout regularity	–	✓	–	–	–	–	–
Component appropriateness	✓	✓	–	–	✓	–	–
Chart appropriateness	✓	✓	–	–	✓	–	–
Image appropriateness	✓	✓	–	–	✓	–	–
Animation and interaction appropriateness	✓	✓	–	–	✓	–	–
Content expressivity	✓	✓	–	–	✓	–	✓
First-screen efficiency	✓	–	✓	–	✓	–	–

Here, *Page* denotes the full-page screenshot, *1st* the first-screen screenshot, *Comp.* component visual inputs, *Text* text and structured information, *Code* component source code, and *Prior* internal preliminary assessment results. ✓ denotes use, ○ conditional use, and – no use. Component visual inputs comprise image/chart crops and animation or interaction videos. Source code is included when available for charts and SVG animations. For Interaction quality, source code is used in the preliminary stage; the final stage uses the recording, structured component information, and preliminary findings.

Component outlines are formulated by the evaluated model and extracted from the generated HTML to describe component content and presentation plans; complete page HTML is not passed directly to the judge. Internal preliminary assessments include basic interaction quality assessment within Interaction quality, and Color harmony and component use assessments within Content expressivity. Their results only support judgments within the respective metrics and are not additional weighted metrics.

B.3 Formation of Metric Scores

Table 11 summarizes how rubric judgments form one raw score per applicable metric and answer.

Table 11: Formation of raw metric scores.

Metric and evaluation scope	Score-formation rule
Text quality, Color harmony, Content expressivity *Page-level*	A single score according to the metric's rubric.
Image, Chart, Animation, Interaction quality *Component-level*	Mean of valid scores for generated components of the corresponding type.
Layout regularity *Page-level*	Mean of spacing, alignment, and space-utilization scores, each in [1, 5] (Section 4.2).
First-screen efficiency *Page-level (first screen)*	Weighted combination of completeness, focus, and component gain, each scored 1, 3, or 5, subject to the completeness constraint in Section 4.2.
Component appropriateness *Page-level (all component types)*	Average generated-component scores within each type, retain omission assessments as separate records, then average all type-level and omission records. A type can contribute both a generated-component record and an omission record. If no component is generated or needed, the score is 5.
Image and Chart appropriateness *Type-specific*	The most severe applicable problem across existing components and unmet needs determines the score (1, 3, or 5); component judgments are not averaged.
Animation and interaction appropriateness *Type-specific*	The lower of the omission-strategy score (1 for severe omission, 3 for minor omission, 5 for none) and the generated-component mean. An unmet need with no relevant component scores 1.

Means and composite metric scores are rounded to one decimal place; Component appropriateness rounds at both averaging stages. Type-specific appropriateness is inapplicable when neither a corresponding component nor a need for one is present.

C Experimental Details and Supplementary Results

This appendix documents the generation and evaluation settings, statistical procedures, common-query robustness analysis, and fine-grained results for the experiments in §5. Metric definitions, applicability rules, and score aggregation are specified in §4 and Appendix B.

C.1 Evaluated Models and Generation Configurations

We evaluate 12 models on all 786 VisExpress queries using a shared generation framework, prompt template, and component interfaces. Table 12 summarizes the model identifiers and inference settings. We request disabled reasoning for the DeepSeek models and Hy4, and use low reasoning effort for the remaining models; Claude uses adaptive thinking with low effort. These provider-specific controls do not imply equivalent reasoning budgets across models. The configured generation token limit is 24,000 for all models. Temperature is omitted for GPT-6 Astra, which does not support this parameter, and Claude Fable 5.1, which only accepts its default setting; all other models use a temperature of 0.5.

The generation framework does not perform best-of-N sampling or score-based answer selection. Within each batch, a model–query pair is generated once unless its execution fails, and completed cases are not automatically regenerated. Output parsing, component realization, and rendering follow the shared procedures described in Appendix C.2.

Table 12: Evaluated models and generation configurations. "Not set" indicates that the parameter is omitted from the request. Token limits follow provider-specific API semantics.

Model	Reasoning	Temperature	Max tokens
DeepSeek V4 Pro	Off	0.5	24,000
DeepSeek V4.1 Flash	Off	0.5	24,000
Qwen 3.8 Flash	Low	0.5	24,000
Qwen 3.8-2.4T-A95B	Low	0.5	24,000
Kimi-k3	Low	0.5	24,000
GLM-5.3	Low	0.5	24,000
GLM-5.3 Flash	Low	0.5	24,000
Hy4	Off(No-think)	0.5	24,000
GPT-6 Astra	Low	Not set	24,000
Claude Fable 5.1	Low	Not set	24,000
Gemini 3.1 Pro	Low	0.5	24,000
Gemini 3.8 Flash	Low	0.5	24,000

C.2 HTML Answer Generation and Rendering

Inputs and presentation decisions. For each query, the framework requests eight web references and supplies the retrieved context together with a shared prompt specifying component interfaces and output requirements. Retrieval failures are retried up to three times, after which generation proceeds without retrieved references. Each model determines the answer content, component selection and configuration, information hierarchy, layout, and color scheme, including grouping, spacing, and visual emphasis. The prompt directs models to use shared typography without font-size overrides, while component renderers provide the underlying runtime implementations.

Structured output and page assembly. The model returns an HTML body through the `html_content` argument of the `WriteReport` tool, containing answer text and component declarations. Component types and JSON configurations are specified through `data-component` and `data-spec`, respectively. The framework parses and validates these declarations, attempts JSON repair when necessary, obtains required media assets, and invokes shared component renderers. It then resolves reference links and assembles a complete page with common styles and runtime dependencies. Invalid optional presentation settings fall back to defaults; unrecognized components or components that fail during component expansion are omitted.

Rendering and screenshot capture. We use Playwright with a Chromium-based browser to render pages at a 390×1000 CSS-pixel mobile viewport with a device scale factor of 2. Navigation has a 60-second timeout. Before capture, the framework waits for image, font, and mathematical rendering, includes a 5.5-second settling period, and checks component readiness. SVG animations are paused at a designated poster time (default: 5.2 seconds) for static screenshots. Full-page screenshots are captured directly or stitched from segments for long pages. Screenshot failures are retried up to twice, and component rendering states are saved for subsequent evaluation.

Animation and interaction recordings. SVG animations are recorded separately from static screenshots, starting from the initial animation state, with recording durations bounded between one and six seconds. The prompt permits at most one animated SVG per page. Local interaction recording supports calculation and random-selection widgets, as well as image-selection and comparison components, using scripted operations such as changing an input, triggering a selection, comparing images, and resetting a widget. Component states are captured at 8 frames per second and encoded into video clips. These recordings provide dynamic evidence rather than exhaustive interaction testing; recording failures are logged without discarding successful screenshots. The generation and rendering implementation is included in the supplementary material.

C.3 Judge Models and Automatic Evaluation Configuration

Each of the 13 leaf-level metrics uses a designated multimodal judge configuration. For a given metric, all evaluated models share the same judge, prompt, scoring criteria, and runtime settings. Model identities are not supplied to the judges.

Table 13: Metric-specific judge models and runtime configurations.

Metric	Judge	Runtime configuration
Text quality	J1	$T = 0.0;\ b = -$
Color harmony	J3	$T = 1.0;\ b = -$
Image quality	J2	$T = 1.0;\ b = 4096$
Chart quality	J1	$T = 1.0;\ b = -$
Animation quality	J2	$T = 0.7;\ b = -$
Interaction quality	J2	$T = 0.7 \rightarrow 1.0;\ b = -$
Layout regularity	J2	$T = 1.0;\ b = -$
Component appropriateness	J1	$T = 1.0;\ b = -$
Chart appropriateness	J2	$T = 1.0;\ b = -$
Image appropriateness	J2	$T = 1.0;\ b = -$
Animation and interaction appropriateness	J1	$T = 1.0;\ b = -$
Content expressivity	J3 + J2	$T = 1.0;\ b = -$ for all stages
First-screen efficiency	J3	$T = 1.0;\ b = -$

J1: Gemini 3 Flash Preview; J2: Gemini 2.5 Pro; J3: Qwen 3.5-397B-A17B. Here, T denotes temperature and b denotes the configured thinking budget. A dash indicates that no reasoning parameter was explicitly configured, not that reasoning was disabled. The temperature arrow for Interaction quality denotes its two evaluation stages.

Evaluation settings and prompts. Evaluation inputs follow Appendix B.2, with a default first-screen screenshot height of 2,000 px. For API failures, timeouts, and structured-output parsing failures, evaluation retries use `max_retry=5` and `sleep_time=7s` with the same target and configuration. Unresolved failures are handled according to §4.2–§4.3. The generation prompt and metric-specific evaluation prompts are included in the supplementary material.

C.4 Validity and Statistical Reporting

A model–query record is included in the overall-score average only if its final evaluation is valid and both Design Normativity and Presentation Effectiveness have valid scores (§4.3). Records that remain invalid after retries are excluded, while valid answers with an overall score of 0 are retained.

Main results use each model's complete valid-answer set from the 786-query evaluation. Valid-answer counts and rates are reported separately; they are not additional terms in the VisExpress score.

Dimension and subdimension scores are first computed for each answer using its valid metrics and their weights, following §4.3, and then averaged over valid answers. The model-level VisExpress score is likewise computed by averaging the answer-level scores $S_i = D_i B_i/100$, rather than by multiplying the separately averaged dimension scores.

For a model m with N_m valid answers, we estimate the 95% confidence interval of its mean VisExpress score using answer-level nonparametric bootstrap. Each iteration resamples N_m answer-level scores S_i with replacement and computes their mean. We use 10,000 iterations, the 2.5th and 97.5th percentiles as interval bounds, and random seed 20260916.

Each resampling unit is one valid model–query record, not an individual component. These intervals characterize uncertainty from the finite query sample, not variation from repeated generation, judge calls, or experimental runs. They are not used for pairwise significance testing.

C.5 Robustness Analysis on the Common Valid Query Set

All 12 models have valid final evaluations on 699 common queries.

Table 14: Results on the 699 common valid queries, ordered by VisExpress score. Bold denotes the highest score in each column.

Model	Vis. ↑	DN ↑	EQ ↑	LR ↑	PE ↑	EC ↑	VE ↑
Claude Fable 5.1	**75.78**	87.91	86.60	91.72	**86.13**	**87.94**	**85.38**
GPT-6 Astra	69.67	**90.39**	**89.06**	93.61	76.62	78.85	76.38
Kimi-k3	68.06	88.51	87.06	92.14	76.76	76.84	77.27
DeepSeek V4.1 Flash	67.87	84.82	81.48	**94.21**	79.93	79.12	80.69
Hy4	66.63	81.02	76.92	92.43	81.97	81.00	82.75
Gemini 3.8 Flash	65.84	83.67	81.55	89.24	78.61	79.74	78.61
DeepSeek V4 Pro	65.70	87.57	85.77	92.06	74.81	73.36	76.23
GLM-5.3 Flash	65.27	85.40	83.23	91.03	76.05	75.30	76.83
GLM-5.3	62.10	85.36	82.76	91.80	72.63	71.55	73.62
Qwen 3.8 Flash	60.80	85.08	82.48	91.61	71.14	66.76	73.99
Gemini 3.1 Pro	60.75	82.08	80.47	86.47	73.84	74.93	73.52
Qwen 3.8-2.4T-A95B	60.46	74.79	67.71	93.67	80.90	81.67	80.87

DN: Design Normativity; EQ: Element Quality; LR: Layout Regularity; PE: Presentation Effectiveness; EC: Expression Coherence; VE: Visual Expressiveness.

We recompute scores on this fixed query set to examine whether differences in model-specific valid-answer sets affect the main results. Table 14 reports overall, dimension, and subdimension scores on the common set; Table 15 compares the cross-model ranges relevant to the observations in §5.2.

Table 15: Cross-model score ranges under the two reporting scopes. Model-specific sets contain all valid answers available for each model; the common set contains the same 699 queries for all models.

Metric or subdimension	Model-specific	Common set
Layout Regularity	86.47–94.17	86.47–94.21
Chart quality	65.19–98.08	65.69–97.87
Animation quality	50.00–76.04	46.97–78.57
Expression Coherence	66.52–88.14	66.76–87.94
First-screen efficiency	86.20–93.17	86.12–93.44
Content expressivity	56.16–77.93	55.84–78.04

Overall mean scores change by at most 0.36 points, and the top nine ranking positions remain unchanged. The bottom three reorder to Qwen 3.8 Flash, Gemini 3.1 Pro, and Qwen 3.8-2.4T-A95B, with a score spread of 0.34 points on the common set.

The patterns underlying §5.2 also persist: Layout Regularity remains less variable than Chart quality, Animation quality, or Expression Coherence, while First-screen efficiency remains higher and less variable than Content expressivity. Thus, the observed contrasts between orderly layouts, component quality, component selection, and whole-answer organization remain apparent when the query set is held fixed.

C.6 Fine-grained Metric Results

Tables 16 and 17 report the 13 leaf-level metrics on each model's own valid-answer set, matching the reporting scope of the main results. Each entry gives the mean answer-level metric score on the 0–100 scale, followed by the number of answers with a valid score for that metric. Counts reflect metric applicability and evaluation validity (§4.2). For component-level metrics, each answer contributes its metric score to the model-level average rather than one observation per component.

The subdimension scores reported in Table 3 and Table 14 are computed separately for each answer using the valid leaf-level metrics and their weights, following §4.3, and are then averaged across answers.

Table 16: Fine-grained results for Design Normativity. Each cell reports score (valid-answer count).

Model	TQ	CH	IQ	CQ	AQ	IntQ	LR
Claude Fable 5.1	93.49 (784)	88.04 (786)	87.78 (224)	81.19 (640)	57.15 (159)	64.74 (268)	91.67 (786)
GPT-6 Astra	96.30 (783)	80.84 (783)	83.90 (59)	98.08 (382)	76.04 (24)	72.46 (69)	93.57 (783)
Kimi-k3	93.44 (785)	83.59 (786)	89.00 (75)	82.10 (411)	61.03 (34)	71.01 (69)	92.01 (786)
DeepSeek V4.1 Flash	85.85 (786)	86.39 (786)	75.79 (53)	72.22 (647)	56.37 (106)	58.89 (315)	94.17 (785)
Hy4	75.54 (785)	83.33 (786)	84.55 (172)	68.81 (713)	65.09 (29)	64.14 (290)	92.47 (786)
Gemini 3.8 Flash	82.52 (785)	86.43 (785)	85.65 (158)	65.19 (498)	57.58 (33)	68.32 (161)	88.93 (784)
DeepSeek V4 Pro	90.65 (786)	83.46 (786)	92.11 (38)	80.68 (486)	56.82 (11)	59.26 (54)	92.14 (786)
GLM-5.3 Flash	87.76 (774)	83.03 (775)	83.45 (72)	79.52 (564)	54.57 (104)	65.70 (86)	90.65 (775)
GLM-5.3	86.62 (779)	80.90 (780)	87.12 (66)	77.66 (432)	50.00 (35)	63.27 (49)	91.64 (778)
Qwen 3.8-2.4T-A95B	45.26 (781)	86.30 (781)	86.34 (97)	73.91 (616)	64.56 (109)	63.86 (101)	93.45 (780)
Qwen 3.8 Flash	82.09 (723)	86.60 (724)	83.71 (133)	71.90 (340)	62.50 (38)	63.48 (115)	91.67 (723)
Gemini 3.1 Pro	81.11 (786)	81.23 (786)	85.72 (185)	74.94 (437)	62.70 (63)	65.43 (81)	86.47 (786)

TQ: Text quality; CH: Color harmony; IQ: Image quality; CQ: Chart quality; AQ: Animation quality; IntQ: Interaction quality; LR: Layout regularity.

Table 17: Fine-grained results for Presentation Effectiveness. Each cell reports score (valid-answer count).

Model	CompA	ImgA	ChartA	AIA	FSE	CE
Claude Fable 5.1	96.58 (786)	78.04 (230)	77.11 (688)	85.98 (437)	91.75 (786)	77.93 (785)
GPT-6 Astra	87.10 (783)	48.40 (94)	69.12 (510)	56.62 (325)	93.17 (783)	56.36 (782)
Kimi-k3	87.29 (786)	55.66 (106)	64.27 (557)	58.38 (325)	89.57 (786)	63.26 (786)
DeepSeek V4.1 Flash	91.64 (786)	35.78 (109)	68.49 (722)	69.86 (506)	88.76 (786)	71.78 (785)
Hy4	93.02 (786)	69.68 (188)	72.93 (750)	66.63 (481)	87.96 (786)	76.75 (785)
Gemini 3.8 Flash	90.52 (784)	67.88 (179)	68.63 (644)	60.65 (385)	86.20 (785)	69.87 (785)
DeepSeek V4 Pro	84.59 (786)	30.50 (100)	62.52 (599)	50.74 (340)	86.64 (786)	63.26 (786)
GLM-5.3 Flash	84.24 (773)	56.02 (108)	65.41 (639)	67.06 (378)	89.45 (775)	61.45 (775)
GLM-5.3	81.46 (779)	51.92 (104)	59.64 (560)	54.32 (336)	87.72 (780)	56.16 (779)
Qwen 3.8-2.4T-A95B	91.98 (781)	62.77 (137)	71.06 (691)	70.94 (363)	87.95 (781)	72.98 (780)
Qwen 3.8 Flash	78.14 (723)	62.20 (168)	50.63 (556)	48.60 (393)	87.73 (724)	56.84 (724)
Gemini 3.1 Pro	82.53 (786)	69.70 (203)	64.31 (559)	61.05 (362)	86.84 (786)	56.93 (786)

CompA: Component appropriateness; ImgA: Image appropriateness; ChartA: Chart appropriateness; AIA: Animation and interaction appropriateness; FSE: First-screen efficiency; CE: Content expressivity.

Component-quality scores are conditional on the components each model generates. Their valid counts therefore need not match across models or match those of the corresponding appropriateness metrics, which also assess missing necessary components.

D HUMAN EVALUATION AND AGREEMENT ANALYSIS

This appendix details the four-model human ranking study in Section 5.3, its agreement measures, and the comparison of alternative score aggregation rules.

D.1 HUMAN RANKING EXPERIMENT

The human evaluation covers four models: Claude Fable 5.1, Kimi-k3, DeepSeek V4 Pro, and GLM-5.3. The study uses all 786 queries and involves 60 annotators. Each query is independently evaluated by five distinct annotators.

For each query, annotators compare the four complete HTML answers and provide relative rankings, with ties permitted when two or more answers are indistinguishable under the annotation criteria. The ranking criteria follow the core VisExpress evaluation requirements. Annotators first identify the user's information need, then compare the answers in terms of readability, information accessibility, information hierarchy and organization, and whether visual components help users understand or use the presented information. They inspect both the first-screen presentation and the complete page; for answers containing animations, videos, or interactive components, they complete the corresponding playback or interaction before ranking. The number of components, page complexity, or visual decoration alone does not provide an advantage.

During annotation, model identities are hidden, the presentation order of answers is randomized, and annotators do not receive VisExpress scores or judgments from other annotators.

Human consensus. For each query, we average each model's assigned ranks across the five annotators and order the models by ascending mean rank. Tied answers receive the average of their occupied rank positions, and equal mean ranks remain tied. This consensus is used for Top-1 Agreement and model-level aggregation; pairwise agreement uses individual annotator judgments directly.

D.2 AGREEMENT MEASURES AND STATISTICAL PROCEDURES

Automatic–human agreement uses a matched set of **770** queries with valid VisExpress scores for all four models and matching query texts and answers across both evaluations. Human–human pairwise agreement uses the same set. Scores are rounded to ten decimal places before comparison to prevent numerical noise from breaking ties; equal scores are treated as tied.

Pairwise agreement. We follow the S1/S2 agreement definitions of MT-Bench (Zheng et al., 2023). Each annotator's four-model ranking yields six pairwise preferences, labeled as preferring one answer, preferring the other, or a tie. VisExpress induces the same three labels from its scores. Comparing each automatic label with all five individual judgments gives 23,100 comparisons. The human reference compares all ten pairs of distinct annotators on each answer pair, giving 46,200 comparisons.

S1 includes all comparisons and counts agreement only when both labels match, including a shared tie. S2 excludes a comparison if either judgment is a tie. Both are computed as the total number of matching judgments divided by the total number of eligible comparisons. S2 retains 22,399 automatic–human comparisons (96.97%) and 43,978 human–human comparisons (95.19%), both covering every query in the matched set. Uniform random choices among the three labels yield a one-third reference for S1; uniform binary choices yield 50% for S2.

Top-1 Agreement. This measure assesses whether VisExpress selects a human-preferred answer. Let H_q^{top} contain the models with the lowest human mean rank and V_q^{top} those with the highest VisExpress score. The query-level score is:

$$\text{hit}_q = \frac{\left|V_q^{\text{top}} \cap H_q^{\text{top}}\right|}{\left|V_q^{\text{top}}\right|}. \tag{4}$$

We average this score over the matched query set, retaining all tied top models. The random reference is the mean of $|H_q^{\text{top}}|/4$, or 26.69%. A separate human reference compares each annotator's

top-ranked set with that of the other four annotators' mean-rank consensus, averaging the five comparisons within each query and then across queries.

Overall annotator concordance. We compute tie-corrected Kendall's W from the five annotators' rankings for each query and report the mean over all 786 queries. As a chance reference, independent uniform random rankings yield an expected W of $1/5 = 0.2000$ for five annotators.

Confidence intervals. We use 10,000 query-level bootstrap samples (seed 20260916), taking the 2.5th and 97.5th percentiles as interval bounds. Each bootstrap sample draws queries with replacement from the corresponding analysis set, keeping its size unchanged and retaining all judgments for each sampled query. Pairwise agreement is recomputed as a pooled count ratio, while Top-1 and W are recomputed as query means. Queries are sorted by ID for the automatic–human comparisons. These intervals describe query-sampling uncertainty conditional on the collected annotations.

D.3 Human Agreement Results

Table 18: Agreement results for the four-model human study. Pairwise comparisons use individual judgments; Top-1 uses human consensus.

Comparison	Statistic	N	Result	95% CI
H–H	Mean Kendall's W	786	0.4596	[0.4416, 0.4775]
V–H	Pairwise S1	770	56.95%	[55.80%, 58.10%]
H–H	Pairwise S1	770	62.15%	[61.14%, 63.16%]
V–H	Pairwise S2	770	58.68%	[57.50%, 59.86%]
H–H	Pairwise S2	770	65.05%	[64.02%, 66.07%]
V–H	Top-1 Agreement	770	42.99%	[39.48%, 46.36%]

Abbreviations: H–H, human–human; V–H, VisExpress–human; N, number of queries.

VisExpress–human S1 agreement is 5.20 percentage points below human–human agreement. This human reference contextualizes agreement with individual judgments under the same counting rules. For Top-1, the leave-one-annotator-out reference is 54.83% [95% CI: 52.52%–57.17%].

The full-study Kendall's W indicates shared preferences among annotators. On the matched set, VisExpress exceeds the S2 random reference and the Top-1 random-selection reference.

Table 19: Model-level aggregated results on the four-model, 770-query matched set.

Model	Mean human rank ↓	Human rank	VisExpress score ↑	VisExpress rank
Claude Fable 5.1	1.9495	1	75.82	1
Kimi-k3	2.5712	2	68.25	2
DeepSeek V4 Pro	2.6012	3	65.38	3
GLM-5.3	2.8782	4	61.87	4

At the model level, the human and VisExpress orderings are identical (Table 19): human mean ranks and VisExpress scores are aggregated independently over the matched set before comparing the resulting rankings.

D.4 Score Aggregation and Human Agreement

We compare the full score $D_i B_i/100$ with Design Normativity alone (D_i), Presentation Effectiveness alone (B_i), and their arithmetic mean ($(D_i + B_i)/2$). All variants reuse the same four-model answers, leaf-level scores, applicability rules, weights, and individual human rankings. Score ties and agreement follow Appendix D.2. S1 retains all 23,100 comparisons; S2 eligibility varies with each scoring rule's ties.

Table 20: Pairwise human agreement under alternative scoring rules. Brackets indicate 95% confidence intervals; S2 counts are eligible individual-judgment comparisons.

Scoring rule	S1	S2	S2 count
VisExpress: $D_i B_i/100$	56.95% [55.80%, 58.10%]	58.68% [57.50%, 59.86%]	22,399
Design Normativity: D_i	45.62% [44.35%, 46.89%]	52.04% [50.74%, 53.33%]	20,120
Presentation Effectiveness: B_i	56.53% [55.29%, 57.74%]	60.11% [58.89%, 61.33%]	21,664
Arithmetic mean: $(D_i + B_i)/2$	56.66% [55.49%, 57.82%]	58.41% [57.21%, 59.60%]	22,389

We estimate 95% confidence intervals for each scoring rule and its difference from the full score using query-level bootstrap (Appendix D.2). All rules use the same resampled queries. Intervals are computed separately for each comparison, without multiple-comparison correction.

The full score improves S1 agreement over Design Normativity alone by 11.33 percentage points [95% CI: 10.00, 12.67]. Its S1 difference from Presentation Effectiveness alone is small and unresolved (0.42 points [−0.66, 1.49]), and its improvement over the arithmetic mean is 0.29 points [0.04, 0.54]. When S2 is restricted to the 19,375 comparisons where all four scoring rules and the human judgment are non-tied, the full score achieves 58.87%, versus 60.18% for Presentation Effectiveness alone (difference −1.31 points [−2.52, −0.11]). These results support the contribution of Presentation Effectiveness beyond basic presentation quality. The full score achieves the highest S1 agreement, while Presentation Effectiveness alone achieves the highest S2 agreement.

E Chart-Appropriateness Rubric and Worked Example

This appendix summarizes the core rubric for **Chart appropriateness** and presents a worked example from evaluation inputs to the metric score. The complete prompt is provided in the supplementary material.

Core Evaluation Rubric

Evaluation Scope and Inputs

- **Goal.** Determine whether the answer uses charts when they are needed and whether each existing chart is relevant, correctly placed, nonredundant, and useful for communicating the information required by the query.
- **Inputs.** The judge receives the user query, the rendered full-page screenshot, extracted answer text containing chart placeholders, and extracted SVG infographic text when applicable. The screenshot takes precedence for chart interpretation and unmet visualization needs; extracted text serves only to locate recognized charts and clarify their intended roles.
- **Recognized charts.** The metric considers ECharts charts, SVG infographics, candlestick charts, and line or bar charts rendered directly from HTML. Ordinary cards, maps, SVG animations, navigation icons, decorative illustrations, and interactive components are not treated as charts merely because they contain chart-like marks.
- **Metric boundary.** Contrast, label legibility, overlap, clipping, incorrect proportions, and missing chart elements are assessed by Chart quality.

Missing-Chart Assessment

Check for missing visual support whether or not the answer already contains a chart.

1. **Explicit requirement or commitment.** Penalize omissions when a chart is explicitly requested but no related chart is provided, or when a promised chart is absent from the rendered page.

2. **Central information needing visualization.** For central information, expect a chart when it materially reduces comprehension cost or exposes relationships not communicated equally well by prose, lists, tables, or cards. Typical cases include complex spatial, logical, dependency, or multistage relationships; numerical comparisons with enough comparable observations to reveal meaningful differences; numerical change over time, including trends, cycles, and anomalies; composition or distribution across multiple parts; and professional financial trend analysis.
3. **Exemptions.** Do not require a chart for incidental values, sparse or heterogeneous data, a single percentage, simple linear processes, or information already communicated efficiently by prose, a list, a table, or an existing chart. A visual form is justified by information gain, not by the mere presence of numbers or structure.
4. **Severity.** Central information dispersed or buried in continuous prose is a stronger omission than central information already exposed in a conspicuous list, card, or table.

EXISTING-CHART ASSESSMENT

Evaluate every chart visible in the rendered answer.

- **Relevance and placement.** Charts must support the user intent or a substantive claim and relate clearly to adjacent text. Place global summaries at the answer level and section-specific charts within the corresponding section.
- **Misuse.** Penalize charts that are unrelated, broken, meaningless, or intrinsically unsuitable for the information they encode. Do not treat a chart-quality defect as misuse unless the rendered chart is effectively unusable.
- **Redundancy and overdesign.** Check for identical or functionally duplicated charts; full duplication between chart data and surrounding text, cards, or tables; unnecessary splitting; unnarrated accumulation; and decorative or repeated encodings that increase comprehension cost without adding information.
- **Information gain.** A chart must satisfy at least one positive-gain criterion: replacing or complementing text, revealing patterns, condensing dispersed information, or clarifying complex relations. Insufficient gain includes summary charts with insufficient coverage, overviews placed after the details they should introduce, and clearly nonoptimal chart types. Also assess whether an SVG infographic would better convey a simple, low-interaction ECharts chart's content.
- **Scientific and dynamic processes.** A static diagram must communicate the mechanism or sequence through spatial organization; an important temporal or causal process that remains unclear may indicate insufficient gain.

SCORING CRITERIA

Apply the lowest score whose condition is satisfied.

Score 0 **(inapplicable):** the answer contains no chart, and no chart is needed. This value is excluded from score aggregation.

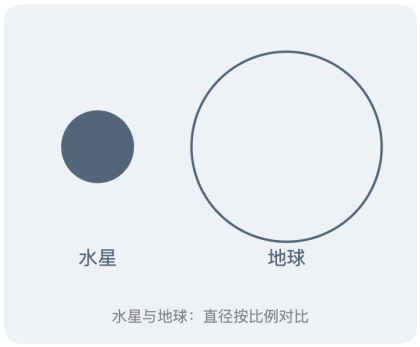
Score 1 a critical omission or misuse is present, including an unmet explicit requirement, a missing committed chart, central information buried in prose, irrelevant or misplaced chart content, a meaningless chart, or structural/functional chart repetition.

Score 3 no score-1 condition applies, but the answer has a partial appropriateness problem: central information remains only in a structured textual form; chart and text/card/table content are redundant; the presentation is overdesigned; or an existing chart provides insufficient gain or uses a clearly nonoptimal representation.

Score 5 no score-1 or score-3 condition applies; the charts are relevant, correctly placed, nonredundant, and provide clear information or efficiency gain.

Explanations should justify the score with specific content from the answer.

Worked example. Figure 4 illustrates how GPT-6 Astra's answer to "How big is Mercury?" receives a chart-appropriateness score of 5. An ECharts chart from GLM-5.3 Flash provides a reference for the rubric's insufficient-gain criterion.



(a) Full-page screenshot

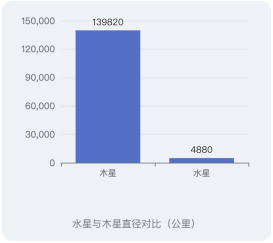
Model: GPT-6 Astra **Score:** 5

Component outline. Using Earth's mean diameter of approximately 12,742 km as a reference, Mercury's diameter is approximately 4,880 km, about 38% of Earth's. The circle diameters are drawn to scale; they do not represent the actual distance between the planets.

Why score 5. The SVG infographic turns the textual statement "about 38% of Earth's diameter" into a visible size comparison. The assessment credits this concrete representation of an abstract ratio, together with the chart's relevance, appropriate placement, and simple design. The visualization makes the relationship more intuitive than the numerical ratio alone (positive gain).

Reference chart: GLM-5.3 Flash (score 3).
Insufficient gain—nonoptimal visualization.

The chart uses substantial visual space to compare only two diameter values, for Mercury and Jupiter. The assessment considers this excessive for the information conveyed.



(b) ECharts chart

Figure 4: Worked chart-appropriateness example for "How big is Mercury?" (a) Full-page screenshot from GPT-6 Astra. (b) Reference ECharts chart from GLM-5.3 Flash.

Answer Text

The actual text input supplied to the judge is translated below.

```
# How big is Mercury?
Mercury has a diameter of about 4,880 km and is the smallest of the eight
planets in the Solar System.
Compared with Earth, Mercury's diameter is about 38% of Earth's; in other
words, approximately 2.6 Mercurys would fit across Earth's diameter.
![(SVG infographic)Mercury and Earth: diameters compared to scale:Using
Earth's mean diameter of approximately 12,742 km as a reference, Mercury's
diameter is approximately 4,880 km, about 38% of Earth's. The circle
diameters are drawn to scale; they do not represent the actual distance
between the planets.]
Treating the planets as approximate spheres, Mercury's volume is about 5.6%
of Earth's, or approximately 1/18. This compares volume, not mass.

@<Closing Component>
In brief: Mercury's diameter is about 4,900 km, under 40% of Earth's.
@</Closing Component>
```