Beyond Accuracy: Unveiling the Cognitive Mechanisms of MLLM Failure in Misleading Visualizations

Anonymous ACL submission

Abstract

While Multimodal Large Language Models demonstrate exceptional proficiency in chart understanding, their robustness against misleading visualizations remains a critical bottleneck. Existing benchmarks predominantly utilize flat taxonomies, conflating visual noise with semantic manipulation, thereby obscuring the deeper cognitive states behind model failures. To address this, we propose CoVis, a benchmark grounded in a four-layer cognitive taxonomy (Perception-Mapping-Reasoning-Logic) to ensure comprehensive coverage of the adversarial landscape. Using our proposed Knowledge-Level Uncertainty (KLU), we identify a systematic Cognitive Bifurcation: model failures collapse into either Cognitive Denial (confusion due to visual obstructions) or Cognitive Hijacking (delusion driven by semantic inducements). Evaluations across 10 state-of-the-art models demonstrate striking cross-model consistency in these failure mechanisms. Notably, we observe a pronounced Textual Dominance effect, where MLLMs often prioritize deceptive text labels over conflicting visual evidence, leading to high-confidence hijacking. Overall, CoVis establishes a mechanism-based evaluation paradigm, shifting the focus from surface-level error correction to the governance of internal cognitive states.

1 Introduction

Data visualizations are essential tools for communicating quantitative evidence and supporting informed decision-making across diverse fields. With the rapid development of Multimodal Large Language Models (MLLMs) (Radford et al., 2021; Alayrac et al., 2022; Team et al., 2023; Achiam et al., 2023; Dubey et al., 2024; Li et al., 2024; Wu et al., 2024), models have demonstrated perception and reasoning capabilities comparable to humans on standard chart understanding tasks (Masry et al., 2022; Xia et al., 2025). However, real-world visualizations are not always benevolent and well-formed. From distorted axes to deceptive labeling, *Misleading Visualizations* have become potent vehicles for misinformation propagation (King, 1986; Pandey et al., 2015). In high-stakes domains such as finance and healthcare, the ability to resist these visual traps is not merely a performance metric, but a prerequisite for trustworthy decision-making (Lauer and O'Brien, 2020).

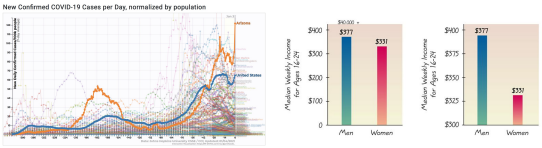

Figure 1: Diverse manifestations of misleading visualizations. (a) **Information Confusion**: Excessive visual noise (e.g., overplotting) that obscures the underlying data. (b) **Intentional Misleading**: Strategic manipulation (e.g., truncated axes) designed to induce incorrect causal conclusions.

Although recent efforts have identified significant performance drops in MLLMs facing chart deception (Chen et al., 2025; Mahbub et al., 2025), existing evaluations typically rely on flat error taxonomies and report only scalar accuracy. Such a broad approach conflates design normative errors with semantic manipulation, thereby failing to distinguish between fundamentally different failure modes, masking the deeper cognitive mechanisms that drive model fragility. As illustrated in Figure 1, these studies predominantly employ a "Flat Taxonomy," conflating *Design Normative Errors* (e.g., cluttering/confusion, Figure 1a) with *Logical Manipulative Errors* (e.g., truncated axes/misleading, Figure 1b). While the former primarily introduces perceptual noise leading to reading difficulties, the latter constructs deep semantic traps. This evaluation approach, driven solely by scalar Accuracy, obscures the internal state of the model: when a

model errs, is it simply confused by illegible pixels or deluded by a malicious narrative? Scalar metrics fail to distinguish between these fundamentally different failure scenarios.

To bridge this gap from pixel-level robustness to cognitive-level alignment, we propose the **CoVis** (Cognitive Visualization) benchmark. Unlike previous classifications based on surface features, CoVis constructs a four-layer cognitive taxonomy covering *Perception-Mapping-Reasoning-Logic*, grounded in cognitive psychology. We meticulously curated 38 core misleading types, ensuring comprehensive coverage from low-level feature extraction blockage to high-level semantic logic traps. By conducting a stress test across 10 representative SOTA models with 8,480 instances, we move limited case studies to perform a systematic auditing of model robustness.

Furthermore, by introducing Knowledge-Level Uncertainty (KLU), an information-theoretic metric that measures the divergence of a model's output distribution from a uniform state, we uncover a **Cognitive Bifurcation** that explains the contrasting failures illustrated in Figure 1. By quantifying the concentration of output logits, KLU serves as a diagnostic probe to distinguish between two orthogonal failure mechanisms:

- **Cognitive Denial**: Triggered primarily by the information confusion seen in Figure 1a. In these cases, excessive visual noise acts as an *Information Blocker*, yielding a distribution near uniform (low KLU). This represents a high-entropy state where the model admits ignorance due to blocked perception.

- **Cognitive Hijacking**: Triggered by the intentional inducements shown in Figure 1b. Here, strategic manipulations act as *Semantic Decoys*, concentrating the probability mass on a single incorrect option (high KLU). This represents a low-entropy state where the model is trapped in confident delusion.

Notably, this discovery exposes a systemic Textual Dominance phenomenon: when visual geometric features conflict with deceptive text labels, models consistently ignore visual evidence and are confidently hijacked by incorrect text.

In summary, our main contributions are as follows:

- **CoVis Benchmark:** We propose CoVis, a benchmark grounded in cognitive psychology containing 38 fine-grained error types designed for systematic robustness diagnosis.

- **KLU Diagnostic Metric:** We introduce Knowledge-Level Uncertainty (KLU), an uncertainty-based probe that decouples failure mechanisms by measuring the information-theoretic distance between model predictions and random guessing.

- **Mechanism Discovery:** We identify a Cognitive Bifurcation, the Denial and Hijacking mechanisms, along with a systemic textual dominance effect, providing empirical evidence to guide future safety alignment.

2 Related Work

2.1 MLLM Benchmarks for Chart Understanding

With the rapid advancement of Multimodal Large Language Models (MLLMs), benchmarks for chart understanding have emerged extensively. Early benchmarks like *PlotQA* (Methani et al., 2020) and *ChartQA* (Masry et al., 2022) and focused on basic capabilities such as data extraction and reasoning from standard charts. Subsequently, comprehensive benchmarks like *Charxiv* (Wang et al., 2024), *Chartbench* (Xu et al., 2023), *Evochart* (Huang et al., 2025), and so on (Wang et al., 2024; Masry et al., 2025), which expanded coverage to broader disciplines, driving models to achieve near-human performance in perception and reasoning tasks. However, these mainstream benchmarks are predominantly composed of normal visualizations. They assume that visual inputs truthfully reflect the underlying data, thereby evaluating the models' performance rather than their robustness. As our study reveals, high accuracy on clean data often masks fragility against adversarial perturbations.

2.2 Misleading Visualization

With the development of MLLMs, the community has begun to address the security threats posed by misleading visualizations. Early research in Human-Computer Interaction (HCI) established the theoretical landscape: Lo et al. (2022) identified 74 distinct misleading elements in real-world charts, while Lisnic et al. (2023) highlighted how visual tricks on social media are coupled with textual framing to manipulate perception.

Recently, automated evaluation for MLLMs has emerged. Lo and Qu (2024); Alexander et al.

(2024) pioneered the probing of models' detection capabilities via prompting. Subsequently, large-scale benchmarks such as *Misleading ChartQA* (Chen et al., 2025), *The Perils of Chart Deception* (Mahbub et al., 2025), *MisViz* (Tonglet et al., 2025b) have quantified the significant performance degradation of state-of-the-art models against various deceptive designs. Furthermore, research has extended to downstream applications, including defense (Tonglet et al., 2025a; Song et al., 2025), interactive correction (Das and Mueller, 2025), and rhetorical intent analysis (Blasilli and Angelini).

Despite confirming the vulnerability of MLLMs, current evaluation paradigms suffer from two critical limitations. First, existing benchmarks predominantly employ a Flat Taxonomy (Chen et al., 2025; Tonglet et al., 2025b), categorizing "Design Normative Errors" (e.g., color misuse) alongside "Logical Fallacies" (e.g., semantic manipulation) as parallel labels without structural distinction. This approach conflates *Visual Noise* (perception level) with *Semantic Manipulation* (logic level). Second, current assessments rely almost exclusively on scalar metrics (e.g., Accuracy), lacking a Cognitive Diagnosis of the model's internal state. While uncertainty estimation is widely used in NLP for hallucination detection (Kadavath et al., 2022; Huang et al., 2023), it remains unexplored in the context of adversarial charts. Consequently, existing works cannot distinguish between *Cognitive Denial* (a high-entropy state of admitted ignorance due to blocked perception) and *Cognitive Hijacking* (a low-entropy state of confident delusion driven by semantic traps). **CoVis** addresses these limitations by introducing a four-layer cognitive taxonomy and the **Knowledge-Level Uncertainty (KLU)** metric, shifting the focus from performance benchmarks to a deeper analysis of internal failure states.

3 The CoVis Benchmark

Rather than relying on web-scraped images (Lo et al., 2022), which can introduce confounding factors like noise and erratic formatting, we utilize a script-based synthesis framework. This approach leverages plotting libraries to generate charts programmatically, providing direct control over visual parameters. Consequently, we can apply specific misleading designs while ensuring absolute alignment between the visual representation and the ground-truth information.

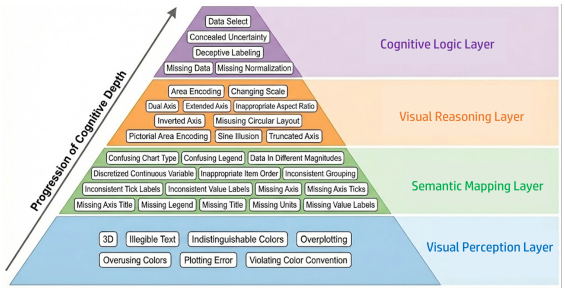

Figure 2: The structural composition of the CoVis four-layer cognitive pipeline. The architecture is organized into four sequential stages: *Perception*, *Mapping*, *Reasoning*, and *Logic*, which categorize 38 core misleading types into a hierarchical system.

3.1 Theoretical Basis: The 4-Layer Cognitive Pipeline

To identify the underlying failure mechanisms triggered by misleading visualizations, we structure our benchmark based on established principles of cognitive psychology and visualization theory. We adapt Pinker's model of graph comprehension (Pinker, 2014; Kosslyn, 2006; Khalil et al., 2005) to investigate the specific stages where MLLMs become vulnerable to deception. By utilizing a four-layer cognitive pipeline, we can diagnose how perturbations at different levels, from initial signal perception to high-level logical synthesis, result in model failure, allowing for a detailed analysis of the internal processes that drive these errors.

As illustrated in Figure 2 and detailed in Appendix A.1, we curate 38 core error types of visual misleadingness mapped to these four stages:

1. **Visual Perception Layer (Signal Acquisition):** Targets the raw extraction of visual features. We inject physical noise (e.g., overplotting, indistinguishable colors) to sever the link between pixels and objects.

2. **Semantic Mapping Layer (Decoding):** Targets the translation of visual variables (e.g., position, length) into data concepts. We remove or obscure the reference systems (e.g., missing axes, inconsistent labels) required for decoding.

3. **Visual Reasoning Layer (Operation):** Targets the geometric comparison and calculation processes. We introduce spatial distortions (e.g., truncated axes, dual scales) that violate standard visual grammar.

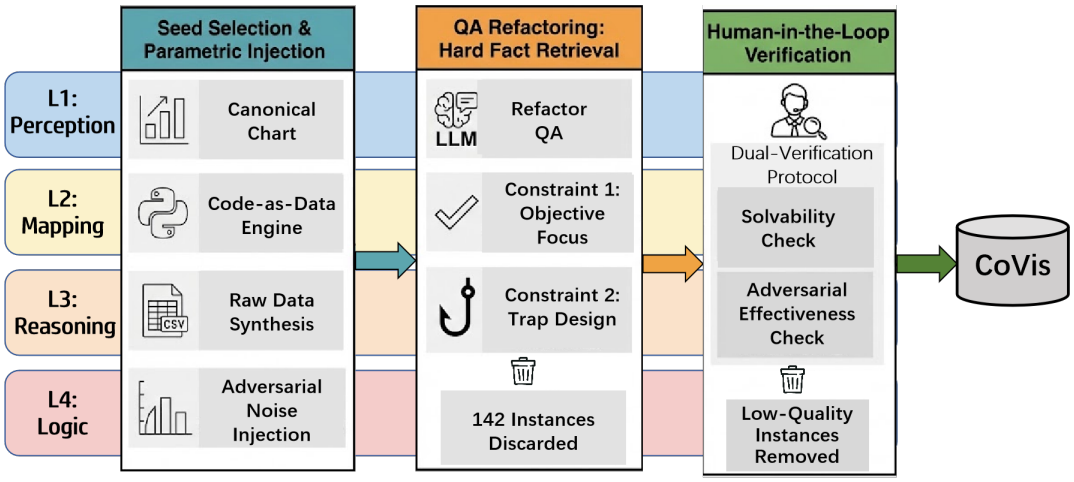

Figure 3: The CoVis data construction pipeline. We utilize a **Code-as-Data** engine to synthesize charts with parametric perturbations, followed by LLM-driven QA refactoring and a stringent human-in-the-loop verification process, resulting in 848 high-quality instances.

4. **Cognitive Logic Layer (Decision):** Targets high-level semantic interpretation. We manipulate contextual information (e.g., data selecting, concealed uncertainty) to induce logical fallacies despite correct visual perception.

3.2 Construction Workflow

3.2.1 Seed Selection & Parametric Injection

Our data generation pipeline begins with the selection of canonical chart schemas. We ground our seed charts in the taxonomy defined by VLAT (Lee et al., 2016), a rigorously validated visualization literacy assessment framework. This ensures that our base charts adhere to standard perceptual principles before perturbation. Leveraging a **"Code-as-Data"** paradigm, we developed a Python-based generation engine utilizing Matplotlib.

To ensure a balanced representation across the four cognitive layers, we initially instantiated a total of **1,250 candidate chart-data pairs**. For each instance, the engine performs two synchronized operations:

- **Raw Data Synthesis**: It generates underlying data tables (CSV) with controlled statistical properties, such as linear trends, clusters, or seasonal fluctuations.
- **Adversarial Noise Injection**: It programmatically alters rendering parameters to create specific error types, such as adjusting `ax.set_ylim()` for *Truncated Axis* or setting `alpha=0.1` for *Indistinguishable Colors*, while maintaining pixel-perfect alignment with the ground-truth numerical data.

3.2.2 QA Refactoring: From Meta-Cognition to Fact Retrieval

A common pitfall in existing benchmarks is the use of meta-cognitive questions (e.g., "Is this chart misleading?"). Such framing introduces a **"suspicion bias,"** priming the model to look for errors regardless of visual evidence. To eliminate this confounder, we refactor the task into **"Hard Fact Retrieval."** We employ GPT-4o to generate multiple-choice questions based solely on the raw data tables. The prompting strategy (detailed in Appendix A.3) enforces specific constraints:

- **Objective Focus**: Questions must target specific numerical data points or categorical trends rather than asking about the chart's design quality.
- **Trap Design**: Distractor options must explicitly include values suggested by the visual manipulation, such as the perceived value in a truncated axis or a distorted area encoding, to act as "lures" for *Cognitive Hijacking*.

During this stage, **142 instances** were automatically discarded due to LLM-generated questions that were either ambiguous or not strictly answerable from the CSV data, leaving 1,108 candidates for human review.

3.2.3 Human-in-the-loop Verification

To ensure the validity of the generated QA pairs, we implemented a rigorous human-in-the-loop verification process via a custom-built Interactive Annotation Platform (see Appendix A.2). Three expert

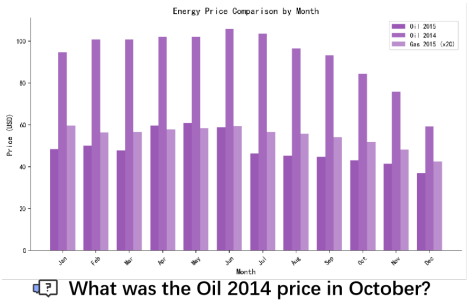

Figure 4: A representative sample from the **CoVis** benchmark. The figure displays a misleading chart with "Indistinguishable Colors" (L1) and its corresponding adversarial QA pair. The distractors are designed as "lures" based on the visual distortion to probe for Cognitive Denial.

annotators with backgrounds in data visualization performed a **dual-verification protocol**:

- **Solvability Check**: Verifying that the Ground Truth option is strictly derivable from the underlying data table.
- **Adversarial Effectiveness Check**: Confirming that the chart visually suggests the "Trap Option" (for Hijacking types) or renders the data genuinely illegible (for Denial types).

Figure 4 illustrates a representative CoVis instance featuring an "Indistinguishable Colors" error (L1), demonstrating how visual obstructions trigger **Cognitive Denial** via our QA refactoring.

4 Experiment

4.1 Experiment Setup

4.1.1 Evaluated Models & Implementation

Model Selection Criteria To ensure a comprehensive evaluation of the MLLM landscape, we curate a set of 10 state-of-the-art models spanning five distinct architectural families, each represented by a small and a large parameter variant (as detailed in Table 1). This paired design allows us to control for architectural differences while isolating the effects of model scaling on cognitive states:

- **Qwen3-VL (8B & 32B):** Represents the latest generation of MLLMs, possessing advanced visual understanding and reasoning capabilities.
- **DeepSeek-VL2 (Tiny & Native):** Represents models optimized for reasoning, incorporating mixture-of-experts (MoE) and reinforcement learning strategies.
- **Llama-3.2-Vision (11B & 90B):** A widely adopted, representative open-weights baseline for multimodal tasks.
- **LLaVA-OneVision (7B & 72B):** Represents the "Strong Encoder" paradigm, utilizing SigLIP-SO400M and a specialized "AnyRes" strategy for high-resolution details.
- **LLaVA-v1.6-Vicuna (7B & 13B):** Serves as a classic baseline representing the widely adopted LLaVA architectures.

Inference Settings We evaluate all models in a zero-shot setting to strictly probe their intrinsic capabilities without the influence of few-shot demonstrations. To ensure the Knowledge-Level Uncertainty (KLU) metric reflects the model's true underlying probability distribution rather than sampling noise, we perform inference using a decoding temperature of 0.01 and a nucleus sampling threshold of 0.001. For each multiple-choice question, we extract the raw logits associated with the option tokens (e.g., "A", "B", "C", "D") and normalize them via Softmax to compute the probability distribution P required for KLU calculation. All experiments were conducted on a cluster of NVIDIA H100 (80GB) GPUs.

4.1.2 Evaluation Metrics

Accuracy We report the standard Accuracy (ACC) to measure the model's success rate in factual retrieval. While ACC reflects the final utility, we argue it is insufficient for adversarial evaluation: a score of zero conflates "ignorant guessing" with "confident hallucination," masking the distinct safety risks associated with each.

Knowledge-Level Uncertainty (KLU) Standard scalar metrics treat all incorrect predictions equally, failing to distinguish between a model that is "confused" and one that is "misled." To decouple these failure modes, we introduce **Knowledge-Level Uncertainty (KLU)**, a diagnostic probe based on the information-theoretic distance between the model's confidence distribution and random guessing.

Model	L1: Perception	L2: Mapping	L3: Reasoning	L4: Logic	Average
DeepSeek-VL2-Tiny	0.320	0.397	0.732	0.160	0.402
DeepSeek-VL2-Native	0.554	0.580	0.727	0.540	0.600
Llama-3.2-11B-Vision	0.117	0.264	**0.939**	0.100	0.355
Llama-3.2-90B-Vision	0.221	0.351	0.797	0.440	0.452
LLaVA-v1.6-Vicuna-7B	0.167	0.299	0.801	0.220	0.372
LLaVA-v1.6-Vicuna-13B	0.162	0.383	0.636	0.380	0.390
LLaVA-OneVision-7B	0.608	0.528	0.519	0.420	0.519
LLaVA-OneVision-72B	0.626	0.574	0.710	**0.660**	0.642
Qwen3-VL-8B	0.685	0.545	0.528	0.580	0.584
Qwen3-VL-32B	**0.716**	**0.597**	0.788	**0.660**	**0.690**

Table 1: Zero-shot accuracy across cognitive layers. Best performance is in **bold**.

Mathematical Formulation: For a given question with K options (where $K = 4$ in CoVis), let $\mathcal{P} = \{p_1, \ldots, p_K\}$ denote the model's softmax probability distribution over the answer space, and $\mathcal{U} = \{1/K, \ldots, 1/K\}$ denote the uniform distribution representing maximum uncertainty. We define KLU as the Kullback-Leibler divergence of $\mathcal{P}$ from $\mathcal{U}$:

$$\mathrm{KLU}(\mathcal{P}) = D_{KL}(\mathcal{P} \parallel \mathcal{U}) = \ln K - H(\mathcal{P}) \quad (1)$$

where $H(\mathcal{P})$ is the Shannon entropy. Physically, KLU quantifies the information gain of the model's prediction relative to a random guess.

Rationale & Interpretation: In the context of misleading visualizations, KLU serves as a probe for the model's certainty landscape when prediction fails:

- **Lower KLU (Distributional Flatness):** This indicates that the model's output approaches uniformity ($\mathcal{P} \approx \mathcal{U}$). It signifies a state of High Entropy, implying that the model fails to extract discriminative features from the visual input and effectively reverts to random guessing (*Cognitive Denial*).

- **High KLU (Probability Concentration):** This indicates that the probability mass is heavily concentrated on a specific option. When the prediction is incorrect, a high KLU signifies a state of False Certainty, implying that the model has been strongly anchored to a specific wrong answer by the adversarial features (*Cognitive Hijacking*).

By analyzing the distribution of KLU, we can determine whether the robustness failure stems from a lack of discriminative power (*Flatness*) or misaligned confidence (*Concentration*).

4.2 Main Results

4.2.1 Overall Vulnerability

We first assess the absolute robustness of MLLMs against the CoVis benchmark. As presented in Table 1, the results reveal a significant performance gap compared to standard chart evaluation benchmarks. The average accuracy across all 10 evaluated models is merely **50.1%**, indicating that nearly half of the adversarial samples successfully bypass current model defenses. Even the state-of-the-art Qwen3-VL-32B achieves only 69.0% accuracy, leaving a substantial 31% robustness gap. More critically, widely used foundational models such as Llama-3.2-11B-Vision (35.5%) and LLaVA-v1.6-7B (37.2%) perform dangerously close to the random guessing baseline (25%). This underscores that robustness against visual misleadingness remains a systemic blind spot, irrespective of the model family.

4.2.2 The Impact of Scale

Analyzing the performance delta between paired small and large models reveals that scaling parameters does not benefit all cognitive layers symmetrically (see Figure 5):

- **Logic Benefits from Scale**: The most substantial gains are observed in Layer 4. For instance, scaling DeepSeek-VL2 from Tiny to Native boosts Layer 4 accuracy by **+38.0%**, and Llama-3.2-Vision from 11B to 90B by **+34.0%**.

- **Perception Hits a Ceiling**: Conversely, gains in Layer 1 are marginal or inconsistent. While LLaVA-OneVision gains only +1.8% when scaling from 7B to 72B, LLaVA-v1.6 sees a slight regression (-0.5%). This suggests a "Perceptual Ceiling" in current architectures.

Error Type	L	State	Score	Error Type	L	State	Score
Missing Normalization	4	Hijack (T)	1.0	inapprop.itemorder	2	Hijack (T)	0.7
overplotting	1	Denial (B)	1.0	changingscale	3	Denial (B)	0.7
discretized-cont.var	2	Hijack (T)	1.0	plottingerror	1	Hijack (T)	0.7
inconsist.valuelabels	2	Hijack (T)	0.9	misuse.circular	3	Denial (B)	0.7
inconsist.ticklabels	2	Denial (B)	0.9	missinglegend	2	Hijack (T)	0.7
Deceptive Labeling	4	Hijack (T)	0.9	truncatedaxis	3	Hijack (T)	0.6
Missing Data	4	Hijack (T)	0.9	sineillusion	3	Denial (B)	0.6
missingaxisticks	2	Denial (B)	0.9	areaencoding	3	Denial (B)	0.6
missingaxis	2	Denial (B)	0.9	missingtitle	2	Hijack (T)	0.6
extendedaxis	3	Denial (B)	0.9	missingabbrev.	2	Denial (B)	0.6
Data Selecting	4	Hijack (T)	0.9	dualaxis	3	Hijack (T)	0.6
overusingcolors	1	Denial (B)	0.8	violating.color	1	Denial (B)	0.6
missingvaluelabels	2	Denial (B)	0.8	*confusingcharttype*	2	Ambiguous	0.5
missingunits	2	Hijack (T)	0.8	*inconsist.grouping*	2	Ambiguous	0.5
invertedaxis	3	Hijack (T)	0.8	*missingaxistitle*	2	Ambiguous	0.5
3d	1	Denial (B)	0.8	*confusinglegend*	2	Ambiguous	0.5
inapprop.aspectratio	3	Denial (B)	0.8	*data-diff.mag.*	2	Ambiguous	0.5
illegibletext	1	Denial (B)	0.8	*pictorialareaenc.*	3	Ambiguous	0.5
Concealed Uncertainty	4	Hijack (T)	0.8	*indisting.colors*	1	Ambiguous	0.5

Table 2: KLU Rank Consistency across 38 error types. **State** indicates the dominant mechanism (Hijacking vs. Denial). **Score** is the Consistency Score across 10 models.

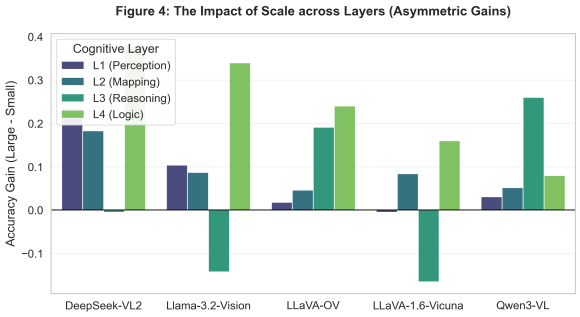

Figure 5: Performance delta across cognitive layers when scaling from small to large model variants. Scaling enhances logical synthesis (L4) but offers diminishing returns for perceptual robustness (L1).

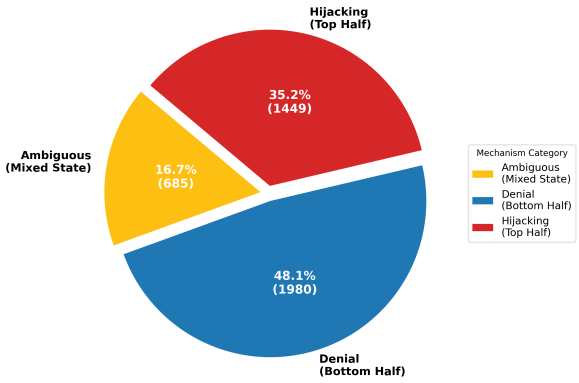

Figure 6: Prevalence of cognitive failure mechanisms. The distribution validates that models under adversarial settings tend to collapse into two states: Denial or Hijacking.

However, we observe a counter-intuitive Reasoning Regression in specific architectures. Notably, scaling Llama-3.2-Vision from 11B to 90B results in a significant performance drop in Layer 3 (Visual Reasoning), falling from 93.9% to 79.7% (**-14.2%**). We hypothesize this represents a "Curse of Prior Knowledge": smaller models may rely on naive, pixel-level geometric measurements, while larger models possess stronger priors that may override immediate visual evidence when facing counter-intuitive designs like inverted axes.

4.3 Cognitive Mechanism Analysis

4.3.1 Universality of Failure Modes

A fundamental question is whether failure modes are model-specific idiosyncrasies or universal phenomena. As detailed in Table 2, our analysis reveals a striking universality: **81.5%** of error types achieve a Consistency Score of ≥ 0.6. This indicates that regardless of architecture or scale, specific visual triggers activate the same underlying uncertainty state across the MLLM family.

4.3.2 Defining the Binary Mechanisms

Based on the observed consistency, we formally define two orthogonal cognitive states that govern MLLM failures:

- **Cognitive Denial**: Characterized by low KLU rankings. Visual noise acts as an *Information Blocker*, forcing the model into a high-entropy state of admitted ignorance.
- **Cognitive Hijacking**: Characterized by high

KLU rankings. Adversarial features act as *Semantic Decoys*, anchoring the model to a wrong answer with false certainty.

As shown in Figure 6, the failure landscape is heavily polarized, confirming that "hesitation" is rare in current MLLMs; they are either universally blinded or universally deluded.

5 Discussion

5.1 The Textual Dominance in Multimodal Conflicts

A critical insight from our consistency analysis (Table 2) is exemplified by the behavior of *Inconsistent Value Labels*. In this error type, the visual signal (the geometric height of a bar) directly contradicts the semantic signal (the text label on the axis). Despite this blatant visual mismatch, this error consistently triggers **Cognitive Hijacking** (High KLU, Top-Half Distribution) across 90% of the evaluated models.

Rather than being an outlier, this phenomenon empirically validates a fundamental Modality Bias(Wu et al., 2025; Zheng et al., 2025) inherent in current MLLM architectures: *Textual Dominance*. When faced with conflicting information from dual modalities, models systematically prioritize the textual path, effectively bypassing visual verification. This suggests that existing MLLMs are influenced by language priors and tend to ignore visual features. The visual encoder acts as a subordinate feature extractor, susceptible to being overridden by any coherent, even if deceptive, semantic narrative found in the text. This "blind trust" in text is a primary engine driving *Cognitive Hijacking*, confirming that the model's certainty is anchored more in what it reads than in what it sees.

5.2 From Error Categories to Cognitive Consequences

Existing benchmarks predominantly rely on *Flat Taxonomies*, treating diverse error types as isolated labels. To address this lack of structure, **CoVis** adopts a 4-Layer Hierarchical Taxonomy (Perception-Mapping-Reasoning-Logic), providing ontological completeness that ensures full coverage of the visual processing pipeline.

However, our empirical findings reveal that while this layered approach is structurally superior, it shares a critical limitation: it categorizes the input trigger rather than the diagnostic outcome. Our KLU analysis demonstrates that regardless of the theoretical layer, the cognitive state of a misled model predominantly collapses into one of two orthogonal attractors: *Cognitive Denial* or *Cognitive Hijacking*. We argue that the true boundary of model robustness lies in this cognitive consequence, not the error stage. A model that remains highly confident when misled (*Hijacking*) is fundamentally more hazardous than one that admits ignorance (*Denial*). Therefore, future evaluations must shift focus from *where* the error occurs to *how* the model reacts, establishing Cognitive Stability as the new standard for safety alignment research.

5.3 Implications for Adversarial Alignment

The striking cross-model consistency in hijacking mechanisms suggests that current RLHF (Reinforcement Learning from Human Feedback) may be insufficient for visual reasoning. Most alignment processes focus on linguistic safety, yet our results show that geometric deception can easily bypass these safeguards. Future alignment should incorporate "Adversarial Chart Reasoning" into the fine-tuning loop, specifically rewarding models that flag visual-textual contradictions rather than those that simply provide a confident (but wrong) answer.

6 Conclusion

In this work, we introduce **CoVis**, a diagnostic benchmark designed to evaluate the cognitive robustness of MLLMs against misleading visualizations. By shifting from flat taxonomies to a 4-layer hierarchical framework, we provide a more structured map of the adversarial landscape. Our evaluation of 10 state-of-the-art models reveals a systemic fragility: even highly capable models are susceptible to Cognitive Hijacking, a state where deceptive semantic cues override visual evidence with high confidence. The introduction of **Knowledge-Level Uncertainty (KLU)** allows us to bridge the gap between error types and failure mechanisms, uncovering a fundamental Cognitive Bifurcation between confusion and delusion. Furthermore, the observed Textual Dominance effect highlights a critical bottleneck in current multimodal architectures. Ultimately, **CoVis** advocates for a paradigm shift in evaluation, moving beyond scalar accuracy toward mechanism-based cognitive alignment. We hope this work provides the diagnostic depth necessary to develop the next generation of cognitively reliable visual agents.

Limitations

Despite the comprehensive scope of the **CoVis** benchmark and the diagnostic depth provided by the **KLU** metric, we acknowledge several constraints that offer avenues for future research:

- **Scope of Visualization Diversity:** While CoVis curates 38 core misleading types across four cognitive layers, it does not yet encompass highly specialized or domain-specific visualizations (e.g., 3D scatter plots in bioinformatics or complex interactive dashboards). Future iterations should expand to these niche categories to test the generalizability of the "Cognitive Bifurcation" theory.
- **Metric Constraints (KLU):** The calculation of Knowledge-Level Uncertainty (KLU) requires access to the model's raw output logits. While this is straightforward for open-weights models and many research APIs, it poses challenges for certain highly restricted black-box systems that provide only text strings without log-probabilities. Approximation methods for these systems remain to be explored.

Ethics Statement

This work does not involve the collection or use of any human-related data. All materials are chart-based and synthesized through our proposed generation engine. To support reproducibility and open science, the CoVis benchmark will be released under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license, and the generation code will be made available under the Apache 2.0 license. We have ensured that the adversarial perturbations and synthesized content do not introduce offensive, biased, or harmful visual-language associations. No ethical concerns were identified in the preparation of this dataset or the subsequent study.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Jason Alexander, Priyal Nanda, Kai-Cheng Yang, and Ali Sarvghad. 2024. Can gpt-4 models detect misleading visualizations? In *2024 IEEE Visualization and Visual Analytics (VIS)*, pages 106–110. IEEE.

Graziano Blasilli and Marco Angelini. True (vis) lies: A preliminary analysis of how generative ai is capable of recognizing visualization lies and their rhetoric.

Zixin Chen, Sicheng Song, Kashun Shum, Yanna Lin, Rui Sheng, Weiqi Wang, and Huamin Qu. 2025. Unmasking deceptive visuals: Benchmarking multimodal large language models on misleading chart question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13767–13800.

Amit Kumar Das and Klaus Mueller. 2025. Misvisfix: An interactive dashboard for detecting, explaining, and correcting misleading visualizations using large language models. *arXiv preprint arXiv:2508.04679*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. 2025. Evochart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3680–3688.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Mohammed K Khalil, Fred Paas, Tristan E Johnson, and Andrew F Payer. 2005. Design of interactive and dynamic anatomical visualizations: the implication of cognitive load theory. *The Anatomical Record Part B: The New Anatomist: An Official Publication of the American Association of Anatomists*, 286(1):15–20.

Gary King. 1986. How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, pages 666–687.

Stephen M Kosslyn. 2006. *Graph design for the eye and mind*. Oup usa.

Claire Lauer and Shaun O'Brien. 2020. The deceptive potential of common design tactics used in data visualizations. In *Proceedings of the 38th ACM International Conference on Design of Communication*, pages 1–9.

Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. 2016. Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics*, 23(1):551–560.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Maxim Lisnic, Cole Polychronis, Alexander Lex, and Marina Kogan. 2023. Misleading beyond visual tricks: How people actually lie with charts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–21.

Leo Yu-Ho Lo, Ayush Gupta, Kento Shigyo, Aoyu Wu, Enrico Bertini, and Huamin Qu. 2022. Misinformed by visualization: What do we learn from misinformative visualizations? In *Computer Graphics Forum*, volume 41, pages 515–525. Wiley Online Library.

Leo Yu-Ho Lo and Huamin Qu. 2024. How good (or bad) are llms at detecting misleading visualizations? *IEEE Transactions on Visualization and Computer Graphics*.

Ridwan Mahbub, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mizanur Rahman, Mir Tafseer Nayeem, and Enamul Hoque. 2025. The perils of chart deception: How misleading visualizations affect vision-language models. *arXiv preprint arXiv:2508.09716*.

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

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, and 1 others. 2025. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*.

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

Anshul Vikram Pandey, Katharina Rall, Margaret L Satterthwaite, Oded Nov, and Enrico Bertini. 2015. How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pages 1469–1478.

Steven Pinker. 2014. A theory of graph comprehension. In *Artificial intelligence and the future of testing*, pages 73–126. Psychology Press.

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

Sicheng Song, Yanjie Zhang, Zixin Chen, Huamin Qu, Changbo Wang, and Chenhui Li. 2025. Vizdefender: Unmasking visualization tampering through proactive localization and intent inference. *arXiv preprint arXiv:2512.18853*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jonathan Tonglet, Tinne Tuytelaars, Marie-Francine Moens, and Iryna Gurevych. 2025a. Protecting multimodal large language models against misleading visualizations. *arXiv preprint arXiv:2502.20503*.

Jonathan Tonglet, Jan Zimny, Tinne Tuytelaars, and Iryna Gurevych. 2025b. Is this chart lying to me? automating the detection of misleading visualizations. *arXiv preprint arXiv:2508.21675*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. 2025. When language overrules: Revealing text dominance in multimodal large language models. *arXiv preprint arXiv:2508.10552*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Renqiu Xia, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. 2025. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*.

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

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, and 1 others. 2025. Mllms are deeply affected by modality bias. *arXiv preprint arXiv:2505.18657*.

A Appendix

A.1 Detailed Taxonomy of Misleading Visualizations

Layer	Error Type	Definition and Cognitive Impact
L1	3D Effect	Using 3D perspective to represent 2D data, causing volume/area distortion.
	Illegible Text	Text rendered at extremely low resolution or overlapping, making it unreadable.
	Indistinguishable Colors	Using color scales with insufficient contrast, making categories visually merged.
	Overplotting	Excessive density of data points or lines that obscures the underlying distribution.
	Overusing Colors	Employing too many colors for categorical data, exceeding cognitive load limits.
	Plotting Error	Raw rendering failures (e.g., misaligned bars) that contradict source data.
	Violating Color Conv.	Using colors counter-intuitively (e.g., red for "good", blue for "hot").
L2	Confusing Chart Type	Using an inappropriate chart (e.g., pie chart for time series) to obscure trends.
	Confusing Legend	Legends that use ambiguous symbols or are placed far from data elements.
	Data in Diff. Mag.	Vast scale differences without clear indication, hiding small values.
	Discretized Cont. Var.	Converting continuous data into arbitrary bins to hide specific fluctuations.
	Inappropriate Item Order	Non-logical ordering of categorical data to imply a non-existent ranking.
	Inconsistent Grouping	Varying the grouping criteria across different parts of the visualization.
	Inconsistent Ticks	Non-uniform intervals between axis ticks, distorting the perceived rate of change.
	Inconsistent Labels	Textual labels on data points that contradict their actual geometric position.
	Missing Abbreviation	Non-standard abbreviations that hinder semantic decoding.
	Missing Axis	Omitting X or Y axes, leaving data points without a spatial reference frame.
	Missing Axis Ticks	Removing tick marks, making it impossible to retrieve precise numerical values.
	Missing Axis Title	Omitting labels for axes, leading to ambiguity in measurement.
	Missing Legend	Omitting the key for colors or shapes, severing the link between visual and data.
	Missing Title	Omitting the main title, depriving the model of the chart's overall context.
	Missing Units	Omitting units (e.g., $, kg), leading to magnitude confusion.
	Missing Value Labels	Omitting direct labels in cluttered charts where axis lookup is difficult.
L3	Area Encoding	Using area to represent data where the scale is non-linear.
	Changing Scale	Dynamically altering the scale within a single axis (e.g., linear to log).
	Dual Axis	Using two different Y-axes to imply a false correlation.
	Extended Axis	Stretching axes unnecessarily to flatten trends and minimize change.
	Inapprop. Aspect Ratio	Distorting height/width ratio to exaggerate or dampen slopes.
	Inverted Axis	Reversing axis direction to flip the perceived trend.
	Misusing Circular	Using radar/pie charts for data that does not sum to a meaningful whole.
	Pictorial Area	Using icons of varying sizes where area vs. height growth is confusing.
	Sine Illusion	Optical illusion where vertical distance between curves is misperceived.
	Truncated Axis	Starting Y-axis at a non-zero value to exaggerate small differences.
L4	Data Selecting	Selective data presentation to support a biased or false narrative.
	Concealed Uncertainty	Presenting estimates as absolute facts without indicating error margins.
	Deceptive Labeling	Using biased language in titles/labels to prime a specific conclusion.
	Missing Data	Strategically omitting unfavorable time periods or data categories.
	Missing Normalization	Comparing raw counts for groups of different sizes, causing unfair bias.

Table 3: Comprehensive Definitions of the 38 Misleading Types.

A.2 Human-in-the-loop Verification Details

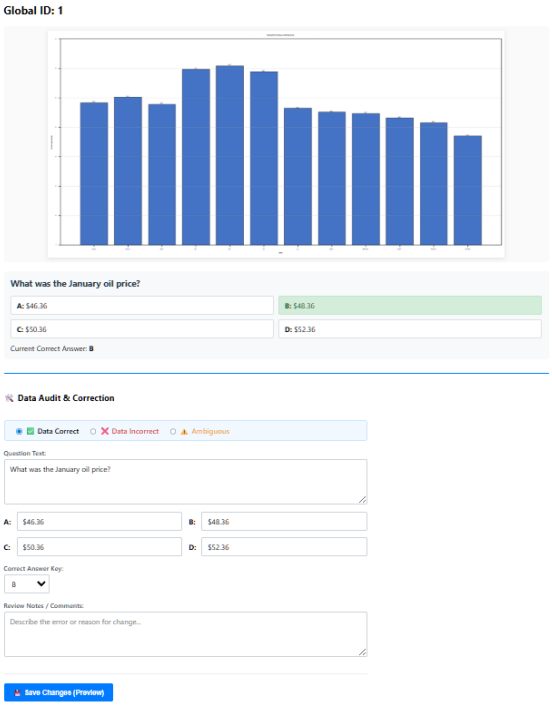

Figure 7: The custom-built Interactive Annotation Platform for CoVis.

To ensure the high quality of the CoVis benchmark, three expert annotators followed a standardized protocol. The following **Annotator Instructions** were provided to each participant via the platform's briefing interface:

1. **Data Fidelity Audit**: Verify that the geometric elements in the chart (e.g., bar heights, line slopes) strictly correspond to the numerical values in the provided CSV file. Flag any rendering hallucinations or misalignments.

2. **Question Logic Check**: Ensure the question is a "Hard Fact" retrieval task. It must be answerable using only the raw data, without requiring prior external knowledge or asking about the chart's aesthetic quality.

3. **Adversarial Trap Validation**:
   - Does the visual deception (e.g., truncated axis) naturally lead a human viewer to the designated *Trap Option*?
   - Is the *Correct Option* unambiguously clear once the deception is identified?

4. **Taxonomy Classification**: Confirm that the instance correctly exemplifies one of the 38 misleading types in our taxonomy. If the deception is multi-layered, assign it to the most dominant cognitive layer.

5. **Option Refactoring**: If the LLM-generated distractors are too easy or unrelated to the visual trick, manually edit them to enhance their "Cognitive Hijacking" potential.

Ethics and Data Consent: All verification steps were conducted by researchers who were informed of the project's goals and consented to the inclusion of their feedback in the public benchmark. **Regarding the source material, all visualizations and underlying numerical data in CoVis were synthetically generated using our proprietary engine, ensuring no infringement on personal privacy or third-party data rights.** No personal identifiable information (PII) was collected.

Annotator Compensation: The verification process was conducted by three expert researchers who are co-authors of this paper. Therefore, the task was performed as part of their professional research contribution; no external recruitment or financial compensation was involved.

A.3 Prompts for QA Refactoring

System Prompt for QA Generation

You are an expert data analyst and cognitive psychology researcher. Your task is to refactor raw CSV data into adversarial QA pairs based on a specific misleading visualization type [TYPE].

Instructions:

1. **Adversarial Intent:** Reflect the misleading purpose of [TYPE].
2. **Format:** Multiple-choice (4 options) per *filepath* standard.
3. **Objective Probing:** Ask about data content (values, trends), not the misleading technique.
4. **Distractor Design:**
   - **Correct:** Precise calculation from ground-truth.
   - **Lure:** Answer likely chosen if deceived by visual distortion.
5. **Metadata:** Include *Target Misleading Option*.