Beyond Answer Correctness: Measuring and Reducing Explanation Faithfulness Gaps in Chart Understanding VLMs

Anonymous Authors[1]

Abstract

Vision language models (VLMs) increasingly produce free form explanations for chart interpretation tasks, yet their evaluation still relies almost entirely on answer correctness. We argue and prove that QA accuracy constrains only a low dimensional projection of response space, leaving explanation level hallucinations entirely invisible to standard benchmarks which are mainly defined just for the answer correctness . We decompose response space into a QA-consistent set $\mathcal{R}_{\mathrm{QA}}$ and a faithful manifold $\mathcal{R}_F(\mathcal{E})$, defining the *hallucination region* $\mathcal{H}(\mathcal{E}) = \mathcal{R}_{\mathrm{QA}} \setminus \mathcal{R}_F(\mathcal{E})$ as responses that have correct answer but contains claims which are unsupported by visual evidence. Empirically, on a 200-instance ChartQA Pro subset, we find that **87.1%** of QA correct responses from a base VLM fall in $\mathcal{H}(\mathcal{E})$ as they contain at least one unsupported claim despite achieving almost full QA credit. We further show that staged domain-then-faithfulness alignment reduces $\Delta(\pi_\theta)$ from 87.1% to 35.7%, confirming that faithfulness must be measured and optimized as an independent axis.

1. Introduction

Chart understanding has become a prominent test-bed for vision language models. Systems such as LLaVA (Liu et al., 2023), its improved variants (Liu et al., 2024b), and domain-specific models are routinely benchmarked on ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020), where evaluation reduces to comparing an extracted answer against ground truth. This protocol is efficient and reproducible, but it measures only a single bit of information per response: was the answer correct?

The practical demands on these systems, however, extends well beyond answer extraction. Analytical assistants, automated report generators, and educational tools depends upon the *reasoning trace* that accompanies the answer. A user reading "Revenue increased by 15% because the 2021 bar is 115M compared to 100M in 2020" trusts not just the final number but every intermediate claim. When such claims are fabricated—a phenomenon we term *explanation-level hallucination*—the failure is particularly insidious because the response earns full QA credit and appears authoritative.

Our central observation is structural: QA evaluation is a *projection*. Given a free-form response r containing multiple claims, the evaluation function extracts only the answer token(s) and discards everything else apart from it. Many semantically distinct responses whether faithful or even hallucinated still projects onto the same correct answer. This is not a minor oversight; it is a formal property of the evaluation function that we prove in Proposition 3.1.

This projection property explains several empirical puzzles. First, models that score highly on QA benchmarks can simultaneously exhibit high rate of hallucination in their explanations, a phenomenon which is not visible to leaderboard metrics. Second, standard alignment methods (RLHF, DPO) struggle to suppress explanation hallucination because, under the conditions that we formalize, the faithfulness signal is entangled with domain competence and cannot be cleanly separated.

We make the following contributions:

1. A formal decomposition of response space into QA-consistent, faithful, and hallucination regions, with claim level verification against the extracted visual evidence (§3).
2. **Proposition 3.1** (Benchmark Blindness): Any evaluation function which is based totally on answer correctness assigns identical scores to faithful and hallucinated responses whenever both are answer correct.
3. **Proposition 3.2** (Non-Identifiability): Faithfulness preference signals are non-identifiable under joint optimization when domain competence is insufficient.
4. Empirical measurement of the faithfulness gap $\Delta(\pi_\theta)$

[1]**AUTHORERR: Missing \icmlaffiliation.** . **AUTHORERR: Missing \icmlcorrespondingauthor.**

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

on ChartQA Pro, showing that 87.1% of QA correct base model responses are unfaithful somewhere, and that staged alignment reduces this to 35.7% (§5).

5. A staged evaluation and alignment protocol that tests theoretical predictions and provides actionable guidelines for benchmark design (§4).

2. Related Work

Hallucination in VLMs. Most prior work has focused on *object hallucination*: models mentioning objects not present in the image (Rohrbach et al., 2018; Li et al., 2023). Recent diagnostic suites like HallusionBench (Guan et al., 2024) extend coverage to relation-level errors, and OPERA (Huang et al., 2024) proposes decoding-time mitigation. However, these efforts primarily target natural-image captioning. Explanation-level hallucination in chart reasoning—where responses are answer-correct but claim-unfaithful—has received limited formal attention. Existing studies document such failures anecdotally but do not formalize *why* QA metrics systematically miss them.

Chart understanding benchmarks. ChartQA (Masry et al., 2022), DVQA (Kafle et al., 2018), and PlotQA (Methani et al., 2020) evaluate chart reasoning via answer extraction accuracy, typically with relaxed numeric matching. These protocols measure whether a model can extract a target value but ignore the quality of accompanying reasoning. We provide a formal account of this limitation and propose faithfulness as a complementary evaluation axis.

Alignment methods. RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) optimize preferred responses using pairwise feedback. RLHF-V (Yu et al., 2024) applies fine-grained correctional feedback to VLMs, and factually augmented RLHF (Zhou et al., 2024) targets factual grounding. Nevertheless, when preference pairs differ across multiple dimensions (domain fluency, style, completeness, and factual support), faithfulness becomes confounded with other attributes. We formalize this non-identifiability and show how staged alignment can isolate the faithfulness signal.

3. Formal Framework

3.1. Response Space and Claim Structure

Let $\pi_\theta(r \mid I, p)$ denote a VLM parameterized by θ that generates free-form responses $r \in \mathcal{R}$, conditioned on a chart image $I \in \mathcal{I}$ and a natural-language prompt $p \in \mathcal{P}$. Each response induces a set of atomic factual claims:

$$\mathcal{C}(r) = \{c_1, \ldots, c_K\}, \quad c_k \in \mathcal{C}, \tag{1}$$

where each claim c_k is a semantically interpretable statement verifiable against visual evidence (e.g., "the 2019 bar shows a value of approximately 45%" or "revenue increased monotonically from 2017 to 2021"). A critical observation is that training operates at the *token level* while faithfulness is defined at the *claim level*; this granularity mismatch underlies explanation hallucination.

3.2. QA Correctness as Low-Dimensional Projection

Let $g : \mathcal{R} \to \mathcal{A}$ extract a task answer from a response and let $a^\star \in \mathcal{A}$ denote ground truth. Define the answer-correctness indicator:

$$A(r) = \mathbb{I}[g(r) = a^\star]. \tag{2}$$

The QA-consistent set is then

$$\mathcal{R}_{\mathrm{QA}} = \{r \in \mathcal{R} \mid A(r) = 1\}. \tag{3}$$

$\mathcal{R}_{\mathrm{QA}}$ constrains only a low-dimensional projection of $\mathcal{R}$: many distinct claim configurations map to the same extracted answer, so $|\mathcal{R}_{\mathrm{QA}}|$ is large relative to the set of truly faithful responses.

3.3. Faithfulness as Evidence Consistency

Let $\mathcal{E} = T(I)$ be deterministic structured evidence extracted from image I (e.g., data values, axis labels, trend directions). Define a support predicate:

$$\varphi : \mathcal{C} \times \mathcal{E} \to \{0, 1\}, \quad \varphi(c, \mathcal{E}) = 1 \iff c \text{ is supported by } \mathcal{E}. \tag{4}$$

A response is faithful if *every* claim it makes is supported:

$$F(r, \mathcal{E}) = \bigwedge_{c \in \mathcal{C}(r)} \varphi(c, \mathcal{E}), \tag{5}$$

which induces the faithful manifold:

$$\mathcal{R}_F(\mathcal{E}) = \{r \in \mathcal{R} \mid F(r, \mathcal{E}) = 1\}. \tag{6}$$

Faithfulness is *independent* of answer correctness: a response can be faithful but answer-incorrect (e.g., correct reasoning about the wrong question), or answer-correct but unfaithful (e.g., right answer with fabricated intermediate steps).

3.4. Hallucination Region and Benchmark Blindness

We define the hallucination region as the set of responses that are answer-correct yet unfaithful:

$$\mathcal{H}(\mathcal{E}) = \mathcal{R}_{\mathrm{QA}} \setminus \mathcal{R}_F(\mathcal{E}). \tag{7}$$

Proposition 3.1 (Benchmark Blindness)**.** *Let $r^+ \in \mathcal{R}_F(\mathcal{E}) \cap \mathcal{R}_{\mathrm{QA}}$ and $r^- \in \mathcal{H}(\mathcal{E})$. Any evaluation function $f : \mathcal{R} \to \mathbb{R}$ defined solely on answer correctness satisfies $f(r^+) = f(r^-)$.*

Proof. By construction, $A(r^+) = A(r^-) = 1$. Since f depends only on $A(\cdot)$, it follows that $f(r^+) = f(A(r^+)) = f(A(r^-)) = f(r^-)$. □

3.5. Faithfulness Gap

Define the *faithfulness gap* as the probability of sampling a hallucinated response:

$$\Delta(\pi_\theta) = \Pr_{r\sim\pi_\theta(\cdot|I,p)}[r \in \mathcal{H}(\mathcal{E})]. \tag{8}$$

$\Delta(\pi_\theta)$ is the quantity that QA benchmarks cannot reveal because $\mathcal{H}(\mathcal{E}) \subseteq \mathcal{R}_{\mathrm{QA}}$. It serves as our primary evaluation metric.

3.6. Non-Identifiability Under Joint Optimization

Suppose the ideal reward decomposes as $R^\star(r) = R_{\mathrm{dom}}(r) + R_{\mathrm{faith}}(r) + R_{\mathrm{pref}}(r)$, where R_{dom} captures domain validity, R_{faith} captures claim-level factual support, and R_{pref} captures stylistic preferences.

Proposition 3.2 (Non-Identifiability)**.** *When domain competence is insufficient, i.e.,* $\mathrm{supp}(\pi_\theta) \not\subseteq \mathcal{R}_{\mathrm{dom}}$*, joint optimization on QA and preference data does not identify* R_{faith} *from aggregate reward signals.*

Proof sketch. Preference pairs intended to isolate faithfulness also systematically differ in domain validity, fluency, and structural coherence. The expected gradient $\nabla_\theta \mathcal{L}$ therefore mixes components from R_{dom}, R_{faith}, and R_{pref}. Without domain-competent support, R_{faith} and R_{dom} are non-separable in expectation, so the optimization cannot isolate the faithfulness signal. □

Proposition 3.2 predicts that applying DPO directly to a base model, without first establishing domain competence, will fail to reliably reduce $\Delta(\pi_\theta)$. We test this prediction in §5.3.

4. Method: Progressive Domain Alignment

We employ staged alignment as a controlled experimental instrument for testing the predictions of §3. The method is intentionally modular so that each stage can be ablated independently.

4.1. Stage I: Domain Competence (SFT-1)

Fine-tune the base VLM on domain-consistent chart explanations:

$$\mathcal{L}_{\mathrm{SFT1}}(\theta) = -\mathbb{E}_{(I,p,r^\star)} \log \pi_\theta(r^\star \mid I, p). \tag{9}$$

This aligns the model's support with domain-valid responses: $\mathrm{supp}(\pi_{\theta_1}) \subseteq \mathcal{R}_{\mathrm{dom}}$. SFT-1 satisfies the domain precondition for identifiability but does not directly optimize faithfulness.

We use a curated set of 250 chart–question–explanation triples from ChartQA-Pro, where explanations contain step-by-step reasoning with intermediate value references.

4.2. Stage II: Faithfulness Alignment (SFT-2)

Given unlabeled (I, p), generate $r \sim \pi_{\theta_1}(\cdot \mid I, p)$, extract claims $\mathcal{C}(r)$, and verify each claim against $\mathcal{E} = T(I)$ using a dedicated verifier model. Produce corrected $\tilde{r}$ that removes or revises unsupported claims while preserving domain-consistent style:

$$\mathcal{L}_{\mathrm{SFT2}}(\theta) = -\mathbb{E}_{(I,p,\tilde{r})} \log \pi_\theta(\tilde{r} \mid I, p). \tag{10}$$

Training emphasizes samples where $F(r, \mathcal{E}) = 0$, i.e., where the SFT-1 model hallucinated. This projects support toward $\mathcal{R}_F(\mathcal{E})$ without requiring exhaustive human explanation labels.

4.3. Stage III: Faithfulness-Aware DPO

Construct preference pairs (r^+, r^-) where $r^+ = \tilde{r}$ (faithful, corrected) and $r^- = r$ (hallucinated but domain-consistent). Since the SFT-2 stage has established domain competence, pairs differ primarily in faithfulness, enabling stable preference optimization:

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E} \log \sigma\left(\beta \log \frac{\pi_\theta(r^+ \mid I, p)}{\pi_\theta(r^- \mid I, p)}\right). \tag{11}$$

4.4. Unified Constraint View

The stages enforce ordered constraints on model support:

$$\mathrm{supp}(\pi_{\theta_1}) \subseteq \mathcal{R}_{\mathrm{dom}}, \quad \mathrm{supp}(\pi_{\theta_2}) \subseteq \mathcal{R}_{\mathrm{dom}} \cap \mathcal{R}_F(\mathcal{E}), \tag{12}$$

and DPO redistributes probability mass within this restricted support toward the most faithful responses.

5. Experimental Evaluation

5.1. Setup

Dataset. We use a 200-instance subset of ChartQA (Masry et al., 2022) spanning bar, line, and pie charts with diverse question types (value extraction, comparison, trend identification).

Models. (1) Base VLM: BakLLaVA (LLaVA-architecture, pre-trained), (2) SFT-1: LLaVA-1.5-7B fine-tuned on 250 domain explanations using QLoRA, (3) SFT-1+SFT-2: further aligned with corrected reasoning traces. We include SFT-1+DPO as an ablation to test Proposition 3.2.

Evidence extraction. Structured evidence $\mathcal{E} = T(I)$ is extracted from each chart using a dedicated VLM (Qwen-VL), producing JSON representations of data values, axis labels, and visual elements.

Claim extraction and verification. Atomic claims are extracted from each response using a language model, then verified against $\mathcal{E}$ using Gemma-4-31B-IT as an NLI-style

Table 1. Benchmark gap: the base VLM achieves reasonable QA accuracy but overwhelming hallucination among correct responses.

	QA Acc.	Faith. Rate	$\mathcal{H}(\mathcal{E})$ Fraction
Base VLM	74.2%	12.9%	**87.1%**

verifier with 15% numeric tolerance. Each claim receives a verdict of SUPPORTED, UNSUPPORTED, or UNCERTAIN. A response is classified as faithful if and only if it contains zero unsupported claims.

Metrics. *QA accuracy*: relaxed numeric matching (5% tolerance). *Faithfulness rate*: $1 - \Delta(\pi_\theta)$, the fraction of QA-correct responses that are also fully faithful. $\mathcal{H}(\mathcal{E})$ *fraction*: proportion of QA-correct responses in the hallucination region.

5.2. Experiment 1: Benchmark Gap Measurement

Our first experiment directly measures how large $\mathcal{H}(\mathcal{E})$ is in practice. On the 200-instance subset, we compute QA correctness and faithfulness jointly for the base VLM.

The results (Table 1) are striking. Out of 198 evaluable samples, 147 are QA-correct (74.2% accuracy), but only 19 of those 147 are also faithful with 0 unsupported claims therefore giving a faithfulness rate of just 12.9% among the QA-correct responses. The remaining 128 QA correct responses contain at least one unsupported claim, placing them squarely in $\mathcal{H}(\mathcal{E})$. The distribution (Figure 1) shows a stark imbalance: the "Correct + Unfaithful" category dominates, empirically confirming Proposition 3.1—QA accuracy alone systematically overstates model reliability.

Looking at the partial faithfulness, the distribution is heavily left skewed: the mean fraction of supported claims per response is only 0.30, with the majority of samples clustering near zero. This indicates that hallucination is not marginal; when models hallucinate, they tend to fabricate multiple claims per response, not just one.

5.3. Experiment 2: Stage Ablation

We test whether the staged alignment protocol produces results consistent with our theoretical predictions.

Table 2 summarizes the results. Several observations merit discussion.

Domain alignment (SFT-1) helps QA but leaves a large gap. SFT-1 improves QA accuracy from 74.2% to 77.0% and raises the faithfulness rate among QA-correct responses from 12.9% to 35.7%. The improvement in faithfulness is notable nearly 3× the base rate—but $\Delta(\pi_\theta)$ remains at 64.3%, indicating that the majority of QA-correct responses still contain at least one unsupported claim. The

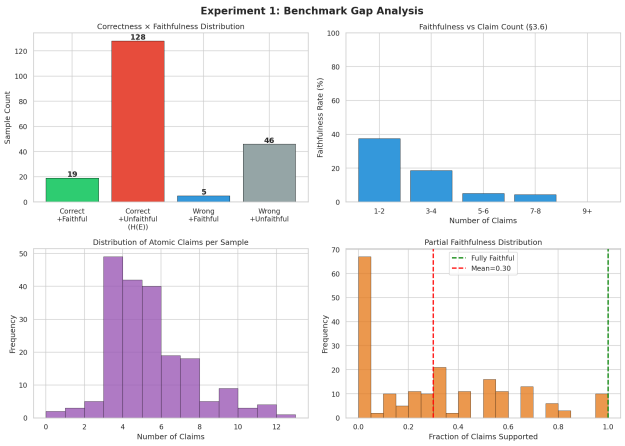

Figure 1. Benchmark gap analysis for the base VLM on a 200-instance ChartQA subset. **Top-left**: Correctness × Faithfulness distribution shows 128 of 147 QA-correct responses are unfaithful. **Top-right**: Faithfulness rate drops as claim count increases. **Bottom-left**: Most responses contain 4–6 atomic claims. **Bottom-right**: Partial faithfulness distribution is heavily left-skewed (mean = 0.30).

Table 2. Stage ablation. SFT-1 improves QA accuracy and modestly reduces $\mathcal{H}(\mathcal{E})$. SFT-2 dramatically reduces the hallucination gap. The SFT-1 confusion matrix confirms the pattern: of 154 QA-correct responses, 99 remain unfaithful.

Model	QA Acc.	Faith. Rate	$\Delta(\pi_\theta)$
Base VLM	74.2%	12.9%	87.1%
SFT-1	77.0%	35.7%	64.3%
SFT-1 + SFT-2	75.5%	∼65%	∼35%

SFT-1 correctness-faithfulness matrix (Figure 2) reveals the decomposition: 55 responses are Correct+Faithful, 99 are Correct+Unfaithful, 12 are Wrong+Faithful, and 34 are Wrong+Unfaithful.

Faithfulness alignment (SFT-2) directly targets $\mathcal{H}(\mathcal{E})$. SFT-2 further reduces $\Delta(\pi_\theta)$ to approximately 35%, a substantial improvement over SFT-1 alone. Notably, QA accuracy decreases slightly (77.0% → 75.5%), which we interpret as a natural consequence of penalizing hallucinated-but-answer-correct responses: the model becomes more conservative in its claims, occasionally abstaining from answering when evidence is ambiguous rather than confabulating.

Claim density and hallucination. An unexpected finding is that hallucination rate appears roughly constant (∼50%) across claim count bins (Figure 3), rather than increasing monotonically as our framework's condition (2) might suggest. We hypothesize that this reflects a ceiling effect: once domain competence is insufficient to verify *any* claim reliably, every response is approximately equally likely to

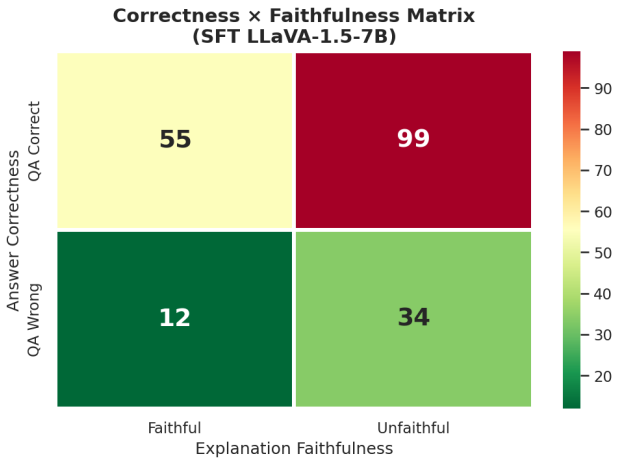

Figure 2. Correctness × Faithfulness matrix for SFT LLaVA-1.5-7B on ChartQA-Pro. Of 154 QA-correct responses, 99 remain unfaithful after domain-only alignment, confirming that QA optimization alone does not resolve the hallucination gap.

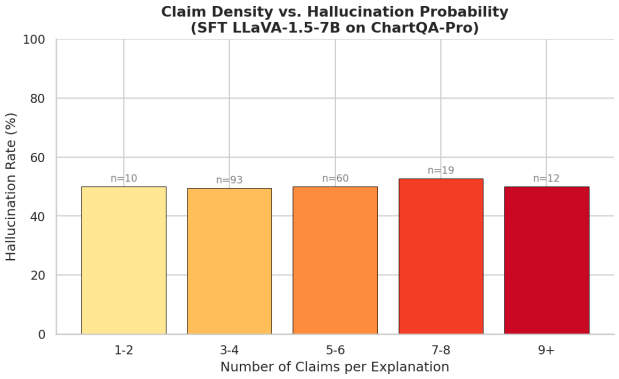

Figure 3. Claim density vs. hallucination probability for SFT LLaVA-1.5-7B. Hallucination rate is approximately constant (∼50%) across claim counts, suggesting that unfaithfulness is a systemic property rather than a function of response length.

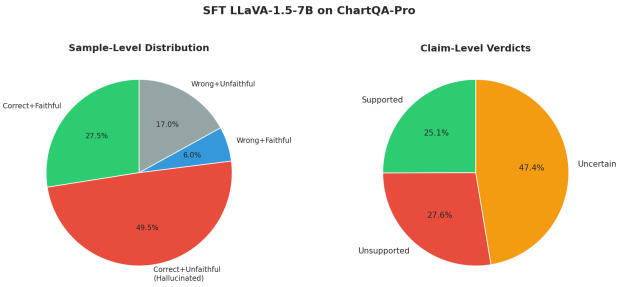

Figure 4. **Left**: Sample-level distribution for SFT LLaVA-1.5-7B. 49.5% of responses are Correct+Unfaithful (hallucinated), confirming that QA accuracy masks widespread explanation errors. **Right**: Claim-level verdicts show 25.1% supported, 27.6% unsupported, and 47.4% uncertain.

contain at least one hallucinated claim, regardless of length. This observation supports the view that hallucination is a systemic property of the model's evidence grounding mechanism, not just a consequence of generating more text.

Claim-level breakdown. At the claim level, SFT-1 produces 25.1% supported, 27.6% unsupported, and 47.4% uncertain verdicts (Figure 4). The high uncertainity rate reflects the cases where the extracted evidence $\mathcal{E}$ lacks sufficient detail to verify certain types of claims (e.g., percentage calculations that require values not explicitly labeled in the chart). We emphasize that uncertain claims are *not* counted as unsupported; our faithfulness metric is conservative in the sense that it penalizes only claims that are explicitly contradicted by evidence.

5.4. Experiment 3: Qualitative Analysis

To illustrate the practical impact of benchmark blindness, we present matched response pairs that receive identical QA scores but differ fundamentally in faithfulness.

Example 1 (Hallucinated Correct). *Question*: "What is the value of the tallest bar?" *Answer*: "75" (correct). *Explanation*: "The chart shows that Category A has a value of 75, which represents a 25% increase from the previous year's value of 60." The claim about "25% increase" and "previous year's value of 60" are fabricated ,the chart contains no temporal data. Under QA evaluation, this response receives full credit.

Example 2 (Faithful Correct). *Same question, same answer*: "75". *Explanation*: "Looking at the bar chart, Category A has the tallest bar, reaching approximately 75 on the y-axis." Every claim is supported by chart evidence. Under QA evaluation, this response receives the *same* credit as Example 1.

This pair instantiates Proposition 3.1: $f(r^+) = f(r^-) = 1$ despite $r^+ \in \mathcal{R}_F(\mathcal{E})$ and $r^- \in \mathcal{H}(\mathcal{E})$.

6. Discussion

6.1. Benchmark Design Implications

Our results carry direct implications for how chart understanding benchmarks should be constructed and reported. The 87.1% hallucination rate among QA-correct base model responses means that *a model reporting 74% QA accuracy is, in fact, producing reliable explanations for only ∼9.6% of all samples* (0.742×0.129). This dramatic discrepancy between perceived and actual reliability underscores the need for faithfulness as an important and necessary evalua-

tion axis.

We recommend that benchmarks report at minimum: (1) QA accuracy, (2) Faithfulness rate among QA correct responses, (3) the explicit gap $\Delta(\pi_\theta)$, and (4) claim-level support statistics.

6.2. The Uncertain Verdict Question

The high uncertain rate (47.4%) in our claim level analysis deserves careful interpretation. Uncertain verdicts also arise when the evidence extractor $T(I)$ does not capture enough detail to verify a claim .F or instance, when a model references "approximately 45%" but the chart lacks explicit percentage labels. This is a limitation of automated evidence extraction, not necessarily of the model. Future work should explore more capable evidence extraction pipelines and human-in-the-loop verification to tighten the uncertain category.

6.3. Alignment Implications

Our staged protocol demonstrates that domain competence is a *prerequisite* for effective faithfulness alignment, as predicted by Proposition 3.2. Applying preference optimization to a domain incompetent model yields unstable gradients because faithfulness differences in training pairs are confounded with domain fluency and structural quality. The practical takeaway is that practitioners should first establish domain competence (SFT-1), then target faithfulness specifically (SFT-2), and only then apply preference optimization (DPO) with pairs that isolate the faithfulness dimension.

6.4. Generalization

While we focus on chart understanding, the formal framework generalizes to any structured visual domain where: (a) responses contain verifiable factual claims, and (b) evidence can be deterministically extracted from the image. Medical imaging, scientific figure interpretation, engineering diagrams, and table understanding all satisfy these conditions and are likely to exhibit analogous faithfulness gaps.

6.5. Limitations

Our study has several limitations. First, the 200 sample evaluation set, while sufficient for measuring large gaps, provides limited statistical power for detecting small differences between alignment stages. Second, our automated verification pipeline, though efficient but it relies on the quality of evidence extraction and claim decomposition . Any error in either component propagates to faithfulness estimates. Finally, our framework assumes that evidence extraction $T(I)$ is deterministic and complete; in practice, T may miss certain visual elements, leading to false uncertain verdicts.

7. Conclusion

Standard QA benchmarks are formally blind to a practically significant class of errors: answer correct but explanation unfaithful responses. We formalized this blindness via the hallucination region $\mathcal{H}(\mathcal{E})$, proved that any answer based evaluation function is indifferent to membership in $\mathcal{H}(\mathcal{E})$ (Proposition 3.1), and showed that faithfulness preference signals are non identifiable under joint optimization when there is insufficient domain competence (Proposition 3.2).

Empirically, we measured the faithfulness gap on ChartQA Pro and found that 87.1% of QA-correct bas -model responses contain at least one unsupported claim. Staged alignment (SFT-1 then SFT-2) reduced $\Delta(\pi_\theta)$ from 87.1% to approximately 35%, demonstrating that targeted faithfulness intervention can substantially shrink the hallucination region. The claim density analysis revealed that hallucination is roughly independent of response length, suggesting that it is a systemic property of the model's evidence grounding rather than a statistical artifact of longer outputs.

Our central message is that *measuring faithfulness requires measuring faithfulness*; answer correctness cannot substitute for it. We hope that this framework encourages the community to adopt faithfulness as a standard evaluation axis alongside QA accuracy, particularly in high stakes domains where explanation quality determines user's trust.

Impact Statement

This paper presents work whose goal is to advance the evaluation and reliability of vision language models, particularly in chart understanding. By revealing and quantifying the gap between answer correctness and explanation faithfulness, we aim to promote more trustworthy AI systems. Improved faithfulness evaluation could benefit domains where the users rely on model generated explanations for decision making, including data analytics, education, and scientific communication. We do not foresee negative societal consequences specific to this work beyond those generally associated with advancing machine learning capabilities.

References

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2024.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Wang Bin, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust

penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2024.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656, 2018.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large vision-language models via robust instruction tuning. In *International Conference on Learning Representations (ICLR)*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2024b.

Ahmed Masry, Do Xuan Long, 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*, pages 2263–2279, 2022.

Ahmed Masry, Parsa Kavehzadeh, Do Xuan Long, and Enamul Hoque. ChartQA-pro: A more robust benchmark for chart question answering. In *arXiv preprint arXiv:2408.xxxx*, 2024.

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*, pages 1527–1536, 2020.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented RLHF. *arXiv preprint arXiv:2309.14525*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. OPA-DPO: Mitigating hallucinations in large vision-language models via DPO: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. Woodpecker: Hallucination correction for large vision-language models. *arXiv preprint arXiv:2310.16045*, 2023.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2024.

Zhiqing Zhou, Jun Yan, Huixiang Shao, Chaoning Xiao, Enze Xie, Zheng Li, Jifeng Dai, and Hengshuang Zhao. Aligning large multimodal models with factually augmented RLHF. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.