When Verification Fails: How Compositionally Infeasible Claims Escape Rejection

Muxin Liu* **Delip Rao** **Grace Kim** **Chris Callison-Burch**
University of Pennsylvania
muxinliu@seas.upenn.edu

Abstract

Scientific claim verification, the task of determining whether claims are entailed by scientific evidence, is fundamental to establishing discoveries in evidence while preventing misinformation. This process involves evaluating each asserted constraint against validated evidence. Under the Closed-World Assumption (CWA), a claim is accepted if and only if all asserted constraints are positively supported. We show that existing verification benchmarks cannot distinguish models enforcing this standard from models applying a simpler shortcut called salient-constraint checking, which applies CWA's rejection criterion only to the most salient constraint and accepts when that constraint is supported. Because existing benchmarks construct infeasible claims by perturbing a single salient element they are insufficient at distinguishing between rigorous claim verification and simple salient-constraint reliance. To separate the two, we construct compositionally infeasible claims where the salient constraint is supported but a non-salient constraint is contradicted. Across model families and modalities, models that otherwise saturate existing benchmarks consistently over-accept these claims, confirming the prevalence of such shortcut reasoning. Via model context interventions, we show that different models and prompting strategies occupy distinct positions on a shared ROC curve, indicating that the gap between model families reflects differences in verification threshold rather than underlying reasoning ability, and that the compositional inference bottleneck is a structural property of current verification behavior that strategy guidance alone cannot overcome.

1 Introduction

Claim verification, the task of determining whether claims are entailed by evidence, is fundamental to grounding discoveries in evidence while preventing misinformation. Claim verification has been widely studied as a natural language inference (NLI) task, with large-scale benchmarks spanning general knowledge verification (Thorne et al., 2018; Aly et al., 2021), scientific literature (Wadden et al., 2020; 2022), clinical trials (Jullien et al., 2023; Vladika et al., 2024), tables (Chen et al., 2019; Lu et al., 2023), and charts (Akhtar et al., 2023; 2024; Masry et al., 2025; Wang et al., 2025). This judgment requires a precise epistemic standard: a claim should be accepted if and only if the evidence provides positive support for it, and absence of support warrants rejection. This standard is formalized as the Closed-World Assumption (CWA) (Clark, 1977; Reiter, 1981). Operationally, this requires checking every constraint the claim asserts (Min et al., 2023), including those derived compositionally from multiple evidence units (Si et al., 2021; Zhang et al., 2024a; Chen et al., 2024).

We evaluate Large Language Models (LLMs) across three established benchmarks, NLI4CT (Jullien et al., 2023; 2024) for clinical trial text, SCITAB (Lu et al., 2023) for scientific tables, and SciVer (Wang et al., 2025) for charts, and we find that models consistently achieve high accuracy on infeasible claims but lower accuracy on feasible ones, as shown in Figure 2. This asymmetry could be interpreted as evidence of conservative, closed-world verification, or could reflect a weaker shortcut observed in the literature for other tasks as well (Gururangan et al., 2018; McCoy et al., 2019; Liu et al., 2023).

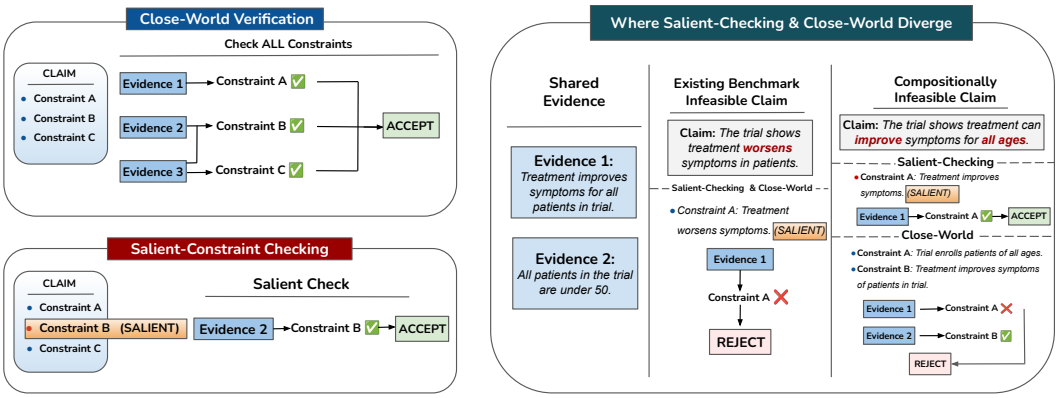


Figure 1: CWA verification checks all constraints; salient-constraint checking only finds the most prominent constraint. On existing benchmarks both agree since violated constraints are always salient. On compositionally infeasible claims, salient-constraint checking accepts while CWA rejects.

The construction of existing benchmarks makes full CWA and a simpler shortcut procedure behaviorally indistinguishable. Negative examples in these benchmarks (also called standard negatives) are created by perturbing a single salient element, such as flipping a number, reversing a trend, or swapping an entity label. This makes the perturbed constraint both the most checkable part of the claim and sufficient for rejection. As a result, as shown in Figure 1, a shortcut is sufficient when verifying **Existing Benchmark Infeasible Claim**: identify the most salient constraint in the claim, check whether it is supported by the evidence, and decide on that basis alone. This salient-constraint checking procedure is an easier, less formal approximation of CWA, as it applies CWA's rejection criterion but only to the most prominent constraint, skipping the exhaustive check that full CWA requires. It also explains the observed asymmetry. For infeasible claims, the salient constraint is violated, so rejection follows directly. For feasible claims, acceptance requires the salient constraint to be clearly retrievable, and any uncertainty defaults to rejection instead. High negative accuracy together with lower positive accuracy is therefore consistent with both full CWA and salient-constraint checking.

To expose the gap between salient-constraint checking and full CWA, we construct claims for which the shortcut yields acceptance but CWA requires rejection, as shown in Figure 1, under the **Compositionally Infeasible Claim** setting. These claims, which we call compositionally infeasible claims (also called adversarial negatives), are designed so that the salient constraint appears supported by the evidence, while the violation lies in a non-salient constraint such as a composed relationship across multiple parts of the evidence. We create hard negatives using a graph-based corrupt-and-propagate procedure, which preserves the salient observations while introducing violations through composing multiple parts of the same evidence source such as two sentences in a clinical report, two trends in a chart, or two columns in a table. Salient-constraint checking verifies a single constraint and accepts, while full CWA checks all constraints including non-salient ones and rejects.

Models that perform strongly on existing benchmarks fail consistently on adversarial examples, as summarized in Figure 2: they accept adversarial negatives at higher rates than standard negatives, and this gap persists across model families and scales. Models accept claims when the salient constraint appears supported, which produces systematic over-acceptance on adversarial negatives.

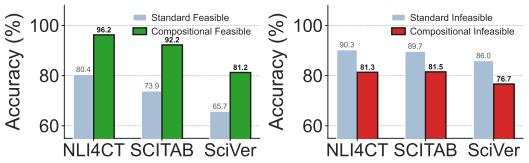


Figure 2: Averaged accuracy across models. Compositionally feasible claims score higher (↑) than standard positives; compositionally infeasible claims score lower (↓). Compositional structure shifts behavior toward acceptance.

We further study this failure with test-time guidance toward CWA, including closed-

world instructions and explicit decomposition. These interventions shift the tradeoff between acceptance and rejection rather than improving performance on adversarial negatives overall. This movement along the ROC curve suggests that different guidance strategies change how the model applies verification, but do not resolve the underlying reasoning bottleneck. To analyze reasoning more directly, we also introduce a rubric-based LLM judge that scores how well a model's reasoning trace follows CWA, including whether it checks all constraints the claim asserts or terminates early at the salient one. These scores reveal systematic differences across models, examples, and interventions, and provide a trace-level analysis of verification reasoning beyond outcome-based evaluation.

We make three contributions. First, we identify a structural limitation of existing verification benchmarks: single-constraint construction makes salient-constraint checking and full CWA behaviorally indistinguishable. Second, we construct compositionally infeasible claims across three modalities that expose this gap, and show consistent over-acceptance across model families and scales. Third, we show this failure reflects a reasoning bottleneck rather than a strategy choice through both intervention and trace analysis.

2 Theoretical Framework

We formalize the verification standards for scientific claim verification, introduce salient-constraint checking as a weaker procedure that existing benchmarks cannot distinguish from full CWA, and define compositional falsification as the condition that separates the two

2.1 Closed-World and Open-World Assumptions

Let $E = \{e_1, \ldots, e_m\}$ be a set of evidence pieces and c a claim asserting constraints $\mathcal{C}(c) = \{c_1, \ldots, c_n\}$. Under **CWA**, c is accepted iff every constraint is supported:

$$\text{CWA}(c, E) = \begin{cases} \text{Accept} & \text{if } \forall c_i \in \mathcal{C}(c),\ \exists S_i \subseteq E \text{ s.t. } \bigwedge_{e \in S_i} e \models c_i \\ \text{Reject} & \text{otherwise} \end{cases} \tag{1}$$

CWA is inherently $\forall$**-type**: every constraint must be checked. Under the **Open-World Assumption (OWA)**, a claim is rejected iff explicitly contradicted:

$$\text{OWA}(c, E) = \begin{cases} \text{Reject} & \text{if } \exists S \subseteq E \text{ s.t. } \bigwedge_{e \in S} e \models \neg c \\ \text{Accept} & \text{otherwise} \end{cases} \tag{2}$$

OWA is inherently $\exists$**-type**: it searches for any contradiction and accepts by default when none is found.

2.2 Salient-Constraint Checking

Definition 1 (Salience)**.** *The salience(c_i, E) of a constraint $c_i \in \mathcal{C}(c)$ measures how directly E speaks to c_i without requiring composition across multiple evidence units. The most salient constraint is $c^* = \arg\max_{c_i \in \mathcal{C}(c)}$ salience(c_i, E), corresponding operationally to the primary subject-predicate assertion verifiable from a single contiguous evidence unit.*

Salient-constraint checking applies the CWA verdict to c^* alone:

$$\text{V}_s(c, E) = \begin{cases} \text{Accept} & \text{if } \exists e \in E \text{ s.t. } e \models c^* \\ \text{Reject} & \text{otherwise} \end{cases} \tag{3}$$

V_s is a hybrid: it inherits CWA's rejection criterion but reduces the $\forall$ over all constraints to a $\exists$ over one salient constraint, making it operationally similar to OWA. Let c^+, c^- denote feasible/infeasible claim. Existing benchmarks construct c^- by perturbing a single salient element, ensuring $\exists!\, c^* \in \mathcal{C}(c^-)$ s.t. $e \models \neg c^*$ for some $e \in E$, with $\forall c_i \neq c^*$: $\exists e \in E$ s.t. $e \models c_i$. We call it *single-constraint regime*.

Proposition 1. *In the single-constraint regime, CWA and V_s produce identical verdicts on both c^- and c^+, assuming all constraints of c^+ are supported.*

Proof. For c^-: both check c^*, find it contradicted, and reject. For c^+: CWA finds all constraints supported; V_s finds c^* supported. Both accept. ■

Thus, existing benchmark performance cannot distinguish CWA from V_s.

2.3 Compositional Falsification

A **compositionally infeasible claim** $c^\dagger$ satisfies:

$$\exists e \in E \text{ s.t. } e \models c^*; \qquad \exists\, c_j \neq c^*,\ \exists S_2 \subseteq E \text{ s.t. } \bigwedge_{e \in S_2} e \models \neg c_j \tag{4}$$

The violation may be direct ($|S_2| = 1$) or compositional ($|S_2| \geq 2$).

Proposition 2. *On compositionally infeasible claims, V_s accepts and CWA rejects.*

Proof. V_s finds c^* supported and accepts; CWA finds c_j contradicted and rejects. ■

Proposition 3. *Degradation on compositionally infeasible claims relative to single-constraint infeasible claims is diagnostic for V_s as the default. Absence of degradation supports CWA.*

Proposition 3 follows from Propositions 1–2. We test this in §4.1 and find consistent degradation across model families and scales, confirming V_s as the default verification procedure.

3 Data Generation Method

A limitation of existing benchmarks is their reliance on surface-level perturbations (e.g., numerical flipping, trend reversal) to construct infeasible claims. Such single-constraint violation procedures can be detectable from merely a single sentence or a table cell. These correspond to the single-constraint regime where salient-constraint checking and CWA are indistinguishable (§2.2). Our method instead constructs compositionally infeasible claims where the salient constraint remains supported and the violation lies in a non-salient constraint, detectable only by combining multiple parts of the same evidence source.

3.1 Graph-Based Adversarial Construction

Our method takes as input a human-verified *feasible* claim together with its evidence source from an existing benchmark, and returns a compositionally infeasible counterpart grounded in the same evidence. We use a SciVer instance as a running example throughout this section and in Figure 3: the evidence is a scatter plot of pole locations at several pruning ratios, and the feasible claim states that as the pruning ratio increases from 0% to 80%, the maximum imaginary part of the poles reaches 80. Construction proceeds in two steps.

Step 1: Graph generation. We prompt an LLM with the evidence and the feasible claim to emit, in a single pass, a three-layer reasoning graph. A *node* is a short natural-language statement about the evidence, and the graph is a directed acyclic structure whose edges encode a *depends-on* relation: a node at one layer cites the lower-layer nodes from which it follows. Figure 3 displays only short entity names for the observation nodes (e.g., *Color Encoding*, *Y-Axis Range*) for space; every node carries full textual content in the released data. The layers are:

- **Observation layer** (O_1–O_5): five directly retrievable facts requiring no inference. In the running example, O_1 records the color-to-pruning-ratio mapping given in the legend, O_2 records the y-axis range, and O_3 records that the high-$|\mathrm{Im}|$ clusters contain points from every pruning ratio.
- **Context layer** (C_1–C_3): three nodes, each citing at least two observations and stating a structural relationship among them. For instance, C_3 combines O_3 and O_5 to state that large-$|\mathrm{Im}|$ points are localized near $\mathrm{Re} \approx 0$ rather than spread across all values of Re.
- **Interpretation layer** (I_1–I_3): three nodes, each citing at least two context nodes and articulating a higher-level reading of the evidence. For instance, I_3 combines C_2 and C_3 to state that the observed increase in $|\mathrm{Im}|$ is driven by specific clusters rather than holding uniformly across the chart. These nodes mirror the conclusions a human reader would draw and are the ones most closely tied to the original feasible claim.

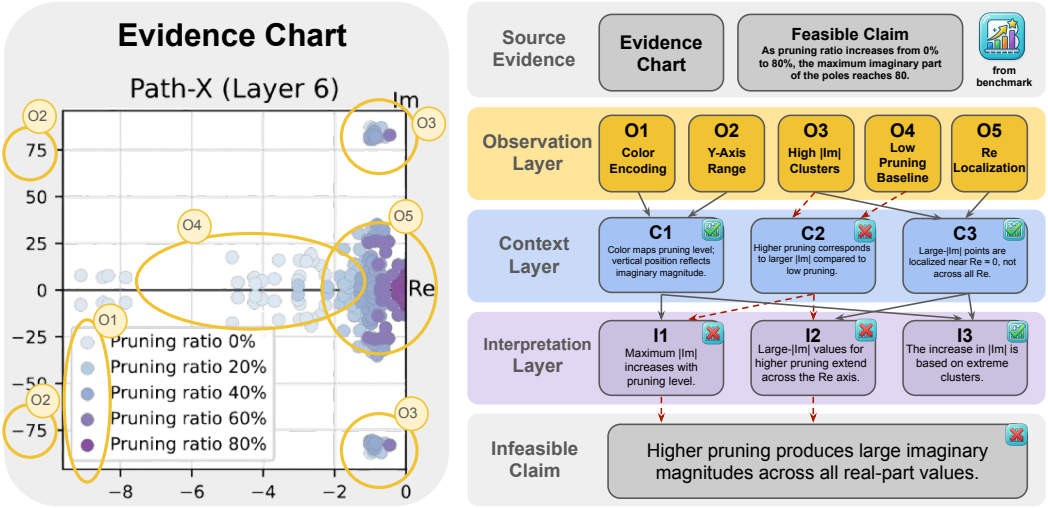


Figure 3: Graph-based adversarial construction. Observation nodes ground directly retrievable facts; context and interpretation nodes combine them into higher-level conclusions. Corrupting one non-salient node produces a compositionally infeasible claim while preserving all observation-layer support.

Step 2: Corruption and propagation. The LLM then applies a *minimal* edit (e.g. scope swapping, overgeneralization, or qualifier omission) to exactly one context or interpretation node, such that the edited node is no longer supported by the observations beneath it. In Figure 3 the corrupted node is C_2; the displayed text, "higher pruning corresponds to larger $|\mathrm{Im}|$ compared to low pruning," is the post-corruption state, and it is inconsistent with O_3, which shows that the high-$|\mathrm{Im}|$ clusters contain points from all pruning ratios rather than only high ones. The corruption is then propagated forward along outgoing edges: I_1 and I_2 both depend on C_2, so their content is rewritten to follow from the corrupted C_2. Finally, the LLM composes the corrupted interpretation layer into a single natural-language infeasible claim—here, "higher pruning produces large imaginary magnitudes across all real-part values." The full generation prompts are given in Appendix G, and Appendix I provides worked examples with explicit original-versus-corrupted labels for every node.

3.2 Datasets

We construct adversarial datasets from three existing scientific claim verification benchmarks: NLI4CT, SCITAB, and SciVer. For each dataset we sample 200 standard positive and negative claims from the original benchmark, then generate a pool of adversarial negatives using our graph-based construction method implemented with GPT-5-mini. After human validation to remove ambiguous and accidentally feasible claims, we retain approximately 200 adversarial negatives per domain. Full dataset construction details are in Appendix A, human annotation details in Appendix B, prompts in Appendix G and generation examples in Appendix I.

4 Experiments

We present three experiments that together characterize salient-constraint checking as the default verification procedure. §4.1 establishes the performance gap under compositional falsification across model families and scales. §4.2 isolates the failure at the reasoning layer. §4.3 tests whether the failure reflects a reasoning bottleneck or a strategy choice.

4.1 Performance Gap under Compositional Falsification

We evaluate eight proprietary models (GPT-5.4, GPT-5-mini, GPT-o4-mini, GPT-4o-mini, Gemini-3.0-Flash, Gemini-2.5-Flash, Claude-Sonnet-4.5, Claude-Haiku-4.5) and the Qwen3 family on NLI4CT, SCITAB, and SciVer, using standard and compositionally infeasible claims

Model	NLI4CT					SCITAB					SciVer				
	Standard			Adv	Δ(%)	Standard			Adv	Δ(%)	Standard			Adv	Δ(%)
	Pos.(%)	Neg.(%)	F_1	Neg.(%)		Pos.(%)	Neg.(%)	F_1	Neg.(%)		Pos.(%)	Neg.(%)	F_1	Neg.(%)	
GPT-5.4	80.2	94.1	0.86	88.7	↓5.4	75.5	93.0	0.83	91.0	↓2.0	55.9	92.5	0.68	87.7	↓4.8
GPT-5-mini	83.7	91.1	0.87	84.7	↓6.4	79.0	92.0	0.84	82.5	↓9.5	78.4	84.5	0.81	73.4	↓**11.2**
GPT-o4-mini	86.1	91.1	0.88	76.8	↓**14.2**	84.0	88.0	0.86	77.0	↓**11.0**	78.9	87.8	0.83	63.4	↓**24.4**
GPT-4o-mini	71.3	85.1	0.77	73.4	↓**11.7**	58.5	81.0	0.66	67.0	↓**14.0**	66.7	59.2	0.64	50.7	↓8.4
Gemini-3.0-Flash	86.1	86.6	0.86	73.9	↓**12.7**	87.0	89.0	0.88	74.0	↓**15.0**	75.6	87.3	0.80	60.6	↓**26.7**
Gemini-2.5-Flash	84.2	91.6	0.87	82.3	↓9.3	66.5	91.5	0.76	83.5	↓8.0	65.3	90.6	0.75	76.1	↓**14.5**
Claude-Sonnet-4.5	75.7	91.6	0.82	86.2	↓5.4	75.0	92.5	0.82	89.5	↓3.0	52.6	94.8	0.67	83.1	↓**11.7**
Claude-Haiku-4.5	76.2	91.1	0.82	84.7	↓6.4	65.5	90.5	0.75	87.5	↓3.0	52.6	91.1	0.65	85.9	↓5.1

Table 1: Evaluation across three domains. Pos./Neg. = accuracy (%) on standard positive/negative claims; F_1 = macro-F_1 on standard negatives; Adv Neg. = accuracy (%) on adversarial negatives; Δ = Adv Neg. − Standard Neg. (%). Δ values shown as ↓; **bold** entries indicate substantial drop ($|\Delta| > 10\%$).

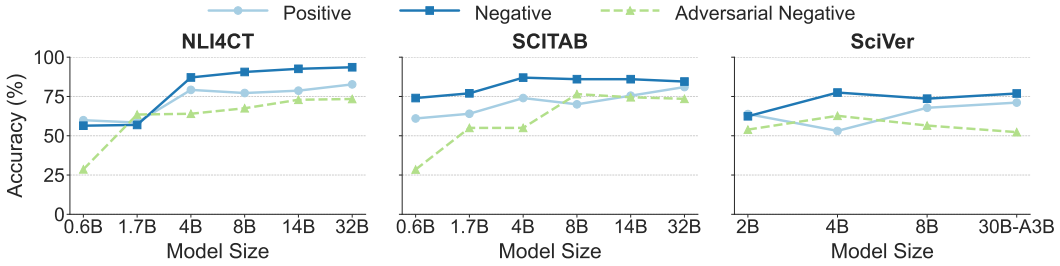


Figure 4: Scaling behavior across Qwen3 (VL) model sizes for three domains. Curves show accuracy for positive, negative, and adversarial negative claims.

from §3.2. Proprietary models are evaluated at temperature 1; Qwen3 uses thinking mode at temperature 0.6. All evaluations use 200 examples per domain; observed gaps exceed sampling uncertainty ($\pm$3% at n=200).

Table 1 shows a consistent performance gap across all models and domains. Every model shows lower accuracy on compositionally infeasible claims than on standard infeasible claims on NLI4CT (Δ from -3.0 to -12.7), with larger drops on SCITAB and SciVer: Gemini-3.0-Flash falls 15.0 and 26.7 points respectively. This pattern is not explained by original-benchmark F_1: Gemini-3.0-Flash achieves $F_1 = 0.88$ yet shows among the largest drops, while Claude-Sonnet-4.5 has lower $F_1 = 0.82$ with consistently smaller drops. The Claude family's robustness is better attributed to safety-induced over-refusal than stronger compositional verification (Röttger et al., 2024; Huang et al., 2025): their positive accuracy is among the lowest in the table, consistent with broad rejection across claim types rather than selective compositional checking.

Figure 4 shows the gap persists with scale. At small scales, standard negative accuracy rises steeply while adversarial negative accuracy lags far behind, consistent with CWA-style rejection without reasoning. As scale increases, adversarial negative accuracy improves, but the gap never closes: at 32B, NLI4CT shows a 20-point gap (93.6% vs. 73.4%), SCITAB a narrower but persistent gap, and SciVer limited gains throughout. The trajectory confirms that scaling sharpens retrieval and direct-falsification rejection but does not resolve failures that require compositional inference. These results support Proposition 3 across model families, scales, and modalities. The consistent pattern of high accuracy on directly falsifiable negatives and substantially lower on adversarial negatives is not an artifact of a particular architecture or scale, but a structural feature of current verification behavior.

4.2 Separating Retrieval from Compositional Reasoning

Models may apply salient-constraint checking for two distinct reasons. First, they may fail to retrieve the non-salient observations that reveal the violation, leaving no basis for checking beyond the salient constraint. Second, the relevant observations may be retrievable, but models cannot compose them into the joint inference required to detect the violation (Barnett et al., 2024). We design an experiment to distinguish between these possibilities.

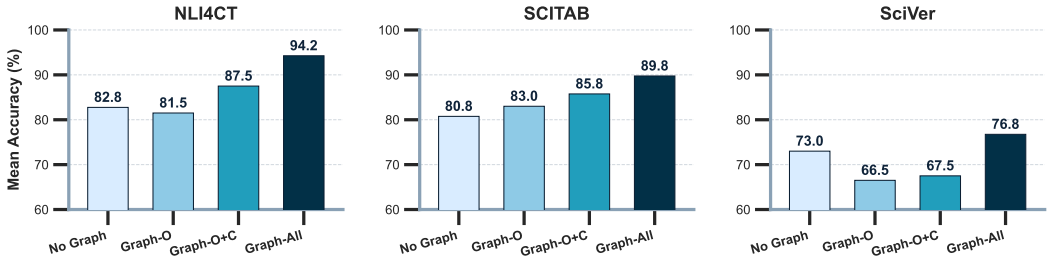


Figure 5: Adversarial negative accuracy under four evidence conditions across eight models. Graph-O provides no consistent improvement over no graph, while Graph-O+C produces intermediate gains and Graph-All approaches ceiling on NLI4CT and SCITAB.

We evaluate four evidence conditions. In **No graph**, the model receives only the claim and the raw document corpus. In **Graph-O**, the model additionally receives isolated atomic observations as the relevant evidence pieces without any compositional structure. In **Graph-O+C**, the model further receives context nodes explaining how observations compose into intermediate relationships. In **Graph-All**, the model receives the full graph including interpretation nodes that encode the conclusion reached by composing the context, and the model must identify where the infeasible claim diverges from this provided conclusion.

Figure 5 shows an incremental pattern. Graph-O shows no consistent improvement over no-graph: performance plateaus on NLI4CT, increases slightly on SCITAB, and drops on SciVer, suggesting retrieval is not the primary bottleneck. Graph-O+C produces intermediate gains across all domains, as context nodes provide the intermediate compositional relationships between observations without supplying the final conclusion. Graph-All approaches ceiling performance on NLI4CT and SCITAB, because interpretation nodes encode the result of full composition directly, reducing the task to identifying where the infeasible claim diverges from the provided conclusion; on SciVer, Graph-All recovers to baseline rather than exceeding it, likely because natural language descriptions interfere with direct visual perception (Sim et al., 2025). The bottleneck lies specifically between observation and interpretation: models have the relevant facts and can follow compositional structure when provided, but cannot perform the compositional inference step themselves (Sim et al., 2025).

4.3 Shifting Verification along the OWA–CWA Axis

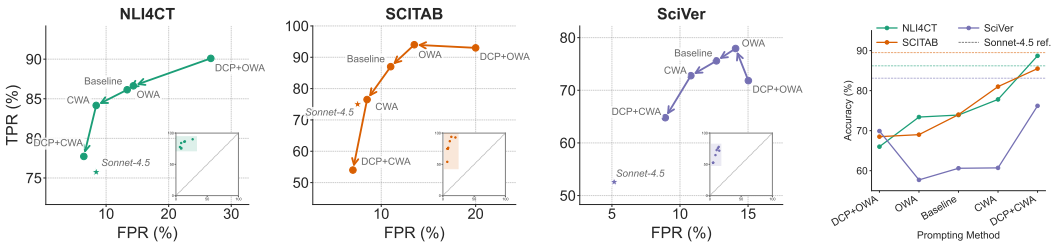


Figure 6: (**a**) Verification trajectories in ROC space (TPR vs. FPR) under different prompting conditions for Gemini-3.0-Flash. Arrows indicate the ordering of prompting variants (DCP+OWA → OWA → Baseline → CWA → DCP+CWA); stars denote the Sonnet-4.5 reference. (**b**) Adversarial negative accuracy under each prompting condition; dashed lines indicate the Sonnet-4.5 reference.

A further question is whether this reflects a reasoning ability limit or a failure to apply the right verification strategy. We intervene on Gemini-3.0-Flash with four prompting conditions with different positions on the OWA–CWA axis: **OWA**, **CWA**, and decomposition (DCP) variants **DCP+OWA** and **DCP+CWA** that explicitly enumerate subclaims before applying the verification rule. Decomposition is included because models prompted with CWA may not actively enumerate constraints without explicit instruction (Zhou et al., 2022; Dhuliawala

	Standard Negative			Adversarial Neg		
	Correct (TN)↑	Incorrect (FN)↑	Avg↑	TN↑	FN↑	Avg↑
GPT-5-mini	86.2	61.7	84.0	92.3	64.3	88.0
Gemini-3.0-Flash	83.0	61.9	80.1	91.4	65.1	84.9
+OWA	87.7	59.2	83.6	92.0	63.6	84.5
+OWA	88.0	63.2	85.9	93.9	63.4	87.1
Claude-Sonnet-4.5	89.1	71.4	87.6	92.1	70.9	89.2

Table 2: CWA trace scores (0–100) on NLI4CT. **Correct (TN)** denotes traces where the model successfully rejected the claim (True Negative); **Incorrect (FN)** denotes traces where the model falsely accepted it (False Negative). +CWA and +OWA are prompt intervention.

et al., 2024). DCP makes this an explicit step, ensuring the ∀-type checking CWA demands is actually carried out .

Figure 6a shows these conditions produce a smooth trajectory in ROC space ordered from most permissive to most strict (DCP+OWA → OWA → Baseline → CWA → DCP+CWA), with the baseline lying between OWA and CWA, which empirically align with §2.2 that salient-constraint checking is a hybrid of the two. Claude-Sonnet-4.5 lies on the same ROC curve as Gemini-3.0-Flash under intervention, and CWA or DCP+CWA prompting moves Gemini-3.0-Flash toward its operating point. This indicates that the performance gap between model families reflects differences in default verification threshold and strategy rather than differences in underlying reasoning ability. Both models share the same reasoning capability frontier, but apply different operating points by default. Claude-Sonnet-4.5's robustness to hard negatives, attributed in §4 to conservative rejection rather than compositional reasoning, is consistent with this picture, as safety alignment shifts its default operating point toward greater conservatism without changing the underlying capability frontier.

The trajectory also reveals the limit of strategy alone. Even under DCP+CWA, the strictest condition, hard negative accuracy does not reach the level achieved on standard negatives, and Figure 6b shows that gains in hard negative accuracy come at the cost of feasible claim acceptance. This tradeoff is not resolved by any prompting condition: stricter verification reduces over-acceptance of compositionally infeasible claims but increases rejection of feasible ones, because models cannot yet reliably distinguish the two through compositional reasoning at any threshold. Prompting shifts where models operate on the curve, but the curve itself reflects a reasoning bottleneck that verification strategy guidance alone cannot overcome (Arditi et al., 2022; Su et al., 2025).

5 Reasoning Trace Analysis

To complement our outcome-based evaluation with trace-level evidence (Lightman et al., 2023), we conduct a supplementary, single-domain analysis on NLI4CT. We score each reasoning trace using a rubric-based judge (Kim et al., 2023) targeting salient-constraint checking versus full CWA verification. Each of four dimensions is scored from -2 (salient-constraint checking) to $+2$ (full CWA) and aggregated to 0–100. Higher scores indicate reasoning closer to full CWA verification. The complete rubric and human validation results are detailed in Appendix E.

Cross-model comparison. CWA scores independently recover the robustness ranking from Table 1: Gemini-3.0-Flash scores lowest (80.1 standard, 84.9 adversarial), GPT-5-mini intermediate (84.0, 88.0), and Claude-Sonnet-4.5 highest (87.6, 89.2). The rubric-based ordering matches the accuracy-based ordering, confirming that robustness to compositional falsification tracks how thoroughly models check constraints in their reasoning traces.

Within-model comparison. Across all models, average scores on compositionally infeasible claims are higher than on standard claims (Gemini-3.0-Flash 84.9 vs. 80.1, GPT-5-mini 88.0 vs. 84.0, Claude-Sonnet-4.5 89.2 vs. 87.6), indicating that the compositional nature of these claims induces more thorough constraint checking in reasoning traces. The gap between correct and incorrect traces is larger on compositionally infeasible claims than on standard

	GPQA		AIME	
	$F_1 \uparrow$	Prec@Rej↑	$F_1 \uparrow$	Prec@Rej↑
Baseline	.218	.253	.287	.256
OWA	.165	.206	.238	.226
CWA	**.262**	**.273**	**.347**	**.290**

Table 3: Self-verification on GPQA Diamond and AIME. A rejection is valid if the solver response is incorrect and the verifier reasoning correctly identifies the exact solver error. F_1: macro-F_1 over valid rejections and valid acceptances. Prec@Rej: precision over valid rejections only. **Bold**: best.

ones, because correctly rejecting a compositionally infeasible claim requires enumerating and composing non-salient constraints, raising the reasoning bar, while failed traces on both claim types reflect a similar pattern of early termination at the salient constraint.

Intervention. CWA prompting raises correct case scores on standard $(83.0 \rightarrow 88.0)$ and adversarial $(91.4 \rightarrow 93.9)$ claims, confirming CWA guidance shifts reasoning toward more thorough constraint checking. Incorrect adversarial scores remain nearly unchanged $(65.1 \rightarrow 63.4)$, indicating a reasoning threshold below which incorrect verdicts occur. OWA prompting does not consistently lower scores despite increasing acceptance rate, confirming the rubric measures constraint coverage in reasoning rather than verdict direction.

6 Discussion

Verification mode in self-verification. Scientific claim verification is a form of external verification as it evaluates external claims against evidence. Verification also arises internally, when models check their own reasoning as a common reasoning strategy. If salient-constraint checking reflects a general verification tendency rather than a property specific to external claims, the same failure mode should appear in self-verification. Models may accept their own incorrect responses without checking all constraints that would reveal the error.

We test this by asking GPT-4o-mini to self-verify its own responses on 200 problems each from AIME (Veeraboina, 2023) and GPQA Diamond (Rein et al., 2024) under baseline and CWA/OWA prompt intervention, with a stronger model judging both verdict and reasoning validity. As shown in Table 3, baseline performance falls between OWA and CWA on both F_1 (over valid rejections and valid acceptances) and Precision at Rejection (Prec@Rej, defined as the proportion of correct rejections accompanied by faithful reasoning), mirroring the ROC trajectory in §4.3. OWA underperforms baseline, consistent with increased over-acceptance. CWA achieves the highest Prec@Rej, indicating that stricter constraint checking improves not just rejection rate but the quality of reasoning behind rejections, which could be a potential direction for fixing confirmation bias (Huang et al., 2023; Zhou et al., 2025; Wan et al., 2025).

Origins of salient-constraint checking. We hypothesize that salient-constraint checking emerges from pretraining on scientific data, where claims are typically supported by directly retrievable features and compositional violations rarely appear. In this setting, checking the salient constraint is nearly always sufficient, providing little training signal to develop full CWA verification (Dziri et al., 2023; Kambhampati et al., 2024). This mismatch between pretraining corpora and the structure of scientific misinformation explains both why salient-constraint checking is the default and why prompting can shift but not overcome the reasoning bottleneck.

Future work. This work is primarily diagnostic, as we identify and isolate a failure mode in scientific verification. Although our graph-based pipeline yields structured reasoning trajectories suitable for post-training (SFT and RLVR), we do not view this as a straightforward solution (Yue et al., 2025). We trained Qwen3-8B on compositional examples using both approaches and shifted the decision boundary toward broader rejection rather than improving compositional reasoning quality. More fundamentally, most training tasks reward $\exists$-type reasoning, where finding any supporting evidence is sufficient for success, leaving the $\forall$-type reasoning that full CWA verification requires systematically undertrained (Dziri et al., 2023; McCoy et al., 2023). Addressing both the compositional inference bottleneck and

this verification tendency is likely necessary for genuine progress, and we leave the design of such interventions to future work.

7 Related Work

Scientific claim verification. Claim verification benchmarks span general knowledge (Thorne et al., 2018; Aly et al., 2021), scientific literature (Wadden et al., 2020; 2022; Lal et al., 2025; Kumar et al., 2025), clinical trials (Jullien et al., 2023; 2024), tables (Chen et al., 2019; Lu et al., 2023), and charts (Akhtar et al., 2024; Wang et al., 2025). Previous works improve claim verification ability of LLM through retrieval (Adjali, 2024; Liu et al., 2024; Chen et al., 2024; Sallami et al., 2025), decomposition (Sahu et al., 2024; Lu et al., 2025; Vladika et al., 2025) and reasoning improvement (Wadden et al., 2025; Freedman et al., 2025). Models are known to exploit surface heuristics rather than genuine reasoning (Gururangan et al., 2018; Clark et al., 2019; Niven & Kao, 2019; McCoy et al., 2019). We identify why existing benchmarks cannot detect a specific failure mode, and construct examples that can detect it.

Compositional adversarial data construction. Compositional reasoning benchmarks use graph-based methods to generate multi-hop question answering pairs (Yang et al., 2018; Tu et al., 2019; Pan et al., 2020; Trivedi et al., 2022). Previous works construct adversarial and counterfactual data to evaluate LLMs through human annotation (Gardner et al., 2020; Kiela et al., 2021; Liu et al., 2023), heuristic rule-based perturbations (Gan et al., 2024; Waghela et al., 2024), and automated LLM-driven synthetic generation (Wu et al., 2024; Bhattacharjee et al., 2024; Zhang et al., 2025). We follow the line of work that applies systematic graph perturbations to generate adversarial data(Sharma et al., 2023; Wu et al., 2024; Zhang et al., 2024b; Hong et al., 2025; Cheng et al., 2025).

8 Conclusion

We identify a gap between the CWA standard and observed model behavior: because existing benchmarks encode a permissive standard, models trained and evaluated under it inherit the same bias, and competitive benchmark accuracy does not survive stricter evaluation. Our contribution is a construction that makes this distinction measurableby applying minimal edits to a single non-salient node of a reasoning graph and leaving all observation-level support intact so that failures are attributable to composition rather than to harder evidence retrieval. The resulting claims mirror the errors that matter downstream: overgeneralization and wrong scoping in paper review, missing qualifiers and conflated subgroups in evidence synthesis from trial reports and experimental tables, and the self-verification failures of LLMs embedded in automated research pipelines, whose generated claims exhibit these same patterns. We therefore view compositional infeasibility not as an adversarial edge case but as the norm in which claim verification is deployed, and release our data and generation pipeline to support evaluation under this standard.

9 Acknowledgement

This research was developed with funding from the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300) and is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 56000026C0019 (the BENGAL program). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Omar Adjali. Exploring retrieval augmented generation for real-world claim verification. In *Proceedings of the Seventh Fact Extraction and VERification Workshop* (*FEVER*), pp. 113–117, 2024.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 399–414, 2023.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. Chartcheck: Explainable fact-checking over real-world chart images. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13921–13937, 2024.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. The fact extraction and verification over unstructured and structured information (feverous) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification* (*FEVER*), pp. 1–13, 2021.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. *URL https://arxiv. org/abs/2406.11717*, 2022.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pp. 194–199, 2024.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. Zero-shot llm-guided counterfactual generation: A case study on nlp model evaluation. In *2024 IEEE International Conference on Big Data* (*BigData*), pp. 1243–1248. IEEE, 2024.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*Volume 1: Long Papers*), pp. 3569–3587, 2024.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

Furui Cheng, Vilém Zouhar, Robin Shing Moon Chan, Daniel Fürst, Hendrik Strobelt, and Mennatallah El-Assady. Understanding large language model behaviors through interactive counterfactual generation and analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language Processing* (*EMNLP-IJCNLP*), pp. 4069–4082, 2019.

Keith L Clark. Negation as failure. In *Logic and data bases*, pp. 293–322. Springer, 1977.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in neural information processing systems*, 36: 70293–70332, 2023.

Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. Argumentative large language models for explainable and contestable claim verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14930–14939, 2025.

Esther Gan, Yiran Zhao, Liying Cheng, Mao Yancan, Anirudh Goyal, Kenji Kawaguchi, Min-Yen Kan, and Michael Shieh. Reasoning robustness of llms to adversarial typographical errors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10449–10459, 2024.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2* (*Short Papers*), pp. 107–112, 2018.

Pengfei Hong, Navonil Majumder, Deepanway Ghosal, Somak Aditya, Rada Mihalcea, and Soujanya Poria. Evaluating llms' mathematical and coding competency through ontology-guided interventions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22811–22849, 2025.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. Nli4ct: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16745–16764, 2023.

Maël Jullien, Marco Valentino, and André Freitas. Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation* (*SemEval-2024*), pp. 1947–1962, 2024.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*, 2024.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. Sciclaimhunt: A large dataset for evidence-based scientific claim verification. In *2025 International Joint Conference on Neural Networks* (*IJCNN*), pp. 1–10. IEEE, 2025.

Yash Kumar Lal, Manikanta Bandham, Mohammad Saqib Hasan, Apoorva Kashi, Mahnaz Koupaee, and Niranjan Balasubramanian. Musciclaims: Multimodal scientific claim verification. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 3285–3307, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Hao Liu, Ali Soroush, Jordan G Nestor, Elizabeth Park, Betina Idnay, Yilu Fang, Jane Pan, Stan Liao, Marguerite Bernard, Yifan Peng, et al. Retrieval augmented scientific claim verification. *JAMIA open*, 7(1):ooae021, 2024.

Nelson F Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025, 2023.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. Scitab: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7787–7813, 2023.

Yining Lu, Noah Ziems, Hy Dang, and Meng Jiang. Optimizing decomposition for optimal claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp. 5095–5114, 2025.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, 2025.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3428–3448, 2019.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4658–4664, 2019.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. Semantic graphs for generating deep questions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1463–1475, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.135. URL https://aclanthology.org/2020.acl-main.135/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First conference on language modeling*, 2024.

Raymond Reiter. On closed world data bases. In *Readings in artificial intelligence*, pp. 119–140. Elsevier, 1981.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*Volume 1: Long Papers*), pp. 5377–5400, 2024.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. Pelican: Correcting hallucination in vision-llms via claim decomposition and program of thought verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8228–8248, 2024.

Dorsaf Sallami, Sabrine Amri, and Esma Aïmeur. Claimveragents: A multi-agent retrieval-augmented claim verification framework. In *2025 IEEE/ACS 22nd International Conference on Computer Systems and Applications* (*AICCSA*), pp. 1–6. IEEE, 2025.

Kartik Sharma, Rakshit Trivedi, Rohit Sridhar, and Srijan Kumar. Temporal dynamics-aware adversarial attacks on discrete-time dynamic graph models. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2023–2035, 2023.

Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. Topic-aware evidence reasoning and stance-aware aggregation for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*Volume 1: Long Papers*), pp. 1612–1622, 2021.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. Can vlms actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24452–24470, 2025.

Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman, and Maarten Sap. Ai-liedar: Examine the trade-off between utility and truthfulness in llm agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (*Volume 1: Long Papers*), pp. 11867–11894, 2025.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long Papers*), pp. 809–819, 2018.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2704–2713, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1260. URL https://aclanthology.org/P19-1260/.

Hemish Veeraboina. Aime problem set 1983-2024, 2023. URL https://www.kaggle.com/datasets/hemishveeraboina/aime-problem-set-1983-2024.

Juraj Vladika, Phillip Schneider, and Florian Matthes. Healthfc: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (*LREC-COLING 2024*), pp. 8095–8107, 2024.

Juraj Vladika, Ivana Hacajova, and Florian Matthes. Step-by-step fact verification system for medical claims with explainable reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (*Volume 2: Short Papers*), pp. 805–816, 2025.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 7534–7550, 2020.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. Scifact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734, 2022.

David Wadden, Kejian Shi, Jacob Morrison, Alan Li, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, et al. Sciriff: A resource to enhance language model instruction-following over scientific literature. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6083–6120, 2025.

Hetvi Waghela, Jaydip Sen, Sneha Rakshit, and Subhasis Dasgupta. Adversarial text generation with dynamic contextual perturbation. In *2024 IEEE Calcutta Conference* (*CALCON*), pp. 1–6. IEEE, 2024.

Yue Wan, Xiaowei Jia, and Xiang Lorraine Li. Unveiling confirmation bias in chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3788–3804, 2025.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. Sciver: Evaluating foundation models for multimodal scientific claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp. 8562–8579, 2025.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*Volume 1: Long Papers*), pp. 1819–1862, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. Causal walk: Debiasing multi-hop fact verification with front-door adjustment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19533–19541, 2024a.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. Darg: Dynamic evaluation of large language models via adaptive reasoning graph. *Advances in Neural Information Processing Systems*, 37:135904–135942, 2024b.

Zhehao Zhang, Weijie Xu, Fanyou Wu, and Chandan K Reddy. Falsereject: A resource for improving contextual safety and mitigating over-refusals in llms via structured reasoning. *arXiv preprint arXiv:2505.08054*, 2025.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models, 2023. *URL https://arxiv. org/abs/2205.10625*, 2022.

Yefan Zhou, Austin Xu, Yilun Zhou, Janvijay Singh, Jiang Gui, and Shafiq Joty. Variation in verification: Understanding verification dynamics in large language models. *arXiv preprint arXiv:2509.17995*, 2025.

A Benchmark Collection

Dataset	Original source	Subset used	Generated	Validated	Notes
NLI4CT	Clinical trial claims	Single-trial examples only	242	203	We exclude multi-trial claims and generate hard negatives only from single-trial evidence.
SciTab	Scientific table claims	Randomly selected table subset	230	200	Candidate pool was sufficient to retain 200 validated hard negatives after manual filtering.
SciVer	Scientific figure claims	Single-chart direct-reasoning subset	213	180	SciVer includes multiple reasoning types and multimodal evidence settings (e.g., chart, text, table, or multiple evidence pieces). We restrict to single-chart examples, which yields 213 source items; after validation, fewer than 200 hard negatives remain.

Table 4: Dataset subsets and annotation outcomes for adversarial hard-negative construction.

	Evidence	Feasible claims	Standard infeasible claims	Our hard-negative perturbations
NLI4CT	Breast cancer clinical trial report excerpts; single section per instance.	Expert-written claims requiring biomedical, numerical, and commonsense reasoning.	Contrast-set perturbations, including numerical, lexical, semantic, and syntactic changes.	Compositional clinical misreadings such as qualifier omission, scope/population shifts, conditional or boundary flips, cross-arm or proxy misattribution, and surfaced assumptions.
SciTab	Scientific tables with captions from NLP/ML papers.	Sentences from the paper grounded in the table; two-annotator verified.	InstructGPT-generated negatives with wrong calculations, value mismatches, and incorrect approximations.	Structural table perturbations such as inconsistent baselines, part–whole and averaging errors, derived-metric overextension, null-as-zero aggregation, and occasional causal or categorical overreach.
SciVer	Chart screenshots from CS/arXiv papers; direct-reasoning subset only.	Expert-written claims grounded in chart content.	Expert-rewritten negatives with hallucinated values, value misattribution, or relationship reversal.	Structural visual misreadings such as local-to-global overreach, ignored panel or layout boundaries, visual analogy or element conflation, prominence-based over-interpretation, and unsupported comparison classes.

Table 5: Benchmark datasets used in this work, including both their original infeasible-claim construction and our compositional hard-negative perturbations.

B Annotation Details

Validation for this work is diagnostic rather than benchmark construction. Our goal is to confirm that generated hard negatives are not accidentally feasible and do not introduce ambiguous cases, not to establish a reusable labeled dataset with publication-grade inter-annotator agreement. In this context, single-annotator validation is standard practice as the annotator's role is to apply a well-defined binary judgment (is this claim clearly infeasible given the document, or is it ambiguous/accidentally true?) rather than to make subjective interpretive assessments.

All generated hard negatives were manually validated by a single author. During annotation, the annotator was given the source evidence, the generated reasoning graph, the perturbation logic (including the corrupted node), and the final infeasible claim. Each example was judged

based on both whether the final claim was clearly infeasible with respect to the source and whether the intermediate corruption constituted a valid adversarial transformation.

The final acceptance rates were 203/242 (83.9%) for NLI4CT, 200/230 (87.0%) for SciTab, and 180/213 (84.5%) for SciVer. Most rejected candidates were ambiguous rather than incorrect. They introduced higher-level interpretations, external assumptions, or broader statistical conclusions that were not clearly licensed by the source, but were also not unambiguously contradicted by it. A smaller portion were genuinely invalid generations, typically because an intended overgeneralization remained accidentally compatible with the evidence.

Validation was conducted by a single annotator with CS-domain expertise. SCITAB and SciVer draw exclusively from CS papers, making the annotator's domain knowledge sufficient and inter-annotator variability is expected to be low. For NLI4CT, judgments were grounded in source-document consistency rather than independent clinical expertise. The task was to determine whether the generated claim and corruption were supported by the provided trial text, not to assess medical validity beyond the evidence. This grounds the judgment in objective textual entailment rather than domain-specific interpretation, limiting the role of annotator background knowledge. We acknowledge that multi-annotator review with inter-rater agreement reporting would further strengthen label quality for future benchmark use, and we intend to address this in future work.

C Decoding Randomness

To assess whether findings are stable across decoding randomness, we evaluate the Qwen3 family at three temperatures: $T = 0$, $T = 0.6$, and $T = 1.0$ on SCITAB. Figure 7 shows accuracy on positive, standard negative, and adversarial negative claims across model sizes for each temperature. Adversarial negative accuracy is lower than standard negative

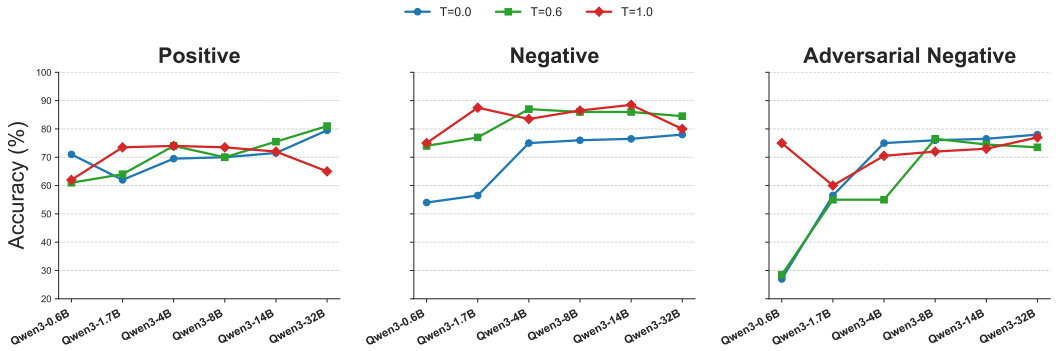


Figure 7: Qwen3 model families performance of SCITAB with decoding at different temperatures

accuracy at every model size and every temperature, confirming that the routing failure is not an artifact of a particular decoding choice. However, the magnitude of the gap varies substantially with temperature. At $T = 0$, Qwen3-0.6B achieves only 27% on adversarial negatives despite 86.5% on standard negatives, a 59.5-point difference, because deterministic decoding commits fully to the model's prior, which at small scale is strong acceptance. As scale increases, the gap narrows monotonically, reaching -8.0% at 32B, producing the clearest scaling signal. At $T = 1.0$, the gap at small scale nearly vanishes: Qwen3-0.6B shows 0% difference, because high-temperature sampling occasionally follows lower-probability reasoning paths that accidentally produce correct rejections. This stochastic masking makes $T = 1.0$ a less reliable diagnostic for the routing failure at small scale. $T = 0.6$ reaches a balance between these extremes, producing stable and monotonically improving adversarial negative accuracy across scale without the extreme compression of $T = 1.0$ or the floor effects of $T = 0$ at small sizes. We therefore use $T = 0.6$ as our primary evaluation setting for the Qwen3 scaling experiment, and report full results across all three temperatures here for completeness.

D Ablation Study

Surface form. Compositionally infeasible claims tend to be longer than standard infeasible claims due to the compositionality. If surface form rather than compositional structure drives the gap, rephrasing standard infeasible claims to match the style of compositionally infeasible claims should produce a similar drop. We take a subset of 50 standard infeasible claims from each domain, and rephrase each standard infeasible claim into a longer, more formal surface form while preserving the identical factual assertion and violation. As shown in Figure 8, rephrased negatives track original standard negatives closely across all three domains, which proves that surface form does not explain the gap between standard infeasible and compositionally infeasible claims.

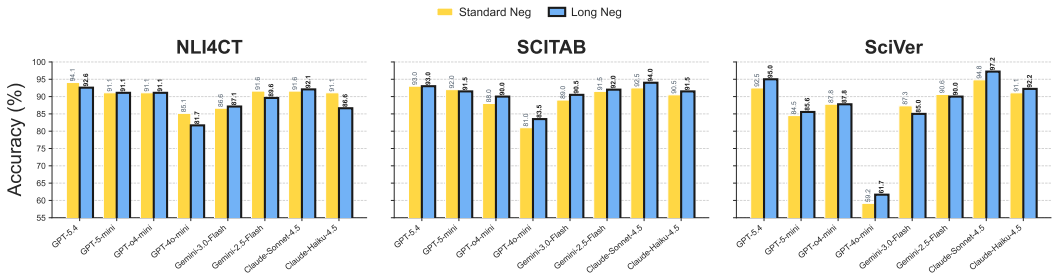


Figure 8: Longer negatives show no improvement or degradation over standard negatives, ruling out surface form as the reason for performance drop on adversarial negatives

Generator-Verifier Bias To assess whether evaluation outcomes are affected by the stylistic artifacts introduced by the generation pipeline, we compare performance on two claim versions across 50 adversarial negatives per domain. In the **rephrased** condition, GPT-5-mini-generated claims are rephrased by Gemini-3.0-Flash while preserving logical content; in the **original** condition, raw GPT-5-mini-generated claims are evaluated without rephrasing. If GPT-5-mini were biased toward its own writing style, it should score consistently higher on original claims than on rephrased ones. Table 6 shows no consistent pattern supporting this. On NLI4CT, both evaluators produce identical scores across conditions, showing no effect. On SCITAB, GPT-5-mini scores higher on original claims (82% vs. 74%), consistent with a weak stylistic preference. On SciVer, GPT-5-mini scores higher on rephrased claims (66% vs. 78%), which is opposite to the bias direction. GPT-5-mini appears more sensitive to surface-level style changes than Gemini-3.0-Flash, but the inconsistency in direction across domains and evaluators indicates no systematic generation-verifier bias. Therefore, the results reported in the main paper for the GPT model family are not inflated by stylistic familiarity with the generation pipeline.

Domain	Evaluator	Rephrased (%)	Original (%)
NLI4CT	GPT-5-mini	86.0	86.0
	Gemini-3.0-Flash	76.0	76.0
SCITAB	GPT-5-mini	74.0	82.0
	Gemini-3.0-Flash	72.0	70.0
SciVer	GPT-5-mini	78.0	66.0
	Gemini-3.0-Flash	66.0	70.0

Table 6: Hard negative accuracy on Gemini-rephrased claims (rephrased) vs. raw GPT-5-mini-generated claims (original), evaluated by two independent model families. Differences are small and inconsistent in direction, ruling out systematic generation-verifier bias.

‘‘‘

E LLM Judge Rubrics and Human Agreement

Dimension	Salient-constraint checking (-2)	Full CWA $(+2)$
Constraint enumeration	Checks one prominent constraint	Lists all constraints before checking any
Non-salient coverage	Stops at salient constraint	Checks non-salient and composed constraints
Evidence boundary	Accepts on local consistency	Rejects when scope exceeds evidence
Verdict warrant	Notes gaps but accepts	Verdict follows from all gaps

Table 7: Rubric dimensions for scoring reasoning traces on the salient-constraint checking vs. full CWA axis. Each dimension is scored $\{-2, \ldots, +2\}$ with anchored examples.

To assess the validity of the rubric-based LLM judge, a single annotator with NLP research experience independently scored a stratified sample of 50 reasoning traces from NLI4CT, drawn to cover correct rejection and incorrect rejections in proportion to their occurrence in the full evaluation set. The annotator scored each trace on the same four dimensions (-2 to $+2$ per dimension, aggregated to 0–100) without access to the LLM judge's scores. Agreement between the human annotator and LLM judge, measured as Spearman's ρ on the aggregated 0–100 scale, was $\rho = 0.43$ for correct traces and $\rho = 0.53$ for incorrect traces, yielding a mean of $\rho = 0.47$ across the two models. This indicates moderate agreement between the human annotator and the automated judge. The overall pooled correlation is lower due to range restriction, as most traces score in a narrow band (75–90), which compresses rank differences and reduces Spearman's sensitivity to genuine agreement. Single-annotator validation is a limitation of this study. Multi-annotator validation with formal inter-annotator agreement reporting remains an important direction for future work.

F Self-Verification Experiment Details

Task and Dataset Selection We evaluate on GPQA Diamond (Rein et al., 2024) and AIME (Veeraboina, 2023). Both datasets require multi-step reasoning where a correct solution must satisfy multiple constraints simultaneously, making them natural analogues to the compositional verification setting in the main paper. We select GPT-4o-mini as the solver because it achieves non-trivial but imperfect accuracy on both tasks, producing a mix of correct and incorrect responses suitable for studying self-verification behavior. A model that always succeeds provides no incorrect responses to verify; a model that always fails provides no correct responses to accept. GPT-4o-mini's accuracy in the range of 25–35% on these tasks provides sufficient signal in both directions.

Experiments For each problem, GPT-4o-mini first produces a solution in a single forward pass. The same model then self-verifies its own response under one of three prompting conditions: baseline, CWA-instructed, or OWA-instructed, using the prompts similar to Appendix G. The self-verification step receives the problem statement and the model's own solution and must output (1) accept or reject (2) reasoning.

Ground truth verdicts and reasoning quality are assessed by GPT-5-mini, which achieves over 90% accuracy on both benchmarks and serves as a reliable judge. The judge evaluates two things independently: (1) whether the solver's accept/reject answer is correct, and (2) whether the verifier's reasoning correctly identifies the source of error when rejecting.

Evaluation Metrics we define the following case taxonomy:

We report macro-F1 and Precision at Rejection (Prec@Rej) defined as TP / (TP + lucky_catch), measuring what fraction of rejections reflect genuine constraint checking rather than incidental rejection.

Solver	Verifier	Reasoning	Label	Description	Count as
Incorrect	Reject	Valid	TP	Verifier correctly identifies the actual error	True Positive
Correct	Accept	—	TN	Verifier correctly accepts a correct solution	True Negative
Incorrect	Reject	Invalid	Lucky Catch	Verifier rejects but for wrong/unrelated reason	False Positive
Correct	Reject	—	FP	Verifier incorrectly rejects a correct solution	False Positive
Incorrect	Accept	—	FN	Verifier incorrectly accepts an incorrect solution	False Negative

Table 8: Case taxonomy for self-verification evaluation.

Dataset	Variant	n	Acc	F1	TP	TN	FP	FN	Prec@Rej
GPQA	baseline	198	0.313	0.218	19	43	82	54	$19/75 \approx 0.253$
	CWA	198	**0.318**	**0.262**	24	39	94	41	$\mathbf{24/88 \approx 0.273}$
	OWA	198	0.283	0.165	14	42	81	61	$14/68 \approx 0.206$
AIME	baseline	200	0.255	0.287	30	21	93	56	$30/117 \approx 0.256$
	CWA	200	**0.285**	**0.347**	38	19	101	42	$\mathbf{38/131 \approx 0.290}$
	OWA	200	0.230	0.238	24	22	87	67	$24/106 \approx 0.226$

Results

The ordering CWA > baseline > OWA on both F1 and Prec@Rej mirrors the ROC trajectory observed in the main paper, with the baseline lying between OWA and CWA. This replicates the external verification pattern in an internal setting, consistent with salient-constraint checking being a general verification tendency rather than a property specific to external evidence evaluation.

CWA prompting improves Prec@Rej on both datasets, indicating that stricter constraint checking improves not just rejection rate but the quality of reasoning behind rejections. The real catch rate under CWA (0.273 on GPQA, 0.290 on AIME) exceeds the baseline (0.253, 0.256), confirming that CWA guidance produces more genuine error identification rather than merely more rejections.

However, the absolute F1 values remain low across all conditions. This reflects the same reasoning bottleneck identified that prompting shifts the operating point but cannot overcome the underlying compositional reasoning limitation. The high false positive counts under CWA (94 on GPQA, 101 on AIME) indicate that stricter verification causes over-rejection of correct solutions, consistent with the tradeoff between acceptance and rejection accuracy observed in the external verification setting.

G Prompts for Data Generation

Graph Generation Prompt

Build a reasoning graph for claim verification. You are given:
- clinical trial: a full document, a target excerpt, and a feasible claim
- table: a table and a feasible claim
- chart: an image and a feasible claim

The claim is the surface. Your job is to write TRUE context beneath it so later we can create subtle misinformation by mischaracterizing that context. Write three layers:

LAYER 1 — OBSERVATIONS (*CLAIM-ANCHORING O1..O5: elements related to or directly supporting the claim*)
- clinical trial: include numerical values, cohort/arm labels, outcome definitions, denominators, eligibility criteria, measurement definitions, timeframes.
- table: describe structural regions or bindings (rows, columns, headers); focus on distributions, hierarchy, or constraints; avoid isolated cells unless they define outliers.
- chart: describe visible anchors (legend, axes, ticks, panels, marks); include positions, scales, labels, and annotated elements.

Rules: all observations must be literal and directly present in the input; no arithmetic, comparison, or inference.

LAYER 2 — CONTEXT (*describe how observations relate structurally*)
- clinical trial: connect observations to broader trial structure.
- table: describe structural relationships (ordering, gaps, trends, scope boundaries).
- chart: describe structural relationships (ordering, gaps, trends, scope boundaries).

Rules: each context step must cite $\geq$2 observations; must be directly supported by the input; do NOT restate the claim.

LAYER 3 — INTERPRETATION (*explain what the contextual structure implies*)
- clinical trial: ground in full document; focus on population, endpoints, conditions, or applicability.
- table: focus on ranking scope, comparison validity, ordering, representativeness.
- chart: focus on trends, generalization, stability, or what values represent.

Rules: each interpretation step must cite $\geq$2 context steps; must remain grounded in the input; do NOT restate the claim.

Infeasible Claim Generation Prompt

Input. Generate a plausible-but-false (**infeasible**) claim. You are given: clinical trial: full document, target excerpt, source claim; table: table, source claim; chart: image, source claim; and a reasoning graph with nodes **O** (observations: concrete facts), **C** (context: structural relationships), and **I** (interpretation: how structure supports conclusions).

Core idea. *Preserve observations:* keep all entities, values, labels, and visible evidence unchanged; the claim must remain grounded in the same observations as the source claim. *Corrupt reasoning:* introduce a minimal but incorrect structural interpretation; the error must come from how evidence is combined, not what is observed; the claim should remain locally plausible.

Structural falsifiability. The claim must (1) not be falsifiable by checking a single value, cell, or sentence; (2) only be falsifiable by combining multiple pieces of evidence; and (3) require integrating different regions, passages, or structures.

Domain-specific corruption. Clinical trial: errors arise from scope, conditions, populations, or trial structure. Table: errors arise from structural relationships such as grouping, units, or aggregation, and should involve multiple regions (e.g., row group + footnote + column). Chart: errors arise from visual structure such as scope, trend, grouping, or segmentation, and should involve multiple regions or patterns in the figure.

What to do. (A) Choose exactly one context (**C**) or interpretation (**I**) node to corrupt. (B) Write `corrupted_node_text` as a minimal change to the original structure; keep observations unchanged and make the error subtle and plausible. (C) Write `propagation_notes` in 1–2 sentences explaining how the corrupted structure leads to the false claim. (D) Write `infeasible_claim` in natural, domain-appropriate language; reuse the same observations as the source claim; do not mention counterevidence; do not rely on exact numeric contradictions.

Self-check (required). The claim must be (1) locally plausible using the same evidence as the source claim; (2) not falsifiable from a single observation or sentence; (3) refutable only by combining multiple pieces of evidence; (4) free of new entities, values, or external knowledge; and (5) not obviously false or ambiguous to a careful reader. If any condition fails, revise.

Disallowed. Changing or fabricating observations; explicit contradictions (numeric flip, label swap); causal or normative claims; external knowledge not in the input; trivial or single-step falsification.

H Prompts for CWA/OWA Intervention

Evaluation Prompt

Determine whether a claim is **FEASIBLE** or **INFEASIBLE** given evidence.

You are given:
- clinical: full trial + target excerpt
- table: table context
- chart: image (+ optional context)
- claim

Type 1: Baseline

Use the evidence to decide whether the claim is supported.

Type 2: OWA

You are under open-world assumption. Use the evidence to decide whether the claim is supported. Do not reject a claim based on missing evidence alone.

Type 3: CWA

You are under closed-world assumption.
- **FEASIBLE** only if all parts of the claim are explicitly supported.
- **INFEASIBLE** if any part is unsupported or contradicted.

Absence of supporting evidence counts as **INFEASIBLE**.

Type 4: DCP+OWA

You are under open-world assumption with decomposition.
STEP 1 — DECOMPOSE
Break the claim into minimal subclaims required for it to hold.

STEP 2 — CHECK CONTRADICTION
For each subclaim:
- mark **INFEASIBLE** only if explicitly contradicted
- absence of evidence does not invalidate it

STEP 3 — VERDICT
- **FEASIBLE**: no subclaim is contradicted
- **INFEASIBLE**: at least one subclaim is contradicted

Type 5: DCP+CWA

You are under closed-world assumption with decomposition.
STEP 1 — DECOMPOSE
Break the claim into minimal subclaims required for it to hold.
STEP 2 — CHECK SUPPORT
For each subclaim:
- it must be fully supported by the evidence
- unsupported or contradicted subclaims fail

STEP 3 — VERDICT
- **FEASIBLE**: all subclaims are supported
- **INFEASIBLE**: any subclaim is not supported

Output format
- (`DCP+OWA` and `DCP+CWA`) `Subclaims`: list
- `Reason`: brief explanation grounded in evidence
- `Verdict`: **FEASIBLE** or **INFEASIBLE**

I Data Generation Examples

Clinical trial example (NLI4CT).

Source. Eligibility criteria from a breast cancer clinical trial: histologically or cytologically confirmed primary invasive breast cancer; primary tumor > 2 cm; HER2 overexpression/amplification confirmed; no prior breast-cancer therapy; *only Japanese women are eligible*.

Feasible claim. Patients must have histologically or cytologically confirmed HER2-positive invasive breast cancer, with a primary tumor larger than 2 cm in diameter, to participate in the trial.

Adversarial negative. To participate in this trial for HER2-positive invasive breast cancer, women with a primary tumor larger than 2 cm in diameter are eligible for enrollment regardless of their geographic or ethnic background.

Reasoning graph.

O
O1: Primary invasive breast cancer is required.
O2: Primary tumor must be larger than 2 cm.
O3: HER2 overexpression/amplification is required.
O4: No prior breast-cancer therapy.
O5: *Only Japanese women are eligible.*

C
C1: The trial requires invasive breast cancer together with HER2-positive status.
C2: The size threshold and no-prior-therapy criterion define a treatment-naive population.
C3 (original): Eligibility is restricted to Japanese women with primary invasive breast cancer.
C3 (corrupted): Eligibility is open to women with primary invasive breast cancer regardless of nationality.

I
I1: The trial targets HER2-positive tumors receiving HER2-directed therapy.
I2: The displayed eligibility constraints characterize who may enroll in the study.

Edit type. Population/scope shift. The corruption preserves the salient disease and tumor-size constraints but removes the nationality restriction encoded in O5 and C3, yielding a globally broader eligibility claim that the source does not support.

Figure 9: Illustrative hard-negative generation for NLI4CT. A single graph node is corrupted while the surface claim remains locally plausible.

Scientific table example (SciTab).

Source. Table 3 from *Sparse and Structured Visual Attention*.

	Att. to image	Att. to boxes	TD Y/N	TD Num	TD Other	TD Overall	TS Y/N	TS Num	TS Other	TS Overall
softmax	✓		83.08	42.65	55.74	65.52	83.55	42.68	56.01	65.97
sparsemax	✓		83.08	43.19	55.79	65.60	83.33	42.99	56.06	65.94
soft-TVmax	✓		83.13	43.53	56.01	65.76	83.63	43.24	56.10	66.11
sparse-TVmax	✓		83.10	43.30	56.14	65.79	83.66	43.18	56.21	66.17
softmax		✓	85.14	49.59	58.72	68.57	85.56	49.54	59.11	69.04
sparsemax		✓	85.40	50.87	58.67	68.79	85.80	50.18	59.08	69.19
softmax	✓	✓	85.33	50.49	58.88	68.82	85.58	50.42	59.18	69.17
sparse-TVmax	✓	✓	85.35	50.52	59.15	68.96	85.72	50.66	59.22	69.28

Feasible claim. Moreover, the model using TVMAX in the final attention layer achieves the highest accuracy, showing that features obtained using the TVMAX transformation are a better complement to bounding box features.

Adversarial negative. The automatic evaluation on VQA-2.0 demonstrates that utilizing the TVmax transformation in the final attention layer consistently yields the highest Test-Standard Overall accuracy across all experimental configurations, regardless of whether the model utilizes attention to images, bounding boxes, or both simultaneously.

Reasoning graph.

O O1: The *sparse-TVmax* row with both image and bounding-box attention has the highest Test-Standard Overall score (69.28).
O2: The *soft-TVmax* row also uses TVmax, but only with image attention, and scores 66.11.
O3: The *softmax* row with both image and bounding-box attention scores 69.17.
O4: The caption states that *sparse-TVmax* and *soft-TVmax* differ in the self-attention mechanism paired with TVmax.
O5: The attention columns specify whether each row uses image attention, bounding-box attention, or both.

C C1: Within the both-attention setting, *sparse-TVmax* slightly outperforms *softmax*.
C2: TVmax rows are not directly interchangeable because they differ in both self-attention pairing and input setting.
C3: The highest score is achieved by one specific TVmax-based configuration, not by every TVmax configuration.

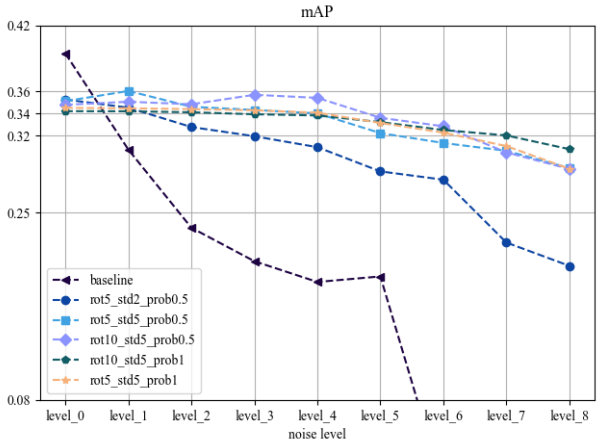
I **I2 (original):** A configuration using TVmax in the final attention layer achieves the highest accuracy among the listed configurations.
I2 (corrupted): Configurations using TVmax in the final attention layer consistently achieve the highest accuracy across all attention settings.

Edit type. Configuration-to-general overreach. A configuration-specific best result is promoted into a universal claim about all TVmax settings.

Figure 10: Illustrative hard-negative generation for SciTab. The evidence table is shown in full, followed by the reasoning graph and the single-node corruption that turns a configuration-specific conclusion into an unsupported generalization.

Scientific chart example (SciVer).

Source.



Feasible claim. At noise level_3, the `rot10_std5_prob0.5` model reaches an mAP of approximately 0.37, which is about 0.15 higher than the baseline's mAP of around 0.22.

Reasoning graph.

O O1: The legend identifies the `rot10_std5_prob0.5` series and the `baseline` series.
O2: The x-axis shows noise levels `level_0` through `level_8`, and the y-axis reports mAP.
O3: At `level_3`, the `rot10_std5_prob0.5` point is around 0.36–0.37.
O4: At `level_3`, the `baseline` point is around 0.21–0.22.
O5: The vertical separation between the two points at `level_3` is visually substantial.

C C1: Using the legend and axis labels, the chart supports a direct local comparison between the two models at `level_3`.
C2: At `level_3`, `rot10_std5_prob0.5` is clearly above `baseline` by roughly 0.15 mAP.
C3: This evidence directly supports a pointwise statement about `level_3`, but does not by itself establish the same relation across the full noise axis.

I **I1 (original):** The chart supports a local claim that `rot10_std5_prob0.5` substantially outperforms `baseline` at `level_3`.
I1 (corrupted): The chart supports treating the substantial advantage observed at `level_3` as representative of performance across most of the plotted noise spectrum.

Adversarial negative. Across the plotted noise spectrum, the `rot10_std5_prob0.5` model maintains a substantial and consistent mAP advantage over the baseline similar to the gap observed at `level_3`, and thus `rot10_std5_prob0.5` outperforms the baseline across most noise levels shown.

Edit type. Local-to-global overreach. A valid local comparison at one noise level is promoted into a broader claim about performance across most of the figure, without sufficient visual support.

Figure 11: Illustrative hard-negative generation for SciVer. A single interpretation-node corruption turns a supported local reading at `level_3` into an unsupported global claim about performance across the plotted noise spectrum.