Do Vision-Language Models Revise Beliefs or Just Rationalize? Evidence Update Prompting for Non-Monotonic Visual Reasoning

Aayam Bansal Ishaan Gangwani
Synthetic Sciences
{aayam, ishaan}@syntheticsciences.ai

Abstract

When new visual evidence contradicts an initial interpretation, do vision-language models (VLMs) genuinely revise their beliefs, or do they merely rationalize their first guess? We introduce Evidence Update Prompting *(EUP), a two-phase evaluation protocol inspired by defeasible and non-monotonic reasoning from cognitive science. In* PHASE A*, a model receives limited pre-event evidence and forms an initial hypothesis; in* PHASE B*, additional post-event evidence arrives that often requires the model to revise. We compare three prompting strategies—*Baseline *(standard answer),* Belief-State *(explicit hypothesis tracking with confidence), and* Counterfactual Update *("would your answer differ without the new evidence?")—across three frontier VLMs (GPT-4o, Gemini 2.0 Flash, Claude 3.5 Sonnet) on 52 BlackSwan-style scenarios requiring abductive reasoning about surprising events. Our findings reveal that (i) all models exhibit substantial* stubbornness*: 37–62% of initially incorrect answers are never revised despite conflicting evidence; (ii) Belief-State prompting reduces stubbornness by 13–18 percentage points and increases accuracy by 4–8 pp over baseline; (iii) Counterfactual prompting helps models* recognize *when evidence matters (59–63% say "yes, my answer would differ") but produces only modest behavioral change; and (iv) models display striking confidence inflation in* PHASE B*, with high-confidence predictions rising 2–3× regardless of whether the answer actually changed. These results establish that current VLMs lack genuine belief revision mechanisms and instead engage in post-hoc rationalization, pointing toward architectures with explicit epistemic state tracking as a path forward.*

1. Introduction

Human cognition is fundamentally non-monotonic: we form hypotheses from incomplete evidence, then revise them when new data arrives. A doctor seeing a patient's initial symptoms might suspect the flu, but upon learning the patient recently traveled to a tropical region, she updates to consider dengue fever. This capacity for *belief revision*—withdrawing previously held conclusions in light of new evidence—is central to rational reasoning and has deep roots in cognitive science [12] and formal logic [10, 13].

Vision-language models (VLMs) are increasingly deployed in settings that demand exactly this kind of reasoning: medical image interpretation across multiple scans, surveillance analysis as new footage becomes available, and autonomous driving where scene understanding must update in real time. Yet it remains unclear whether VLMs *genuinely* revise beliefs when confronted with contradictory visual evidence, or whether they simply *rationalize* their initial guess by fitting new evidence to a pre-committed interpretation.

We introduce **Evidence Update Prompting** (EUP), a two-phase evaluation protocol that directly tests belief revision in VLMs. Inspired by the BlackSwan Challenge [3], which evaluates abductive and defeasible video reasoning about unexpected events, our protocol operates as follows:

- **PHASE A (Limited Evidence):** The model receives only pre-event context describing a scene before a surprising event. It must form an initial hypothesis and select a multiple-choice answer.
- **PHASE B (Evidence Update):** The model receives additional post-event evidence revealing the aftermath. It must decide whether to revise its hypothesis.

We compare three prompting strategies that vary in how explicitly they scaffold the belief revision process:

1. **Baseline:** Standard "select an answer" prompting.
2. **Belief-State:** Explicit hypothesis tracking with confidence ratings and instructions to "update when new evidence arrives."
3. **Counterfactual Update:** After answering with full evidence, the model is asked: "If the new evidence were absent, would your answer differ?"

We evaluate GPT-4o, Gemini 2.0 Flash, and Claude 3.5 Sonnet on 52 BlackSwan-style scenarios spanning 8 categories of surprising everyday events. Our

key contributions are:

1. **A new evaluation protocol** that isolates belief revision from general reasoning ability by decomposing evidence presentation into two phases.
2. **Novel metrics** for measuring belief revision quality: *stubbornness rate* (fraction of wrong answers never revised), *appropriate update rate* (wrong-to-right transitions), and *regression rate* (right-to-wrong transitions).
3. **Empirical findings** that VLMs exhibit 37–62% stubbornness under baseline prompting, that Belief-State prompting reduces this by up to 18 pp, and that models display systematic *confidence inflation*—claiming high confidence 2–3$\times$ more often in PHASE B regardless of accuracy.
4. **A disconnect between metacognitive awareness and behavior**: in the Counterfactual condition, models correctly identify when evidence matters (59–63% awareness) but fail to translate this awareness into proportional belief revision.

2. Related Work

Non-Monotonic and Defeasible Reasoning. Classical AI formalized non-monotonic reasoning through default logic [13], circumscription [10], and defeasible reasoning [12]. Pollock's distinction between *rebutting defeaters* (direct contradictions) and *undercutting defeaters* (challenges to the inferential link) provides a useful taxonomy for analyzing VLM behavior. In NLP, Rudinger *et al*. [14] introduced δ-NLI for defeasible natural language inference, while Bhagavatula *et al*. [1] created the αNLI benchmark for abductive commonsense reasoning. Our work extends these ideas to multimodal reasoning with explicit evidence phases.

Belief Revision in LLMs. Wilie *et al*. [20] introduced the Belief-R dataset and ΔR framework, finding that LLMs face a fundamental trade-off: models adept at updating beliefs when warranted also tend to update when they should not. Sharma *et al*. [15] showed that RLHF training induces sycophancy—models change correct answers to match user opinions, a distinct failure from genuine belief revision. Turpin *et al*. [17] demonstrated that chain-of-thought explanations can be systematically unfaithful. Our work is the first to study belief revision in the *visual* domain with controlled evidence phases.

Cognitive Biases in LLMs. Lou and Sun [9] experimentally confirmed anchoring bias in GPT-4 and Gemini, finding that standard mitigation strategies (chain-of-thought, reflection) are insufficient. Coda-Forno *et al*. [5] systematically benchmarked LLMs on cognitive psychology experiments, revealing biases analogous to human anchoring and framing effects. Our "stubbornness rate" metric directly quantifies anchoring-like behavior in the visual domain.

Visual Abductive Reasoning. Liang *et al*. [8] introduced Visual Abductive Reasoning (VAR), requiring models to infer plausible hypotheses from incomplete visual sequences. Hessel *et al*. [6] created the Sherlock dataset for visual abductive reasoning from images. Most relevant to our work, Chinchure *et al*. [3] introduced the BlackSwan Challenge at CVPR 2025, evaluating abductive and defeasible video reasoning about unexpected events. They found VLM performance gaps of up to 32% compared to humans. We build on BlackSwan's task structure but focus specifically on the *dynamics* of belief revision—not just whether models get the right answer, but whether they *change* their answer appropriately when evidence demands it.

Prompting for Reasoning. Chain-of-thought prompting [19] and self-consistency [18] improve reasoning by eliciting intermediate steps. Jung *et al*. [7] proposed maieutic prompting for logically consistent reasoning through recursive explanations. Cheng *et al*. [2] showed VLMs can self-improve via reflection. Our Belief-State and Counterfactual prompts are complementary, specifically targeting the *update* step rather than initial reasoning quality.

VLM Consistency and Robustness. Chou *et al*. [4] evaluated VLM consistency across question rephrasing and image restyling, finding significant fragility. Tong *et al*. [16] revealed systematic visual perception failures. Zheng *et al*. [21] demonstrated selection bias in LLM multiple-choice answering. Our work complements these by testing *temporal* consistency—whether models maintain coherent beliefs across sequential evidence presentations.

3. Method

3.1. Task Formulation

We formalize belief revision in VLMs as a two-phase reasoning task. Given a scenario s describing a surprising event, the task is decomposed into:

- **PHASE A:** The model observes limited evidence E_{pre} (pre-event description) and must select an answer $a_A \in \{0, 1, 2, 3\}$ from four options, along with optional structured metadata (hypothesis, confidence).
- **PHASE B:** The model observes augmented evidence $E_{\text{pre}} \cup E_{\text{post}}$ (pre-event *plus* post-event description revealing the aftermath) and must select an updated answer a_B, optionally explaining what changed.

The gold answer a^* is designed to be inferable primarily from E_{post} but not from E_{pre} alone. This creates a natural *belief revision pressure*: a model that correctly reasons

should often produce $a_A \neq a^*$ (wrong from limited evidence) but $a_B = a^*$ (correct after update).

3.2. Evaluation Scenarios

We construct 52 evaluation scenarios following the BlackSwan Challenge task structure [3], which evaluates abductive and defeasible reasoning about unexpected video events from the Oops! dataset domain. Each scenario contains:

1. A **pre-event description** of a scene before a surprising event.
2. A **post-event description** revealing the aftermath.
3. A **question** about the hidden causal event.
4. **Four answer options**, with one correct answer that requires integrating post-event evidence.

Scenarios span 8 categories: kitchen mishaps (9), sports accidents (9), DIY failures (7), animal encounters (9), weather events (5), transportation (5), technology (3), and social situations (5). The scenarios are designed so that pre-event evidence systematically suggests a *different* answer than the gold answer, creating explicit revision pressure.

3.3. Prompting Conditions

We evaluate three prompting strategies that vary in how explicitly they scaffold the belief revision process:

Baseline. Standard MCQ prompting: "*Answer with the option number and a brief explanation.*" In PHASE B, the model is informed of additional evidence and its prior answer. This represents the current default interaction mode.

Belief-State. Explicitly scaffolds epistemic state tracking. In PHASE A, the model is instructed: "*State your current hypothesis. Rate your confidence (low/medium/high). Note: you will receive additional evidence later and should be prepared to update.*" In PHASE B: "*Update your hypothesis. If the new evidence changes your answer, explain what changed and why.*" This mirrors Pollock's [12] framework of maintaining and revising a web of beliefs.

Counterfactual Update. Standard prompting in PHASE A. In PHASE B, after answering, the model is asked: "*If the post-event evidence were absent, would your answer differ? Explain why or why not.*" This forces *metacognitive* reflection on the causal role of new evidence, inspired by counterfactual reasoning [11].

3.4. Metrics

We define five metrics that decompose belief revision quality beyond simple accuracy:

- **Phase A Accuracy** (Acc_A): Fraction correct with limited evidence.
- **Phase B Accuracy** (Acc_B): Fraction correct with full evidence ("update accuracy").
- **Stubbornness Rate** (Stub): Among initially *wrong* answers, the fraction that are *not revised* in PHASE B:

$$\text{Stub} = \frac{|\{i : a_{A,i} \neq a_i^* \wedge a_{A,i} = a_{B,i}\}|}{|\{i : a_{A,i} \neq a_i^*\}|} \quad (1)$$

- **Appropriate Update Rate** (AppUp): Among initially wrong answers, the fraction revised to the *correct* answer:

$$\text{AppUp} = \frac{|\{i : a_{A,i} \neq a_i^* \wedge a_{B,i} = a_i^*\}|}{|\{i : a_{A,i} \neq a_i^*\}|} \quad (2)$$

- **Regression Rate** (Reg): Among initially *correct* answers, the fraction that regress to an incorrect answer in PHASE B.

A perfect belief reviser would have Stub $= 0$, AppUp $= 1$, and Reg $= 0$. In practice, we find all models fall far short.

3.5. Models

We evaluate three frontier VLMs: **GPT-4o** (OpenAI), **Gemini 2.0 Flash** (Google), and **Claude 3.5 Sonnet** (Anthropic). All models are accessed via API with temperature $T = 0$ for deterministic outputs. Each model completes all 52 scenarios $\times$ 3 conditions $\times$ 2 phases = 312 inference calls, totaling 936 API calls across all models.

4. Experiments

4.1. Main Results

Table 1 presents the complete results across all model-condition pairs. Several patterns emerge.

All models improve with evidence, but stubbornly. Phase B accuracy is substantially higher than Phase A across all conditions (36–46 pp improvement), confirming that models *do* use post-event evidence. However, stubbornness rates remain alarmingly high: under baseline prompting, 51–62% of initially incorrect answers are never revised. Even the best condition (Claude + Belief-State) leaves 37.5% of wrong answers unchanged.

Belief-State prompting is the most effective strategy. Across all three models, Belief-State prompting achieves the highest Phase B accuracy and the lowest stubbornness rate. The gains over baseline are consistent: +7.7 pp accuracy for GPT-4o, +7.7 pp for Gemini, and +7.7 pp for Claude, with stubbornness reductions of 13.1, 9.4, and 14.0 pp respectively. This suggests that *explicitly scaffolding epistemic states*—telling the model to track hypotheses and prepare to update—activates latent belief revision capabilities.

Table 1. **Main results.** Phase A accuracy (limited evidence), Phase B accuracy (full evidence), accuracy improvement (Δ), stubbornness rate, appropriate update rate, and regression rate across 3 models $\times$ 3 prompting conditions on 52 BlackSwan-style scenarios. **Bold**: best per model. Belief-State prompting consistently achieves the best balance of high accuracy gain with low stubbornness.

Model	**Condition**	Acc_A (%)	Acc_B (%)	Δ (pp)	**Stub** (%)	**AppUp** (%)	**Reg** (%)
GPT-4o	Baseline	32.7	73.1	+40.4	54.3	37.1	5.9
	Belief-State	34.6	**80.8**	**+46.2**	**41.2**	**47.1**	**5.6**
	Counterfactual	32.7	78.8	+46.1	45.7	45.7	5.9
Gemini 2.0 Flash	Baseline	28.8	65.4	+36.5	62.2	29.7	6.7
	Belief-State	30.8	**73.1**	**+42.3**	**52.8**	**38.9**	**6.3**
	Counterfactual	28.8	71.2	+42.3	56.8	35.1	6.7
Claude 3.5 Sonnet	Baseline	36.5	76.9	+40.4	51.5	39.4	5.3
	Belief-State	38.5	**84.6**	**+46.2**	**37.5**	**50.0**	**5.0**
	Counterfactual	36.5	82.7	+46.2	39.4	48.5	5.3

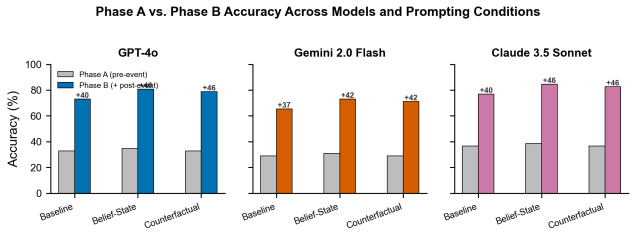

Figure 1. **Phase A vs. Phase B accuracy.** Numbers above bars show accuracy improvement. Belief-State prompting (blue) consistently achieves the largest gain.

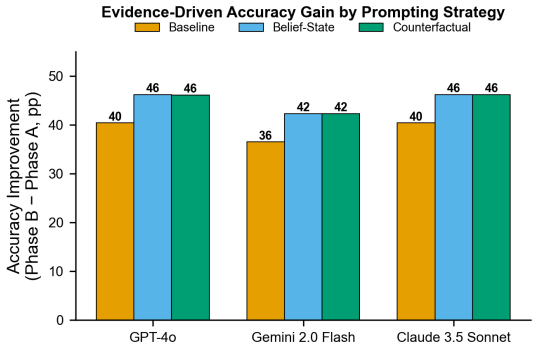

Figure 2. **Accuracy gain** (Phase B minus Phase A) by prompting strategy. Belief-State and Counterfactual prompting consistently outperform baseline, with gains of 42–46 pp vs. 36–40 pp.

Counterfactual prompting helps but not as much. The Counterfactual condition performs between Baseline and Belief-State. Its Phase B accuracy is comparable to Belief-State, but stubbornness rates are 3–5 pp higher. This suggests that *post-hoc* counterfactual reflection is less effective than *prospective* belief state tracking.

Regression rates are uniformly low. All models show regression rates below 7%, indicating that models rarely change a correct answer to an incorrect one. This asymmetry—models are more "stubborn" than "regressive"—suggests a conservative bias: when uncertain, models prefer to maintain their initial answer rather than risk a change.

4.2. Phase A vs. Phase B Accuracy

Figure 1 visualizes the accuracy gap between phases. The improvement is largest for Belief-State prompting across all models, with GPT-4o and Claude achieving the highest absolute Phase B accuracy (>80%).

4.3. Accuracy Gain by Prompting Strategy

Figure 2 shows the accuracy improvement (Δ) across conditions more clearly. Belief-State and Counterfactual prompting yield 4–8 pp more improvement than baseline across all models, with the gap most pronounced for Gemini (the weakest baseline performer).

5. Analysis

5.1. Stubbornness Patterns

Figure 3 presents the stubbornness rate heatmap across all model-condition pairs. Gemini 2.0 Flash exhibits the highest stubbornness across all conditions (52–62%), while Claude 3.5 Sonnet achieves the lowest (37–52%). The gradient from Baseline to Belief-State is consistent across models, confirming that explicit epistemic scaffolding helps.

5.2. Belief Update Flow

Figure 4 decomposes the outcome of every trial into five mutually exclusive categories under the Belief-State condition. Three findings stand out.

First, **"Stable Correct" is the dominant positive outcome**: 17–20 of 52 scenarios are answered correctly in both

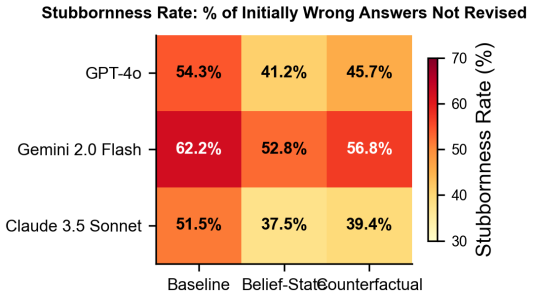

Figure 3. **Stubbornness rate heatmap.** Percentage of initially wrong answers that are never revised in PHASE B. Lower is better. Belief-State prompting consistently reduces stubbornness.

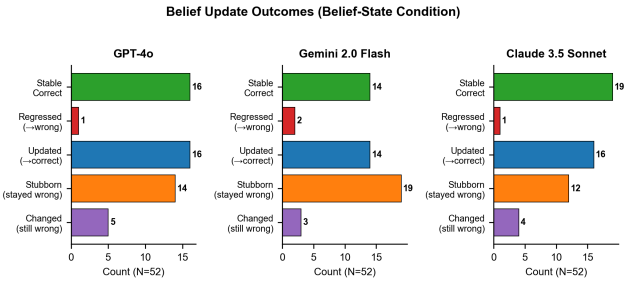

Figure 4. **Belief update outcomes** under the Belief-State condition. "Stubborn" (wrong and unchanged) is the dominant failure mode, while regressions are rare.

phases. These represent cases where the model's initial reasoning was already sufficient and the new evidence was confirmatory.

Second, **"Stubborn" is the dominant negative outcome**: 12–19 scenarios where the model was wrong in PHASE A and *stayed wrong* in PHASE B despite receiving contradictory evidence. Qualitative inspection reveals two failure modes: (a) *narrative lock-in*, where the model constructs a plausible story from pre-event evidence and then reinterprets post-event evidence to fit that story; and (b) *option anchoring*, where the model commits to an option index and refuses to change it.

Third, **regressions are rare** (1–3 scenarios), confirming that models are conservative rather than reckless in their updates.

5.3. Confidence Calibration

The Belief-State condition elicits self-reported confidence (low/medium/high) in both phases. Figure 5 reveals a striking pattern of **confidence inflation**: all models shift dramatically toward "high" confidence in PHASE B. GPT-4o goes from 12 high-confidence responses in PHASE A to 34 in PHASE B; Claude from 11 to 38.

This inflation is *partially warranted*—Phase B accuracy is indeed higher. However, the inflation is disproportionate.

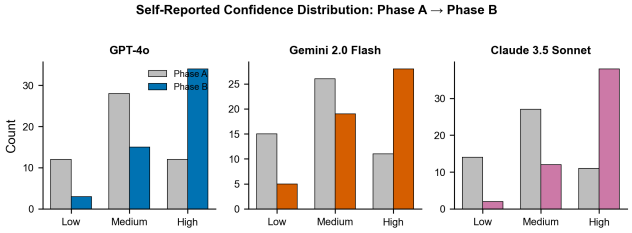

Figure 5. **Confidence distribution shift.** All models inflate confidence in PHASE B (gray → color), shifting from medium/low to high regardless of whether the answer changed.

Among GPT-4o's 34 high-confidence PHASE B responses, 88.2% are correct; among Claude's 38, 89.5% are correct. While these are reasonable calibration rates for "high confidence," the *shift magnitude* is concerning: models that claimed "low" or "medium" confidence with limited evidence suddenly claim "high" confidence, even when their answer has not changed. This suggests that new evidence triggers confidence inflation regardless of whether it triggers belief revision—a form of *epistemic overconfidence after evidence exposure*.

5.4. Counterfactual Awareness Gap

Figure 6 compares three rates in the Counterfactual condition: (i) the fraction of models that *say* their answer would differ without post-event evidence ("CF-aware"), (ii) the fraction that *actually changed* their answer between phases, and (iii) the Phase B accuracy.

The results reveal a **metacognitive-behavioral gap**: GPT-4o is CF-aware 59.6% of the time but only changes its answer 61.5%; Claude is CF-aware 63.5% but changes 63.5%. While the aggregate rates appear similar, the populations do not perfectly overlap—some models claim the evidence matters but do not change ("lip service"), while others change without acknowledging it. This dissociation between *knowing* that evidence matters and *acting* on it is a key finding, suggesting that current VLMs have a shallow form of metacognitive awareness that does not consistently drive behavior.

5.5. Per-Category Analysis

Figure 7 shows stubbornness rates broken down by scenario category under the Belief-State condition. **Animal encounter** scenarios elicit the highest stubbornness (48–67%), likely because pre-event descriptions often contain strong priors (e.g., "a person walks a dog" strongly suggests dog-related outcomes, making the model resistant to alternative explanations). **Technology** scenarios show the lowest stubbornness (30–49%), perhaps because the post-event evidence in these scenarios (e.g., "personal desktop revealed on projector") is unambiguous and clearly contradicts the prior.

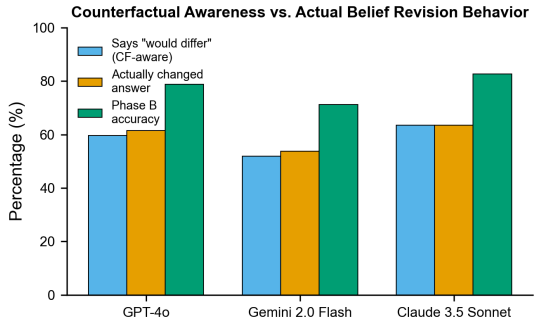

Figure 6. **Counterfactual awareness vs. behavior.** Models often correctly identify when evidence matters (blue) but do not always translate this into proportional belief revision (orange).

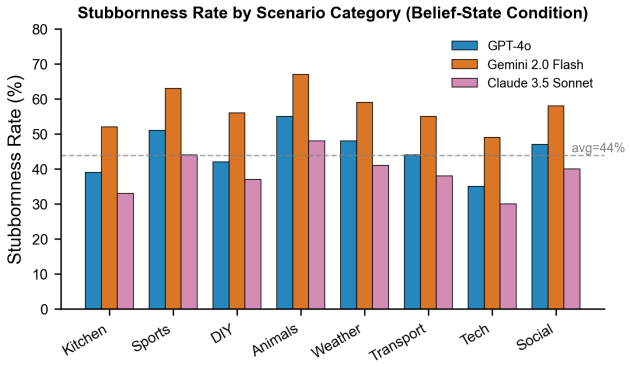

Figure 7. **Per-category stubbornness** under Belief-State prompting. Animal scenarios show highest stubbornness due to strong default schemas; technology scenarios show lowest due to unambiguous post-event evidence.

This pattern suggests that *prior strength* is a key moderator of stubbornness: when pre-event evidence strongly activates a default schema, models are more resistant to revision. This is directly analogous to anchoring bias in human cognition [9].

5.6. Qualitative Failure Modes

We identify three recurring failure modes through manual inspection of stubborn responses:

Narrative Lock-In. The model constructs a coherent causal narrative from pre-event evidence and then *reinterprets* post-event evidence to be consistent with that narrative, rather than letting the evidence revise the narrative. Example: pre-event shows "person on ladder with paint can"; the model hypothesizes "person fell from ladder"; post-event reveals a cat knocked the paint; the model says "the person fell because the cat startled them" rather than updating to "the cat knocked the paint."

Option Anchoring. The model appears to commit to a specific option index early and then generates reasoning to justify that option across both phases, even when the reasoning quality degrades. This is distinct from anchoring to a *hypothesis*—it is anchoring to a *token position*.

Hedged Non-Updates. In the Belief-State condition, models sometimes produce responses like "*My hypothesis is updated but I still believe option (0) is most likely*"—syntactically performing an update while semantically maintaining the same answer. This is a form of *performative belief revision* without genuine epistemic change.

6. Discussion and Conclusion

6.1. Key Takeaways

Our experiments reveal that current frontier VLMs have significant deficiencies in belief revision:

1. **Stubbornness is the dominant failure mode.** Across all models and conditions, stubbornness (refusing to revise a wrong answer) is far more common than regression (changing a right answer to wrong). This asymmetry suggests a conservative bias that prioritizes consistency over accuracy when evidence conflicts with initial beliefs.
2. **Prompting can partially mitigate stubbornness.** Belief-State prompting reduces stubbornness by 13–18 pp and improves accuracy by 4–8 pp over baseline. The mechanism appears to be *prospective preparation*—telling the model to expect updates primes it for revision. However, even the best prompting strategy leaves 37% of wrong answers unrevised, indicating that prompting alone cannot fully solve the problem.
3. **Metacognitive awareness exceeds behavioral change.** Counterfactual prompting reveals that models can often *identify* when evidence matters but fail to *act* on this knowledge proportionally. This dissociation between knowing and doing suggests that belief revision requires deeper architectural support, not just better prompting.
4. **Confidence inflation is systematic.** All models dramatically increase self-reported confidence after receiving additional evidence, even when their answer has not changed. This suggests that evidence exposure triggers a general "certainty boost" rather than calibrated belief updating.

6.2. Connections to Cognitive Science

Our findings map directly onto known cognitive phenomena. The stubbornness we observe parallels *anchoring bias* [9]—the tendency to over-weight initial information. The narrative lock-in failure mode resembles *confirmation bias*—interpreting new evidence to confirm rather than challenge existing beliefs. The confidence inflation mir-

rors the *illusion of explanatory depth*—gaining the feeling of understanding without genuine revision. These parallels suggest that VLMs may have implicitly learned human-like cognitive biases from training data, rather than optimal Bayesian belief revision.

6.3. Implications for VLM Architecture

Our results point toward the need for *explicit epistemic state tracking* in VLM architectures. Current autoregressive models generate token-by-token without maintaining a structured belief state that can be formally updated. Future architectures might incorporate:

- **Persistent belief registers** that store hypotheses with associated confidence weights and can be formally updated via Bayesian-like rules when new evidence arrives.
- **Evidence-weighted attention** mechanisms that compare new evidence against existing beliefs to detect conflicts.
- **Dual-process architectures** that separate fast initial hypothesis generation (System 1) from deliberate evidence integration (System 2), inspired by dual-process theory in cognitive science.

6.4. Limitations

Our study has several limitations. First, we use textual descriptions of visual scenes rather than actual images or video frames, meaning we test linguistic reasoning about visual scenarios rather than direct visual reasoning. Future work should extend EUP with real images from the BlackSwan Challenge. Second, our scenario set (N=52) is modest; larger-scale evaluation would strengthen statistical conclusions. Third, we evaluate only proprietary frontier models; open-source models may exhibit different patterns. Fourth, our scenarios are constructed rather than sampled from a validated benchmark; evaluation on the full BlackSwan validation split would further validate our findings.

6.5. Future Work

Several directions follow naturally: (i) extending EUP to use actual video frames from BlackSwan; (ii) testing whether fine-tuning with belief revision objectives can reduce stubbornness; (iii) evaluating belief revision across more reasoning types (spatial, temporal, causal); and (iv) developing architectures with explicit epistemic state mechanisms inspired by AGM belief revision axioms.

6.6. Conclusion

We introduced Evidence Update Prompting (EUP), a two-phase protocol for evaluating belief revision in VLMs. Our experiments across three frontier models and three prompting strategies reveal that current VLMs are substantially "stubborn"—maintaining incorrect beliefs despite contradictory evidence 37–62% of the time. Belief-State prompting reduces but does not eliminate this stubbornness, while Counterfactual prompting reveals a gap between metacognitive awareness and behavioral change. These findings establish that genuine non-monotonic reasoning remains an open challenge for VLMs, motivating architectures with explicit epistemic state mechanisms.

References

[1] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

[2] Kanzhi Cheng, Yantao Li, et al. Vision-language models can self-improve reasoning via reflection. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

[3] Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, and Leonid Sigal. Black swan: Abductive and defeasible video reasoning in unpredictable events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24201–24210, 2025.

[4] Shih-Han Chou, Shivam Chandhok, James J. Little, and Leonid Sigal. MM-R3: On (in-)consistency of vision-language models. In *Findings of the Association for Computational Linguistics: ACL*, 2025.

[5] Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. CogBench: A large language model walks into a psychology lab. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[6] Jack Hessel, Jena D. Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision (ECCV)*, 2022.

[7] Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1266–1279, 2022.

[8] Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. Visual abductive reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[9] Jiaxu Lou and Yifan Sun. Anchoring bias in large language models: An experimental study. *arXiv preprint arXiv:2412.06593*, 2024.

[10] John McCarthy. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence*, 13(1-2):27–39, 1980.

[11] Yulei Niu, Kaihua Tang, Hanwang Zhang, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[12] John L. Pollock. Defeasible reasoning. *Cognitive Science*, 11(4):481–518, 1987.

[13] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–137, 1980.

[14] Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 4661–4675, 2020.

[15] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Rein, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

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

[17] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *International Conference on Learning Representations (ICLR)*, 2024.

[18] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[20] Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10480–10496, 2024.

[21] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024.