CROSSREF-VQA: BENCHMARKING CROSS-MODAL REFERENCE RESOLUTION IN STRUCTURED GRAPHICS

Anonymous authors
Paper under double-blind review

ABSTRACT

While MLLMs perform well on everyday visual question answering (VQA) tasks, they remain unreliable when required to combine visual and textual information at the level of individual elements. Interpreting structured graphics demands exactly this: elements are marked with short references such as legend keys and colour codes with the corresponding semantic identities in a separate text section, leaving the reader to cross-reference. To study this limitation, we propose **CrossRef-VQA**, a benchmark explicitly designed to isolate element-precise cross-modal reference resolution. **CrossRef-VQA** renders structured data as graphics, replaces every semantic label in the image with an anonymous reference, supplies the reference-to-element mapping as text, and computes every answer programmatically from the underlying data. We intentionally incorporate human designed templates to avoid circular LLM evaluation and support scalable data generation, enabling stable large-scale assessment. Unlike traditional VQA benchmarks that evaluate a single, confounded setup – where full semantic information is redundantly present in both the image and the text – our framework introduces a set of controlled experimental conditions. By keeping the anonymised graphic fixed and systematically varying only the need to perform cross-modality *referencing* and *de-referencing*, we can directly compare performance across conditions to cleanly isolate cross-modal integration from pure visual perception. Across 8 chart types and 11 MLLMs, we find that requiring cross-referencing causes a significant overall performance drop of 15–30% among the eight models used in the primary analysis. Notably, the two directions fail differently: referencing is imprecise when the question specifies an element, whereas de-referencing is skipped, where models correctly locate the visual element but return its raw label most of the time. These findings demonstrate that the fundamental failure point is not visual understanding, but the cross-modal reference resolution itself.

1 INTRODUCTION

Linking complementary visual and textual information is a fundamental requirement for Multimodal Large Language Models (MLLMs) (Baltrušaitis et al., 2018; Yin et al., 2024). In CAD drawings, patent diagrams, and architectural blueprints, short visual identifiers refer to descriptions in separate parts lists or schedules. Answering questions about these graphics requires **referencing** (semantics $\rightarrow$ reference), **de-referencing** (reference $\rightarrow$ semantics), or both. Humans incur a split-attention cost during this lookup (Chandler & Sweller, 1991; Ginns, 2006); whether MLLMs reliably perform the same resolution remains underexamined.

High accuracy on ChartQA and MMMU does not establish this ability (Masry et al., 2022; Yue et al., 2024). MMStar and MMMU-Pro highlight textual shortcuts and inline-image reading as alternative routes to successful answers (Chen et al., 2024; Yue et al., 2025). Existing benchmarks emphasize overlapping information, either redundant across modalities or integrated in an image, or disjoint but conflicting inputs (Methani et al., 2020; Zhang et al., 2024; Deng et al., 2025; Yan et al., 2025). The disjoint-complementary setting instead requires connecting separate, non-conflicting sources. Prior work typically prints semantic annotations on the figure (Wang et al., 2024b) or distributes separate facts across modalities (Wang et al., 2026), leaving the element-level lookup unisolated.

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.

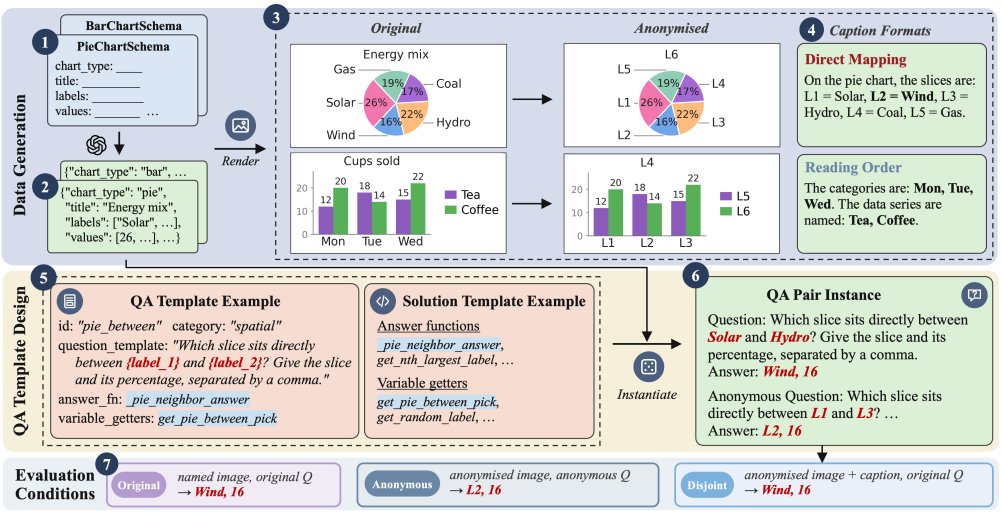


Figure 1: **CrossRef-VQA pipeline.** (1, 2) GPT-4o populates chart schemas with semantically coherent data. (3) Charts are rendered in Original (semantic labels) and Anonymised (dummy labels) formats. (4, 5) Text annotations map dummy labels to semantics, paired with programmatic QA templates. (6, 7) These generate instances evaluated across three information-equivalent conditions.

To address this gap, we propose **CrossRef-VQA** (Figure 1), a benchmark generator covering eight chart types and 92 human-authored template designs with deterministic solution functions. An LLM supplies generated data from schema; programs render the graphics, separate semantic labels into a text key, and compute answers. Our evaluated set contains 154 chart instances and 1,305 QA pairs spanning five task categories. We evaluate 11 MLLMs under Original (O), Anonymous (A), and Disjoint (D) conditions. The latter two share an anonymised image but differ in whether questions and answers use references directly or require a separate semantic key.

By comparing performance across these carefully controlled conditions, **CrossRef-VQA** disentangles basic visual perception from the higher-order cognitive task of multimodal fusion. Our evaluation of 11 state-of-the-art MLLMs reveals a severe and persistent mapping gap when modalities are decoupled. While baseline visual accuracy on the Anonymous condition (A) remains high for many chart types, enforcing cross-modal reference in the Disjoint condition (D) triggers a significant degradation in performance. This evaluation exposes two distinct architectural bottlenecks: an alignment bottleneck where models process visual geometries but fail at color-legend binding (frequently outputting the visual dummy label instead of the semantic answer), and a low-level perception bottleneck on dense geometries like error bars. Furthermore, through a controlled scaling ablation, we demonstrate that this fusion deficit is an inherent limitation of current vision-language binding mechanisms, rather than a capacity issue resolvable by simply scaling the language backbone.

2 Related Work

Existing benchmarks for multimodal reasoning predominantly consider relationships between information sources that are Overlapping (Redundant or Integrated) and Disjoint (Conflicting), with the Disjoint-Complementary setting remaining largely underexplored.

The Overlapping Regimes. Most established multimodal benchmarks exhibit significant information overlap between modalities. From early works such as PlotQA (Methani et al., 2020) and ChartQA (Masry et al., 2022) to more recent ones including MMMU (Yue et al., 2024), MathVision (Wang et al., 2024a) and ChartQA-Pro (Masry et al., 2025), textual queries usually contain overlapping cues presented in visual evidence, making it possible to directly infer the answer without understanding the image input. To counter textual shortcuts, MathVerse (Zhang et al., 2024) and MMMU-pro (Yue et al., 2025) render the query directly in the image, forcing the models to read from pixels since no textual information is provided. While these benchmarks necessitate visual perception, the

information remains spatially integrated, making the task solvable as a uni-modal vision problem rather than cross-modal synthesis.

The Complementary Regimes. Another line of research presents contradictory visual and textual inputs to quantify modality preference and interference, revealing a persistent bias toward text or parametric priors Liu et al. (2024); Deng et al. (2025). In contrast to these interference studies, we examine cross-modal cooperation when visual and textual channels provide complementary and disjoint information. MultiModalQA (Talmor et al., 2021) was one of the pioneering works introducing document-level reasoning with text and images, followed by MMCR (Tian et al., 2025) and WikiMixQA Foroutan et al. (2025) which benchmark cross-source reasoning in scientific papers or long documents. However, these benchmarks test multi-step reasoning across separate sources without a paired same-image control, and most evaluate models via multiple choice. Closest to our work, Compose and Fuse Wang et al. (2026) varies how logical facts are distributed and composed across modalities and locates the bottleneck in cross-modal integration rather than perception. While they split distinct propositions across modalities but leave each visual input semantically self-contained, we decouple a single graphic into visual primitives and a separate semantic context, where the image contains disparate information from the caption, and score visual-textual resolution from open-ended model answers.

Keyed graphics and symbol binding. Prior work on *visual binding* either matches plotted marks to legends that remain inside the image Xu et al. (2023); Tang et al. (2026), grounds alphanumeric tags on objects where the tag itself is the target answer Yan et al. (2024), or evaluates keyed graphics end-to-end without isolating the lookup step Wang et al. (2024b). Building on this literature, CrossRef-VQA moves the semantic key into external text and uses a paired same-image control to separate visual perception from both referencing and de-referencing.

3 METHODOLOGY

3.1 DESIGN PHILOSOPHY

Our construction separates visual primitives from semantic metadata, uses algorithmic ground truth, and supports schema-driven generation.

Structural Disentanglement. Many existing benchmarks allow models to exploit overlapping text within an image, bypassing true multimodal integration. To counter this, we treat visual primitives (geometry, color, spatial layout) and textual metadata (legends, labels) as strictly separate information streams. By decoupling these channels, we construct evaluation conditions where neither modality alone is sufficient, compelling the model to actively synthesize both sources to derive an answer.

Algorithmic Ground Truth. Relying on LLMs to both generate questions and evaluate answers risks circular logic, where generation and evaluation errors unpredictably compound (Yang et al., 2025). We eliminate this by enforcing a strict separation of concerns: humans design the reasoning templates to guarantee cross-modal dependence, while deterministic programs compute exact answers directly from the underlying structured data. This ensures that all ground truths are externally verifiable and completely independent of model behavior.

Scalability. To enable large-scale evaluation while preventing test-set memorization, we adopt a generative, schema-driven approach. Probabilistic schemas specify chart type, series structure, layout, and semantic topic, allowing us to produce virtually unlimited chart instances while preserving the structural constraints necessary for programmatic verification. For this evaluation, we curated a vetted set spanning eight chart topologies, 154 chart instances, and 1,305 QA pairs.

3.2 EVALUATION CONDITIONS

To isolate reasoning, perception, and cross-modal alignment capabilities, we evaluate models under five controlled input conditions. Fix pre-generated structured data, let the chart image rendered be V_O, its twin with anonymised labels be V_A (with dummy labels L_n's), and V_A's textual complement be T (with mappings such as "Category L1 is Tech Sector"), our benchmark draws conclusions by evaluating the following conditions:

- **Original Condition (O).** MLLMs receive the image rendered from the generated data directly, preserving all labels and semantic meanings on the image. This condition reduces the benchmark into a standard VQA task, which serves as a measure to interpret LLM's capability.
- **Anonymous Condition (A).** MLLMs receive the image V_A, and questions are phrased using anonymised labels (e.g., "Which series among L8, L9, L10 has the lowest total sum?"). Since labels are seen in images, the model can answer with those labels.
- **Disjoint Condition (D).** MLLMs receive the image V_A and its description T as input. Questions are phrased using semantic labels (e.g., "Which series among Q1 2023, Q2 2023, Q3 2023 has the lowest total sum?"). To answer, the model must consult T to align "Q1 2023" with "L8 (blue)" in the image, then perform visual reasoning.
- **Unimodal Ablations (V only; T only).** The model receives a single modality. For V_A-only, questions use semantic labels but the model lacks T to resolve the mapping. For T-only, the model receives the text key with a noise image. These serve as validity checks: high performance would indicate information leakage or dataset bias, while near-chance performance confirms that the benchmark requires cross-modal integration.

Crucially, O, A, and D are *information-equivalent* and *reasoning-equivalent*—they all contain sufficient information and require identical logical operations to compute the answer, differing only in their cross-modal binding requirements. Specifically, the performance gap between A and D isolates errors in cross-modal binding and de-referencing. Because both conditions demand identical visual parsing steps to read the chart, any degradation in D directly reflects the model's inability to integrate the textual mapping.

3.3 Data Generation and Modality Decoupling

We design an automated data generation pipeline that balances semantic realism with strict structural control. Following prior visualization benchmarks (Yang et al., 2025), we use GPT-4o to generate chart data conditioned on high-level topics[1]. To avoid inconsistencies inherent in free-form generation, we enforce a predefined JSON schema specifying all attributes required for deterministic rendering. The pipeline consists of three stages:

1. **Data Synthesis.** GPT-4o populates the schema with numeric values, distributions, and textual metadata (titles, axis labels, tick labels, legends).
2. **Chart Rendering (Generating V_O).** Charts are rendered using Matplotlib with specified metadata (groups, colours, titles, axis, etc) under fixed layout templates to ensure visual diversity while preserving structural consistency.
3. **Modality Decoupling (Generating T and V_A).** For each instance V_O, we construct a tuple containing its anonymised twin, V_A, and a pure-text semantics description, T, whose information may be used to perform referencing and/or de-referencing. As shown in Figure 1, pie and treemap charts use explicit reference-to-identity mappings, while bar, line, box, violin, error-bar, and error-point charts list category or series names in reading order.

Formally, let $I(\cdot)$ denote the information content of an input *with respect to its query*. Our construction satisfies $I(V_A) \cup I(T) = I(V_O), \quad I(V_A) \cap I(T) \approx \emptyset$. To minimize information leakage, we randomize axis scaling, assign non-semantic colors, and remove implicit numeric encoding. This ensures that the questions require explicit cross-modal referencing rather than unimodal inference.

3.4 Question–Answer Template Design

Figure 1 illustrates our paired template system; representative question examples appear in Table 5a and Table 5b in Appendix A. Each QA instance is generated from a *Question Template* (a natural language query instantiated with schema variables) and a paired *Solution Template* (an executable program that computes the exact answer from the underlying structured data). As detailed in Appendix A, **`CrossRef-VQA`** spans five reasoning categories: **Value Lookup** (retrieving specific data points), **Aggregation** (computing sums, averages, or variances), **Comparative** (ranking magnitudes

[1]Only text is generated to fill in the schema, thus using other performative LLMs such as Claude is equivalent; we generically select GPT-4o by following prior works and balancing generation speed, cost and quality.

Table 1: Required-operation distribution within each question category.

Question category	Referencing only	Dereferencing only	Both	Templates
Extrema	5.3%	47.4%	47.4%	38
Value lookup	100.0%	0.0%	0.0%	18
Comparative	77.8%	16.7%	5.6%	18
Aggregate	72.7%	27.3%	0.0%	11
Spatial	0.0%	0.0%	100.0%	7
Overall	45.7%	26.1%	28.3%	92

between entities), **Extrema** (identifying maximums, minimums, or second-order extrema), and **Spatial** (identifying positional relationships).

Beyond cognitive complexity, we explicitly categorize templates by the required operations of their cross-modal mapping, as detailed in Table 5b. This orthogonal dimension tracks the flow of information between modalities:

- **Referencing (Semantics → Reference):** The model is given a semantic entity, must ground it visually, and extract a computed value.
- **De-referencing (Reference → Semantics):** The model is given visual or geometric criteria (e.g., the largest visual element), must locate it, and return its semantic identity.
- **Bidirectional (Both):** The model must locate semantic anchors visually, execute a spatial or comparative operation to find a target, and map that target back to its semantic identity.

As **Table 1** illustrates, the distribution of required mapping operations is strictly dependent on the question's reasoning category. For instance, **Value lookup** tasks strictly demand referencing (100.0%), as they query the visual properties of a given text entity. Conversely, **Spatial** queries exclusively force bidirectional mapping (100.0%) to resolve positional relationships between entities. Categories like **Extrema** naturally skew towards dereferencing or bidirectional operations, requiring the model to visually search for a maximum or minimum before translating that target back into the semantic text space.

The Anonymous condition preserves the underlying visual reasoning task while skipping text-vision mapping: questions and answers use the chart's dummy labels directly.

4 Evaluation Design

4.1 Benchmark Details and MLLMs Evaluated

We evaluate 11 MLLMs as follows: GPT-6 Sol and GPT-6 Luna (OpenAI, 2026), Claude Sonnet 5 (Anthropic, 2026), Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025), DeepSeek V4.1 Flash (DeepSeek-AI, 2026), GLM-4.6V (GLM-V Team, 2025), Qwen3.8-27B (Qwen Team, 2026), Llama 4 Maverick (AI, 2025), GPT-4o (OpenAI, 2024), Grok 4.3 (xAI, 2026), and Gemma 3 27B (Gemma Team, 2025). During preliminary analysis, we identified three models—GPT-4o, GPT-6 Luna, and Gemma 3 27B—as "capability-capped readers"; they select the queried element on the anonymized chart less than half the time (Condition A accuracy ranging from 25.7% to 47.5%). We leave these three models out of our pooled results to avoid skewing the variance, focusing our primary analysis on the remaining eight capable models.

4.2 Evaluation Metrics

Scoring Protocol. To ensure consistent parsing, we instruct all models to enclose their final response within the {{answer}} tags. We then compare the answers from the MLLMs with the ground truth, which we retrieve programmatically from the underlying data D. Following Yang et al. (2025), we evaluate numerical answers with a 5% tolerance, while textual responses (e.g., categories) are assessed via regular expression matching to accommodate minor formatting variations. If a model fails to produce a valid tagged answer, the response is marked as incorrect. All evaluation scripts are deterministic and will be released with the benchmark for reproducibility.

Table 2: Main results. O−A measures the effect of anonymisation on reading (positive: the original image is read better). A−D is the mapping gap, over all items and by required operation: referencing only, de-referencing only, and both. Models are ordered by D. †Capability-capped readers (A < 50): their own reading caps the gap that can be observed.

	Accuracy (%)			Anonymisation	Mapping gap, A−D (points)			
Model	O	A	D	O−A	**All items**	Referencing only	Dereferencing only	Both
GPT-6 Sol	78.8	76.7	62.0	−0.7	**14.7**	6.4	31.8	6.8
Claude Sonnet 5	72.9	71.0	47.2	+0.6	**23.8**	5.0	42.2	31.3
Qwen3-VL-235B	69.1	64.9	45.8	+3.5	**19.1**	9.9	30.3	19.8
DeepSeek V4.1 Flash	61.8	62.8	39.9	+0.5	**22.9**	10.3	33.2	30.4
GLM-4.6V	68.0	65.4	39.5	+1.0	**25.9**	4.0	39.8	43.7
Qwen3.8-27B	65.8	61.7	34.1	+4.1	**27.6**	12.7	38.4	38.1
Llama 4 Maverick	56.5	56.6	33.9	−1.4	**22.7**	7.2	37.2	29.5
GPT-4o†	50.9	47.5	26.8	+1.1	**20.7**	9.2	38.2	17.4
Grok 4.3	56.2	56.3	26.3	+0.9	**30.0**	9.0	44.8	45.4
GPT-6 Luna†	31.0	25.7	24.5	+4.0	**1.2**	−5.3	6.6	5.0
Gemma 3 27B†	33.0	34.0	17.2	−2.6	**16.8**	2.9	27.0	26.3

Mapping Gap (Δ_{map}). We quantify a model's cross-modal fusion capability using the **Mapping Gap**, defined as the absolute performance difference between the visually equivalent, anonymized, and disjoint conditions: $\Delta_{\text{map}} = \text{Acc}(A) - \text{Acc}(D)$.

The Mapping Gap categorizes model behavior into three regimes:

- **Mapping Deficit ($\Delta_{\text{map}} > 0$):** The model's performance degrades when required to map decoupled text to visual elements.
- **Decoupling Gain ($\Delta_{\text{map}} < 0$):** The model performs better under disjoint conditions. Given our strict enforcement of high-quality rendering, this typically reflects an instability in the model's visual reasoning over limited samples rather than any genuine advantage to decoupling.
- **Mapping Robustness ($\Delta_{\text{map}} \approx 0$):** The model perfectly bridges the modalities.

Ultimately, Δ_{map} serves as a direct diagnostic measure of multimodal integration efficiency by strictly isolating mapping-specific degradation. Because the Anonymous condition inherently controls for task difficulty and visual reasoning capacity, the absolute gap provides a fair and interpretable comparison across models and question types without requiring further normalization.

Benchmark Robustness. We report Acc(V_A-only) and Acc(T-only) as accuracies under the visual-only and text-only ablation conditions, respectively. These ablations serve as a dataset validity diagnostic. Because the question templates are explicitly constructed to require complementary information from both modalities, high performance under either ablation condition would indicate information leakage or exploitable unimodal shortcuts. Conversely, near-chance performance under these conditions provides empirical evidence that the benchmark instances genuinely require cross-modal synthesis, confirming the robustness of our data generation design.

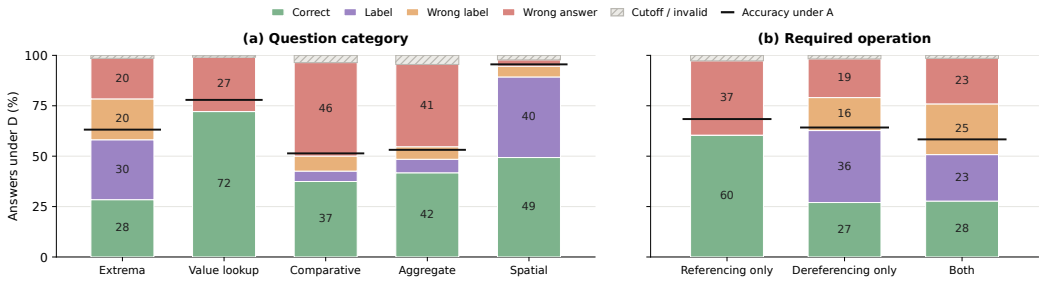
5 Results and Discussion

5.1 Headline Results: The Pervasive Mapping Gap

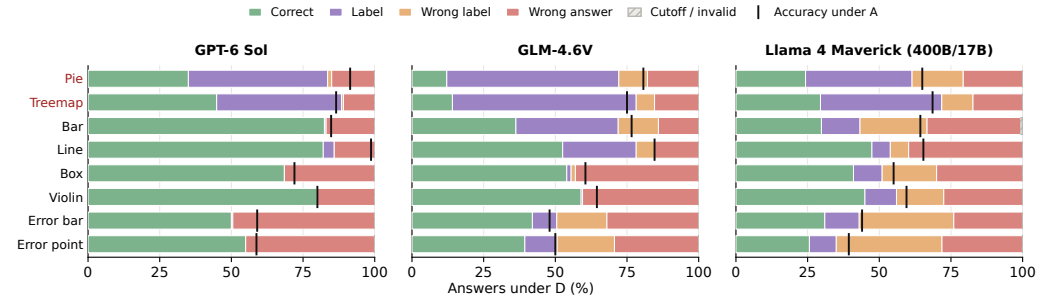
Table 2 reports the evaluation of **`CrossRef-VQA`** across all 11 models. Our primary mapping analysis focuses on the eight capable readers (Section 4). The visual disruption of anonymisation is negligible: the performance difference ($O - A$) remains within 1.4% for six capable models, confirming that replacing semantic labels with anonymous references does not degrade baseline perception.

In contrast, enforcing cross-modal reference in the Disjoint condition triggers severe and universal degradation. Every capable model exhibits a substantial mapping gap ($A - D$), ranging from 14.7% (GPT-6 Sol) to 30.0% (Grok 4.3). This pervasive gap confirms that current architectures fundamentally struggle with multimodal synthesis, even when baseline visual recognition succeeds.

Crucially, mapping robustness does not strictly correlate with parameter count. Grok 4.3 and Llama 4 Maverick exhibit nearly identical Anonymous baselines (56.3% and 56.6%), yet Grok suffers



(a) Error decomposition by reasoning category (left) and mapping direction (right). When operations require dereferencing, failures are heavily dominated by skipped mappings (Label) and incorrect visual selections (Wrong label).



(b) Disjoint outcomes across chart topologies for three representative models. Alignment bottlenecks (high Label errors) dominate direct-mapping charts (e.g., pie, treemap), while dense geometries (e.g., error bars) expose low-level perception bottlenecks.

Figure 2: Performance breakdown under the Disjoint (D) condition. Stacked bars decompose model responses to isolate specific multimodal failure modes: correct answers, skipped text-vision mappings (Label), incorrect visual target selections (Wrong label), and standard logical/perceptual failures (Wrong answer). The black markers denote the baseline Anonymous (A) accuracy, isolating the exact performance gap introduced by decoupling the text from the image.

a 30.0% mapping gap compared to Llama's 22.7%. Similarly, Qwen3-VL-235B and GLM-4.6V share baseline scores near 65%, but Qwen's gap is 6.8 points smaller. This variance indicates that architectural decisions—such as cross-attention routing—are significantly more determinative of fusion capability than raw model scale.

5.2 The Directionality of Binding: Referencing vs. Dereferencing

To isolate mapping mechanics, we decompose Disjoint errors (Figure 2a) into four modes: skipped mapping (Label, returning the dummy reference of the correct element), incorrect target selection with skipped mapping (Wrong label), incorrectly selected answers (mapped but wrong, Wrong answer), and invalid responses. This categorization exposes a severe directional vulnerability. On *Referencing only* tasks (text-to-vision), Label and Wrong label errors are exactly 0%; the performance drop is driven entirely by Wrong answers. This indicates MLLMs effectively project textual queries into the visual space to locate targets.

However, when operations require *Dereferencing* (vision-to-text), performance collapses. Instead of completing the cross-modal loop, models frequently terminate reasoning in the visual domain. On *Dereferencing only* tasks, 52% of responses incorrectly yield a visual dummy label (36% Label, 16% Wrong label). This asymmetry dictates question-level outcomes: referencing-heavy Value lookup queries are bottlenecked only by logical failures, whereas Spatial queries—which force dereferencing—plummet to 49% Disjoint accuracy despite a 95% visual baseline.

We hypothesize this reflects a fundamental limitation in cross-modal working memory. Once a model grounds its attention to execute visual logic, its latent representation loses the associative link back to

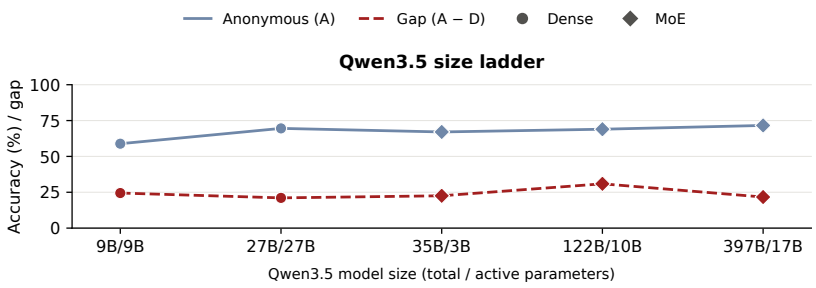


Figure 3: Anonymous accuracy and the accuracy gap across Qwen3.5 variants. Despite massive parameter scaling and architectural shifts, both baseline visual accuracy and the mapping gap plateau.

the disjoint text. Outputting the dummy label as the terminal answer suggests that text-vision binding is not a persistent state, but a fragile, unidirectional pointer.

5.3 Structural Vulnerabilities: Perception vs. Alignment

The severity of the dereferencing failure is highly dependent on chart topology (Figure 2b), revealing two structural bottlenecks.

The Alignment Bottleneck. *Direct-mapping* charts (pie and treemap) arrange data in non-sequential 2D layouts, forcing the model to map a detached semantic legend to an arbitrary region. While their baseline Anonymous accuracy is high, Disjoint performance is decimated by Label errors across GPT-6 Sol, GLM-4.6V, and Llama 4 Maverick. This highlights a pervasive vulnerability in pure color-legend binding. Conversely, *reading-order* charts (bar, line) provide a 1D coordinate system that acts as a spatial crutch. While top-tier models like GPT-6 Sol exploit this sequential alignment to resolve mappings, weaker models still suffer substantial reference-based errors, proving spatial positioning alone does not guarantee robust binding.

The Perception Bottleneck. *Error-bar* and *error-point* charts present a fundamentally different failure mode. Anonymous accuracy on these charts frequently falls below 60%. Crucially, their mapping gaps are dominated by Wrong answers and Wrong labels rather than isolated skipped mappings. This indicates an inability to resolve dense geometries. For these structures, the reasoning process fails at the low-level vision encoder long before higher-order semantic alignment can occur.

5.4 The Scale Illusion: Parameter Count vs. Spatial Grounding

A central assumption in modern multimodal architecture design is that scaling model parameters yields proportional performance gains. To test whether the mapping gap is merely a symptom of insufficient capacity, we evaluate performance across the Qwen3.5 family size ladder. This provides a controlled ablation—spanning from a 9B dense model to a 397B Mixture-of-Experts (MoE) architecture—that isolates language and routing capacity scaling while maintaining a consistent vision-encoder front-end.

Figure 3 plots the baseline Anonymous (A) accuracy alongside the Disjoint mapping gap ($A - D$). Following a modest initial improvement from the 9B to the 27B model, baseline visual perception completely flattens near 70%. As parameter counts increase into the massive MoE architectures (up to 397B), the steep scaling laws typical of text benchmarks are entirely absent. Furthermore, the mapping gap remains persistent between 20% and 30% across the entire size ladder, regardless of model scale or MoE routing.

This plateau is a highly diagnostic feature of current MLLM design, corroborating emerging research on multimodal bottlenecks. Recent studies demonstrate that visual reasoning failures are predominantly rooted in the static vision encoder rather than the language model (Tong et al., 2024). While architectures like Qwen3.5 integrate modalities through massive multimodal training, their scaling strategy heavily biases the downstream reasoning layers. As the model expands to 397B parameters, developers scale the LLM backbone and MoE routing pathways, but the front-end visual processing

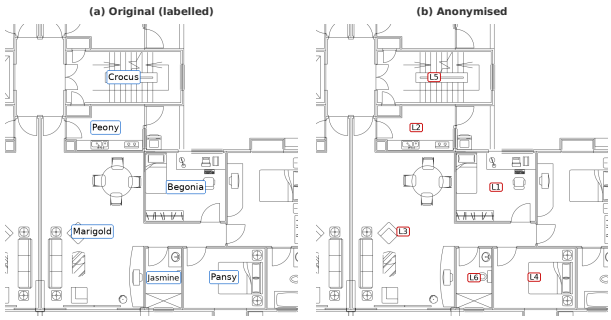


Figure 4: An example from the floorplan evaluation.

Table 3: Floorplan accuracy (%) on 507 questions across 24 plans.

Model	O	A	D	**A−D**
GPT-6 Sol	80.0	77.7	73.2	**4.5**
Claude Sonnet 5	78.8	67.5	62.9	**4.5**
Qwen3-VL-235B	85.5	79.7	41.2	**38.5**
DeepSeek V4.1 Flash	67.9	61.3	53.1	**8.3**
Qwen3.8-27B	80.6	79.7	72.4	**7.3**
Qwen3.5-35B-A3B	78.2	73.6	46.2	**27.4**
Qwen3.5-27B	85.5	83.6	31.4	**52.3**
Qwen3.5-9B	77.6	80.1	19.3	**60.7**

still bottlenecks through a fixed Vision Transformer (ViT) with inherent spatial blindness (Alam et al., 2026).

Because our Anonymous condition isolates spatial grounding from world knowledge, we hypothesize that the smaller dense models already possess sufficient logic to execute the reasoning tasks. The performance ceiling is instead dictated by the vision encoder's resolving power. Expanding the downstream language capacity to 397B parameters cannot retrospectively reconstruct dense 2D spatial relationships that the front-end ViT failed to parse (Liu et al.). The flat scaling curve strongly suggests that scaling language parameters alone is insufficient to bypass low-level perception limits. Overcoming this barrier will likely require a fundamental redesign of how visual representations are encoded and bound to text.

5.5 Uni-modal Ablation and Beyond Charts

Benchmark Validity via Uni-modal Ablation. Across QA instances, the mean Visual-Only and Text-Only accuracies over 11 models are 9.40% and 6.18%, respectively (Appendix B, Table 4). These low accuracies provide a sanity check against unimodal shortcuts in the benchmark.

Extensibility: The Alignment Bottleneck Beyond Charts. To verify this bottleneck extends beyond charts, we applied our decoupling procedure to architectural floorplans (Figure 4). Results across 507 questions reveal a stark architectural divergence (Table 3). Top-tier models (e.g., GPT-6 Sol, Claude Sonnet 5) show high resilience with single-digit mapping gaps, likely aided by the lower visual density and simpler layouts of floorplans. Conversely, the Qwen families still suffer severe degradation, losing up to 60.7 percentage points. This confirms that spatial alignment vulnerabilities are not chart-specific artifacts, but generalized multimodal deficits dictated fundamentally by model architecture.

6 Conclusion

We introduced **`CrossRef-VQA`** to isolate the fundamental ability of MLLMs to fuse disjoint text and vision modalities. By enforcing strict cross-modal dereferencing, we reveal a persistent mapping gap across frontier models driven by two architectural vulnerabilities: an alignment bottleneck in 2D color-legend binding (e.g., pie charts) and a low-level perception bottleneck on dense geometries (e.g., error bars). Crucially, this fusion capability does not scale with parameter count. Our Qwen3.5 ablation demonstrates that expanding the language backbone to 397B parameters fails to close the mapping gap, indicating that native vision-language binding mechanisms determine success far more than raw LLM size. **`CrossRef-VQA`** highlights the urgent need to shift from language-only scaling toward architectures enabling robust cross-modal alignment.

Limitations. First, our core benchmark focuses on structured graphics (charts and floorplans); extending this disjoint methodology to natural images or unstructured documents requires fundamentally different semantic decoupling strategies. Second, our scaling ablation was restricted to the Qwen3.5 family to exploit its consistent size ladder; future work must verify if these plateaus persist across different vision-encoder paradigms. Finally, while spanning five reasoning categories, our question templates may not capture all real-world multimodal reasoning patterns.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data (gpt-4o was used to fill in the dataschema in order to generate real-life alike charts and figures, but the rendering is programmatic), implement methods (with the help of codex and claude code), support qualitative and thematic data analysis (writing code and executing statistical calculations of confidence intervals), interpret results (slightly, AI are asked to search for other works that support or are against the findings in this work).

We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulate mathematical claims, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset.

And provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, are not applicable to this work.

Additionally, we used generative AI tools for: create or modify scientific figures or images, create or edit software code, summarize or analyse existing literature, sourcing/searching for information, propose a title or keywords for a research paper. We have reviewed all AI-assisted work: LLM-generated code was verified and tested for correctness by 2 authors, all AI written code comes with a large amount of testing and visualization of the results, then manually reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025.

Nahid Alam, Leema Krishna Murali, Siddhant Bharadwaj, Patrick Liu, Timothy Chung, Drishti Sharma, Kranthi Kiran, Wesley Tam, Bala Krishna S Vegesna, et al. The spatial blindspot of vision-language models. *arXiv preprint arXiv:2601.09954*, 2026.

Anthropic. Claude Sonnet 5 system card, June 2026. URL https://www.anthropic.com/claude-sonnet-5-system-card.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Paul Chandler and John Sweller. Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4):293–332, 1991.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. *arXiv preprint arXiv:2609.19969*, 2026. URL https://arxiv.org/abs/2609.19969.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3867–3876, 2025.

Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. Wikimixqa: A multimodal benchmark for question answering over tables and charts. *arXiv preprint arXiv:2506.15594*, 2025.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Paul Ginns. Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and instruction*, 16(6):511–525, 2006.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. URL https://arxiv.org/abs/2507.01006. Technical report recommended by the GLM-4.6V model card.

Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. On the perception bottleneck of vlms for chart understanding, 2025. *URL https://arxiv. org/abs/2503*, 18435.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. Insight over sight: Exploring the vision-knowledge conflicts in multimodal llms. *arXiv preprint arXiv:2410.08145*, 2024.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pp. 1527–1536, 2020.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

OpenAI. GPT-6 Astra system card, 2026. URL https://deploymentsafety.openai.com/gpt-6-astra/sec:appendix-sol-luna. GPT-6 Sol and GPT-6 Luna appendix, added September 22, 2026.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8. Family citation recommended by the Qwen3.8-27B model card.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021.

Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Liu, Zayne Sprague, et al. Chartmuseum: Testing visual reasoning capabilities of large vision-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Yang Tian, Zheng Lu, Mingqi Gao, Zheng Liu, and Bo Zhao. Mmcr: Benchmarking cross-source reasoning in scientific papers. *arXiv preprint arXiv:2503.16856*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024a.

Yucheng Wang, Yifan Hou, Aydin Javadov, Mubashara Akhtar, and Mrinmaya Sachan. Compose and fuse: Revisiting the foundational bottlenecks in multimodal reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 11272–11301, 2026.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024b.

xAI. Grok 4.3. API documentation, 2026. URL `https://docs.x.ai/developers/models/grok-4.3`. Accessed September 26, 2026.

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

An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, et al. List items one by one: A new data source and learning paradigm for multimodal llms. *arXiv preprint arXiv:2404.16375*, 2024.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. *arXiv preprint arXiv:2502.16033*, 2025.

Yuwei Yang, Zeyu Zhang, Yunzhong Hou, Zhuowan Li, Gaowen Liu, Ali Payani, Yuan-Sen Ting, and Liang Zheng. Effective training data synthesis for improving mllm chart understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2653–2663, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Table 5: Two classifications of the same 92 recoverable designed templates for the eight chart types.

(a) By task category.

Category	Example template	Templates
Extrema	Which category has the highest value for {series}?	38
Value lookup	What is the value of series {series} at category {category}?	18
Comparative	How many slices have a value larger than {slice}?	18
Aggregate	What is the mean of the values of series {series}?	11
Spatial	Which section is directly above {section}?	7
Total		**92**

(b) By required resolution direction.

Type	Example	Templates
Referencing only	How many slices have a value larger than Gas Giants? **Answer: 5**.	42
Dereferencing only	Which slice has the largest value? **Answer: Super-Earths**.	24
Both	Which slice sits directly between Gas Giants and Terrestrial? Give the slice and its percentage. **Answer: Hot Jupiters, 15.7**.	26
Neither	No templates in this category.	0
Total		**92**

A Question Template Examples

B Unimodal Ablation Results

C Chart construction and examples

Structured data and rendering. Chart data are generated by asking GPT-4o to populate chart-specific structured schemas. These schemas specify the element names, numerical values, and chart metadata used by the renderer. The resulting data are rendered into charts, and question templates are instantiated from the same structured data, with answers computed by template-specific functions. The evaluated chart types comprise pie, treemap, bar, line, box, violin, error-bar, and error-point charts. Figure 5 shows representative anonymised charts together with their associated text keys.

Table 4: Unimodal ablation accuracy (%).

Model	V-only (%)	T-only (%)
GPT-6 Sol	14.06	10.69
Claude Sonnet 5	11.13	9.66
Qwen3-VL-235B	8.20	5.56
DeepSeek V4.1 Flash	10.25	5.42
GLM-4.6V	11.27	9.37
Qwen3.8-27B	7.76	7.47
Llama 4 Maverick	9.22	4.98
GPT-4o	8.93	5.42
Grok 4.3	8.20	2.20
GPT-6 Luna	7.47	3.07
Gemma 3 27B	6.88	4.10
Average	**9.40**	**6.18**

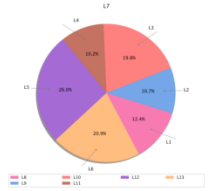
Anonymisation. The rendering pipeline produces an original image containing names and an anonymised counterpart in which selected text elements are replaced by references such as `L1`, `L2`, and so forth. Numerical ticks and data annotations designated for retention remain visible. The images are derived from the same structured chart data. This separates the visual evidence needed to identify or measure an element from the textual evidence needed to recover its name. In pie charts, a slice label and its legend entry can have distinct references associated with the same identity; the key preserves both associations.

Text keys. Keys are generated from the structured labels and the reference mapping produced during anonymisation. Pie and treemap charts use a *direct-mapping* format, explicitly associating references with identities, for example, `L1 = Biomass Energy`. Bar, line, box, violin, error-bar, and error-point charts use a *reading-order* format, listing the relevant category or series names in their corresponding order. Keys may also describe the chart title and axes. These formats are properties

of the chart construction: because format varies with chart type, comparisons between them do not isolate an effect of caption format alone.

Construction checks. The preparation pipeline checks asset availability, question and answer consistency, reference availability, and agreement between question identities and the text key. It also checks that identities intended to be hidden do not remain in the reference-phrased questions. Visual inspection and rendering checks address overlapping labels and insufficiently distinguishable series colours. Eligibility rules exclude unsuitable cases such as continuous axes whose values cannot serve as discrete element identities. These checks support consistency between the data, render, key, and instantiated questions; they do not guarantee that every rendered element is equally easy to read.

pie direct mapping



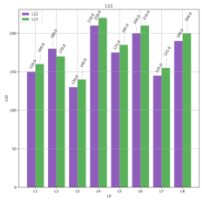
The chart title is: Global Distribution of Renewable Energy Investments in 2023. On the pie chart, the slices are labeled: L1 = Biomass Energy, L2 = Ocean Energy, L3 = Hydropower, L4 = Geothermal Energy, L5 = Wind Energy, L6 = Solar Energy, L8 = Biomass Energy, L9 = Ocean Energy, L10 = Hydropower, L11 = Geothermal Energy, L12 = Wind Energy, L13 = Solar Energy.

treemap direct mapping



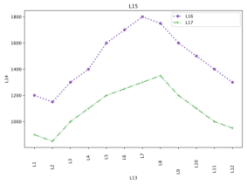
The chart title is: Distribution of Energy Consumption in Particle Accelerators. On the treemap, the sections are labeled: L1 = Beam Diagnostics, L2 = Magnet Power Supplies, L3 = RF Cavities, L4 = Control Systems, L5 = Vacuum Systems, L6 = Cryogenics, L7 = Experimental Detectors, L8 = Cooling and Ventilation.

bar reading order



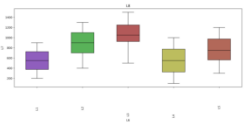
The chart title is: Popularity of Art Movements in Social Media Discussions. The x-axis label is: Art Movements. The y-axis label is: Mentions on Social Media (in thousands). The categories are: Impressionism, Cubism, Abstract, Surrealism, Modernism, Postmodernism, Pop Art, Conceptual Art. The data series are named: 2022, 2023.

line reading order



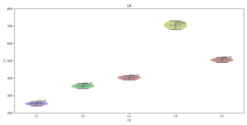
The chart title is: Trends in Art Gallery Visitor Attendance Over a Year. The x-axis label is: Month. The y-axis label is: Number of Visitors. The x-axis tick labels are: January, February, March, April, May, June, July, August, September, October, November, December. The data series are named: Modern Art Gallery, Contemporary Art Space.

box reading order



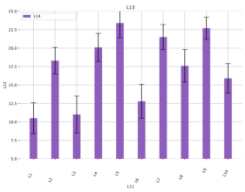
The chart title is: Population Distribution of Major Ancient Civilizations. The x-axis label is: Civilizations. The y-axis label is: Population (in thousands). The data series are named: Maurya Empire, Roman Empire, Han Dynasty China, Achaemenid Empire, Ancient Egypt.

violin reading order



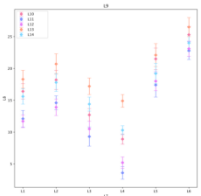
The chart title is: Population Distribution in Major Roman Cities (100 AD). The x-axis label is: Cities. The y-axis label is: Population (in thousands). The data series are named: Ephesus, Carthage, Antioch, Rome, Alexandria.

error bar reading order



The chart title is: Average Monthly Temperature Variations in Major Cities. The x-axis label is: City. The y-axis label is: Average Temperature (°C). The categories are: New York, Los Angeles, Chicago, Houston, Phoenix, Philadelphia, San Antonio, San Diego, Dallas, San Jose. The data series are named: Average Temperature.

error point reading order



The chart title is: Average Monthly Temperature Variations in Major Cities. The x-axis label is: Cities. The y-axis label is: Temperature (°C). The categories are: January, February, March, April, May, June. The data series are named: Tokyo, New York, Paris, Sydney, Cape Town.

Figure 5: Examples of the chart types in the Disjoint condition. Each panel shows an anonymised chart and its associated text key. Pie and treemap keys explicitly associate references with names; the remaining types list names in reading order. The examples illustrate the construction and the two key formats.

Table 6: Main and Qwen3.5 ladder model identifiers, temperature, leading recorded hosts, and total token counts.

Identifier	Leading recorded hosts	Temp.	Total tokens
`openai/gpt-6-sol`	OpenAI	0	8,544
`anthropic/claude-sonnet-5`	Claude Platform on AWS	0	4,889
`qwen/qwen3-vl-235b-a22b-instruct`	Alibaba, Parasail, DeepInfra	0	2,627
`deepseek/deepseek-v4.1-flash`	DeepInfra, Together, Novita	0	1,073
`z-ai/glm-4.6v`	Z.AI, Novita	0	6,153
`qwen/qwen3.8-27b`	Darkbloom, DeepInfra, Reka	0	5,734
`meta-llama/llama-4-maverick`	DigitalOcean, DeepInfra, Parasail	0	2,420
`openai/gpt-4o`	OpenAI, Azure	0	985
`x-ai/grok-4.3`	xAI	0	2,690
`openai/gpt-6-luna`	OpenAI	0	8,542
`google/gemma-3-27b-it`	Parasail, Novita, Nebius	0	345
Additional Qwen3.5 ladder models			
`qwen/qwen3.5-9b`	DeepInfra	0	1,057
`qwen/qwen3.5-27b`	DeepInfra	0	1,058
`qwen/qwen3.5-35b-a3b`	DeepInfra	0	1,057
`qwen/qwen3.5-122b-a10b`	Alibaba	0	2,595
`qwen/qwen3.5-397b-a17b`	DeepInfra	0	1,057

D MODEL IDENTIFIERS, HOSTS, AND TOKENS

Table 6 reports the identifiers used through OpenRouter, temperature, leading recorded serving hosts, and median token counts for the main model suite. Prompt-token counts refer to Anonymous calls and include the image and question; completion-token counts refer to Disjoint calls. They are provider-reported counts for different request components, rather than measurements in a common cross-provider tokenisation. For models routed through multiple hosts, the listed hosts are the most frequent recorded hosts; the list is not an exhaustive routing inventory, and some batch records do not identify a serving host.