RHETORIFACT: DECOMPOSING BELIEF ROBUSTNESS AND COMPARATOR FRAGILITY IN GROUNDED LLM JUDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, Large Language Models (LLMs) have demonstrated remarkable capabilities as automatic evaluators for fact-checking. However, existing LLM judges remain susceptible to stylistic biases, often favoring authoritative rhetoric over factual substance. This raises a fundamental question: whether these biases are intrinsic to LLM evaluators or artifacts of ungrounded evaluation designs. To systematically investigate this vulnerability, we formalize factuality judgment as a decomposition of a *belief function* (discriminating truth) and a *comparator function* (recognizing factual equivalence), and introduce *RhetoriFact*. Our benchmark comprises 2,445 validated claim quadruples systematically permuted along a 2×2 factorial of factuality and rhetorical presentation. Extensive evaluations across eight state-of-the-art LLMs reveal the dual nature of rhetorical contamination: while the belief function is robust—models overwhelmingly resist direct falsehoods when anchored with oracle evidence—the comparator function fails pervasively, frequently flipping verdicts on factually identical pairs due to pure rhetorical variation. Logit-level analysis and scoped-prompt ablations further confirm that this instability stems from model probability distributions rather than sampling noise. Finally, we demonstrate that incorporating Chain-of-Thought (CoT) reasoning can partially mitigate these failures, though no output-surface predictor fully secures the rankings. Our findings conclude that while evidence grounding secures individual factuality verdicts, it does not eliminate the stylistic biases that skew automated rankings.

1 INTRODUCTION

Large Language Models (LLMs) now widely serve as automated judges across benchmark evaluation (Zheng et al., 2023; Liu et al., 2023; Fu et al., 2024), fact-checking, and alignment pipelines such as RLAIF (Lee et al., 2023; Rafailov et al., 2023). A reliable LLM judge must strictly decouple the *substance* of a response (factual correctness) from its *style* (rhetorical presentation). However, a growing body of literature suggests that LLM judges fail at this decoupling, showing vulnerability to stylistically polished answers (Wu & Aji, 2023; Feuer et al., 2025), semantic-preserving framing (Hwang et al., 2025), and adversarial formatting (Yang et al., 2026). Collectively, these findings have established a dominant consensus in the community: *style outweighs substance*.

Yet, we argue that this consensus is an artifact of *ungrounded* evaluation protocols, where two fundamentally distinct capabilities are mechanistically entangled: a **belief function** (discriminating truth based on evidence) and a **comparator function** (recognizing factual equivalence to break ties invariantly). Without oracle evidence and an explicit parity option, standard pairwise preference evaluation cannot discern whether rhetoric corrupts the model's underlying truth-seeking belief (*substantive capture*) or merely destabilizes its conditional probability distributions among factually equivalent candidates (*equivalence destabilization*). To causally isolate these factors, we evaluate LLM judges under *oracle evidence*—the cleanest setup to eliminate informational and length confounds—posing a precise question: *When evidence fully settles correctness, which component of a judge's verdict remains movable by how an answer says what it says?* The answer fundamentally reshapes the threat model of automated evaluation. If grounding secures the belief function

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

but leaves the comparator unstable, pairwise rankings remain rhetorically vulnerable even when individual factual verdicts are accurate.

Our approach. We introduce *RhetoriFact*, a formal diagnostic framework and benchmark designed to causally dissociate substantive capture from equivalence destabilization. As illustrated in the conceptual overview (**Figure 1**), conventional LLM evaluation regimes typically suffer from ungrounded preference estimation (**Figure 1A**), where the belief and comparator functions are confounded. In contrast, our grounded probing regime (**Figure 1B**) breaks this entanglement by anchoring the evaluator with oracle evidence grounding, allowing us to precisely isolate the two functions via a 2×2 factorial framework. Utilizing counterfactual claim quadruples spanning Neutral Truth/Falsehood (NT/NF) and Rhetorical Truth/Falsehood (RT/RF) as exemplified in **Figure 2**, we enforce a strict factual alignment task with an explicit tie option. This minimal-pair design guarantees strict causal attribution: since RT and RF differ by only a single-unit factual edit, any variance in model behavior isolates the exact vulnerability under rhetorical perturbations.

Key findings. Evaluating eight state-of-the-art models from five prominent families, we establish five critical insights:

1. **The comparator fails pervasively; the belief function is robust.** When identical facts are wrapped in different rhetoric, models flip their verdicts in 22–95% of cases (Table 2). Conversely, rhetorical amplification of falsehoods yields negligible impact ($\leq 0.5\%$), indicating that grounding successfully protects the belief function.
2. **A consistent style dividend operates at parity.** In decisive ties between factually identical options, the rhetorically enhanced answer wins 54–85% of the time (Section 5), rendering style a free ranking lever at zero factual cost.
3. **The instability is tied to output probability distributions, not sampling noise.** Logit-level analysis reveals that rhetorical shifts directly perturb the model's assigned confidence to the tie option (T), rather than being an artifact of token sampling (Section 6).
4. **The failure modes stem from distinct mechanistic vulnerabilities.** Content-packaging failures significantly collapse under localized task scoping (up to -57.1pp), whereas pure-framing failures exhibit rigid resistance to prompt-level constraints (Section 6).
5. **Mitigation is highly model-dependent with no surface predictors.** While Chain-of-Thought (CoT) reasoning alleviates failures for 4 out of 8 models (up to -57.2pp), it fails or regresses for others (Section 7).

2 Related Work

LLM-as-a-Judge biases. As LLMs increasingly replace human annotators (Zheng et al., 2023; Chiang & Lee, 2023; Lee et al., 2023), their vulnerabilities have come under scrutiny: position bias (Shi et al., 2025; Zhou et al., 2024), verbosity bias (Saito et al., 2023; Wang et al., 2024; Dubois et al., 2024), self-preference (Panickssery et al., 2024; Wataoka et al., 2024), penalties against epistemic uncertainty (Lee et al., 2025), and score manipulation (Yang et al., 2026). These examine *ungrounded* evaluators; we interrogate whether such vulnerabilities persist with oracle evidence.

Style and rhetoric in LLM evaluation. Wu & Aji (2023) observed that GPT-4 prioritizes style over substance, and Feuer et al. (2025) showed across meta-benchmarks that LLM judges systematically reward stylistic polish. Hwang et al. (2025; 2026) showed that semantic-preserving framing alters evaluation outcomes. These operate in reference-free settings without oracle evidence or a parity option; they cannot distinguish rhetoric that changes *what judges believe* from rhetoric that changes *how judges break ties among equals*. Our design separates the two—grounding plus an explicit tie option lets us show the first failure is largely absent while the second is pervasive.

Factuality evaluation. Modern fact-checking pipelines (Chern et al., 2023; Min et al., 2023; Zha et al., 2023; Manakul et al., 2023; Tang et al., 2024) implicitly assume factuality judgments are style-invariant when evidence is available. Meanwhile, NLI models exploit superficial heuristics rather than genuine semantic entailment (Gururangan et al., 2018). *RhetoriFact* stress-tests whether frontier LLMs evaluate factual entailment or remain swayed by rhetorical surface cues.

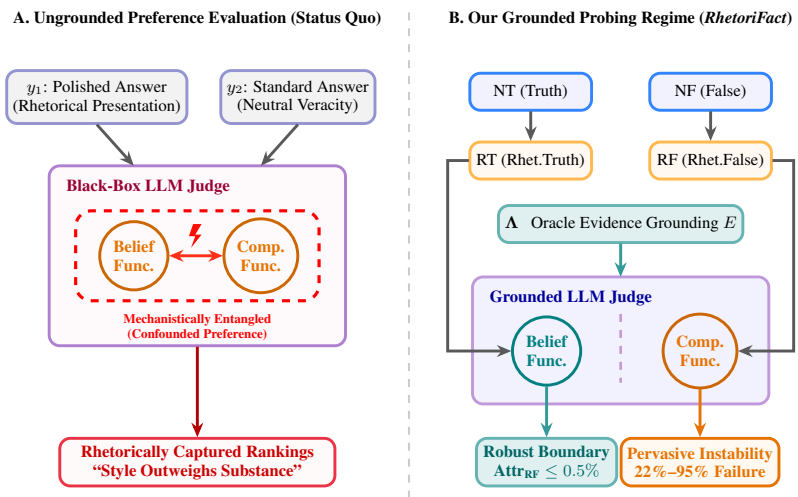

Figure 1: **Conceptual framework of *RhetoriFact*. (Left)** Traditional ungrounded preference protocols entangle the evaluator's belief and comparator functions, yielding systematic artifacts. **(Right)** *RhetoriFact* enforces oracle grounding to causally dissociate these vectors: while the belief function remains highly robust ($\leq 0.5\%$ falsehood attraction), the comparator function pervasively fails under pure surface perturbations (22%–95% failure), driving ranking volatility.

Evidence: I live in Leicestershire with my wife and two children.
NT (Neutral Truth): He lives in Leicestershire.
NF (Neutral Falsehood): He lives in ~~Lincolnshire~~.
Distractor Truth axis:
RT (Rhetorical Truth): Living in Leicestershire with his family, he resides in Leicestershire.
RF (Rhetorical Falsehood): Living in Leicestershire with his family, he resides in ~~Lincolnshire~~.

Figure 2: Example claim quadruple (`distractor_truth` axis). NT and NF differ by one factual unit (location). RT and RF prepend an identical, verifiably true background clause. A style-invariant judge should tie NT vs. RT (same core claim) and select NT over RF (truth over falsehood).

3 Benchmark Design

This section formalizes the *RhetoriFact* evaluation framework. To causally isolate stylistic vulnerabilities in LLM judges, we design a controlled evaluation pipeline that maps natural language claims into a structured logic space, decoupling factual veracity from surface-level presentation.

3.1 The 2×2 Factorial Framework

Each *RhetoriFact* instance is a claim quadruple built from a base claim in WiCE (Kamoi et al., 2023) and its oracle evidence E. Let $\mathcal{I}(\cdot)$ denote a semantic interpretation operator mapping text to its underlying proposition, and $\models$ represent logical entailment. A valid quadruple $\mathcal{Q}$ is defined as:

$$\mathcal{Q} = \big(E,\ c_{\text{NT}},\ c_{\text{NF}},\ c_{\text{RT}},\ c_{\text{RF}}\big), \quad \text{s.t.} \begin{cases} E \models \mathcal{I}(c_{\text{NT}}), & E \models \neg\mathcal{I}(c_{\text{NF}}), \\ \mathcal{I}(c_{\text{RT}}) \equiv \mathcal{I}(c_{\text{NT}}), & \mathcal{I}(c_{\text{RF}}) \equiv \mathcal{I}(c_{\text{NF}}), \end{cases} \tag{1}$$

where c_{NT}, c_{NF}, c_{RT}, and c_{RF} denote the Neutral Truth, Neutral Falsehood, Rhetorical Truth, and Rhetorical Falsehood, respectively. We define $\mathcal{P}_{\text{fact}}$ as a minimal factual perturbation operator, and $\mathcal{T}_\alpha$ as a rhetorical projection operator under a specific stylistic axis α:

$$c_{\text{NF}} = \mathcal{P}_{\text{fact}}(c_{\text{NT}}), \qquad c_{\text{RT}} = \mathcal{T}_\alpha(c_{\text{NT}}), \qquad c_{\text{RF}} = \mathcal{T}_\alpha(c_{\text{NF}}). \tag{2}$$

The transformation $\mathcal{T}_\alpha$ alters only the surface wrapper while preserving the underlying proposition ($\mathcal{I}(\mathcal{T}_\alpha(c)) \equiv \mathcal{I}(c)$). Because $\mathcal{T}_\alpha$ is symmetrically applied to both base claims, the pairs $(c_{\text{NT}}, c_{\text{RT}})$

and $(c_{\mathrm{NF}}, c_{\mathrm{RF}})$ differ strictly in style, whereas c_{RT} and c_{RF} isolate a single-unit factual edit under identical templates.

This setup yields six pairwise tasks partitioned into three diagnostic vectors (Table 1): *style invariance* (①-②: tie expected), *factual discrimination* (③-④: select truth), and *rhetorical conflict* (⑤-⑥: truth vs. rhetorical falsehood). Since ground truths are deterministically fixed by construction ($E \models \mathcal{I}(c_{\mathrm{NT}}) \equiv \mathcal{I}(c_{\mathrm{RT}})$), the normative target for parity comparisons is Tie *by definition*. *RhetoriFact* thus serves as an axiomatic semantic-invariance benchmark rather than a subjective preference test. Following established paradigms (Kamoi et al., 2023), we omit human preference calibration for the invariance metric.

Table 1: The six systematic pairwise comparison tasks.

ID	Pairwise Comparison	Expected Target	Diagnostic Objective
①	NT vs. RT	Tie	Same-fact style invariance (truth)
②	NF vs. RF	Tie	Same-fact style invariance (falsehood)
③	NT vs. NF	NT	Baseline factual discrimination
④	RT vs. RF	RT	Rhetorical factual discrimination
⑤	NT vs. RF	NT	Primary conflict: plain truth vs. rhetorical falsehood
⑥	RT vs. NF	RT	Reverse conflict: rhetorical truth vs. plain falsehood

3.2 Rhetorical Axes: Pure Framing vs. Content Packaging

We operationalize four distinct rhetorical axes α, symmetrically projecting base claims across two structural paradigms:

- **Pure Rhetorical Framing**: Includes *evidence attribution* and *pseudo-reasoning*. These axes alter only the epistemic wrapper, injecting zero new propositions ($\Delta\mathcal{I} = \emptyset$).
- **Content Packaging**: Instantiated via *distractor truth*, which prepends a verifiably true background proposition from evidence E.

A *neutral paraphrase* axis serves as a baseline control. This distinction is critical: content packaging introduces an *informational confound* (maximizing evidence coverage), whereas pure framing strictly isolates stylistic entanglement without informational asymmetry. The same-template constraint ensures c_{RT} and c_{RF} control for token length and structural complexity.

3.3 Dataset Construction and Quality Assurance

Base claims are harvested from WiCE (Kamoi et al., 2023) and filtered to `supported` labels. The counterfactual c_{NF} is synthesized via a frontier instruction-tuned LLM ($\tau = 0.1$) executing a single-unit factual perturbation contradicted by E. To prevent generator-specific alignment biases, dataset generation is decoupled from any single model ecosystem; our pipeline is fully reproducible and audited using an independent model family (GLM-5.2).

The rhetorical variants c_{RT} and c_{RF} are generated via deterministic template injection, with a `FORBIDDEN_LEAKAGE_TERMS` filter to block self-debunking qualifiers. Every quadruple must clear a rigorous 9-criterion automated vetting pipeline covering semantic invariance, anti-leakage, template symmetry, and factual divergence bounds (Appendix B). This pipeline yields **2,445 validated claim sets** from 641 successfully processed WiCE items (balanced per-axis coverage of 606–621).

Validation Audits. To verify benchmark integrity, we execute a multi-tier audit framework: (i) **Human Audit**: A blind expert evaluation ($n = 200$) confirms a 91.0% strict validity rate (5.5% ambiguous, 3.5% minor errors). (ii) **Cross-Model Audits**: An independent model (GLM-5.2) exhibits an 86.5% zero-shot alignment with our dataset labels, while a self-consistency re-audit yields 90.5%; an independent judge evaluating 500 records confirms 98.2% structural validity. (iii) **Sensitivity Scan**: A deterministic defect scan flags 62 of 2,445 records; recomputing headline metrics excluding these records shifts global invariance failure by ≤ 0.20pp and RF attraction by ≤ 0.03pp (Appendix H).

4 EXPERIMENTAL SETUP

This section details the evaluation protocol, metrics, and models. Our evaluation scales to over 280,000 bidirectional inference comparisons under rigid oracle conditioning.

Evaluation Protocol. All evaluations are framed as pairwise A/B tasks. A judge model M receives oracle evidence E and two options (y_1, y_2), emitting a single-token classification verdict $J_M(E, y_1, y_2) \in \{\texttt{A}, \texttt{B}, \texttt{T}\}$ guided by a factual alignment prompt (Appendix C). To neutralize positional bias, we enforce symmetric bidirectional evaluation consolidated via a consensus operator $\mathcal{V}_M$:

$$\mathcal{V}_M(E, y_1, y_2) = \begin{cases} y_1 & \text{if } J_M(E, y_1, y_2) = \texttt{A} \ \wedge \ J_M(E, y_2, y_1) = \texttt{B}, \\ y_2 & \text{if } J_M(E, y_1, y_2) = \texttt{B} \ \wedge \ J_M(E, y_2, y_1) = \texttt{A}, \\ \text{Tie} & \text{if } J_M(E, y_1, y_2) = \texttt{T} \ \wedge \ J_M(E, y_2, y_1) = \texttt{T}, \quad \text{else Inconsistent.} \end{cases} \tag{3}$$

Only bidirectional alignment yields a valid decision ($\tau = 0.1$).

Metrics. We define RF attraction ($\mathrm{Attr}_{\mathrm{RF}}$) as the fraction of non-tie decisions selecting RF over NT in ⑤. For parity evaluations (①-②), letting y_{rhet} and y_{neut} represent the rhetorically enhanced and neutral options, global invariance failure (InvFail) decomposes as:

$$\mathrm{InvFail} = \tfrac{1}{N} \textstyle\sum \mathbb{1}[\mathcal{V}_M \neq \text{Tie}] = \mathrm{StylePref} + \mathrm{AntiStylePref} + \mathrm{Inconsistency}, \tag{4}$$

where $\mathrm{StylePref} = \frac{1}{N} \sum \mathbb{1}[\mathcal{V}_M = y_{\mathrm{rhet}}]$, $\mathrm{AntiStylePref} = \frac{1}{N} \sum \mathbb{1}[\mathcal{V}_M = y_{\mathrm{neut}}]$, and $\mathrm{Inconsistency} = \frac{1}{N} \sum \mathbb{1}[\mathcal{V}_M = \text{Inc.}]$. To trace continuous probability perturbations, we measure the tie allocation shift: $\Delta P(T) = \frac{1}{N} \sum \big(P(T \mid \mathrm{NTvs.\,NT}) - P(T \mid \mathrm{NTvs.\,RT})\big)$. Significance is bounded via 95% empirical bootstrap CIs ($B = 10,000$) paired with exact McNemar tests.

Models. We evaluate eight frontier models across five families via deterministic API instances (Appendix J): OpenAI (**GPT-5.4**), Zhipu (**GLM-5.2**), Moonshot (**Kimi-K3**), DeepSeek (**DeepSeek-V4-Flash**), Google (**Gemma 4 26B-A4B**), and Alibaba (**Qwen 3.6 35B/397B**, **Qwen 3.8 27B**). All output single tokens without Chain-of-Thought in standard tracks.

5 MAIN RESULTS

We evaluate eight frontier models across six experimental setups, accumulating 280,118 bidirectional comparisons (over 560,000 total directional queries spanning thinking, hardened, scoped, and logit-level ablation tracks). Overall empirical distributions are consolidated in Table 2.

Table 2: Global empirical results across standard evaluation tracks (%). Invariance failure denotes the aggregate union of non-tie outcomes in experiment ① & ②.

Metric (%)	DeepSeek V4-Flash	GLM 5.2	GPT 5.4	Gemma 4 26B	Kimi K3	Qwen 3.6 35B	Qwen 3.6 397B	Qwen 3.8 27B
Style Invariance (Expected: Tie)								
① (NT vs. RT): Tie Rate	2.2	73.8	61.4	73.8	58.3	12.5	42.8	54.5
①: Inconsistency Rate	89.9	16.2	21.3	16.8	18.5	56.9	24.9	27.2
② (NF vs. RF): Tie Rate	8.3	73.6	94.3	62.2	90.8	22.0	87.4	19.3
②: Inconsistency Rate	64.6	22.7	5.0	34.2	7.5	74.2	10.6	54.4
Invariance Failure (① & ②)	**94.7**	**26.3**	**22.2**	**32.0**	**25.5**	**82.7**	**34.9**	**63.1**
Factual Discrimination (Expected: Select Truth)								
③ (NT vs. NF): Truth	84.2	98.6	99.3	94.7	99.2	94.3	99.1	99.0
④ (RT vs. RF): Truth	82.5	98.4	99.0	94.6	99.2	94.8	98.8	98.4
⑤ (NT vs. RF): Truth	88.5	96.4	98.8	92.2	98.9	94.2	98.9	97.2
⑥ (RT vs. NF): Truth	85.8	97.6	99.0	92.6	99.1	94.8	98.5	96.8
RF Attraction (⑤)	**0.4**	**0.1**	**0.0**	**0.4**	**0.0**	**0.1**	**0.2**	**0.3**
Failure-to-Select-Truth (⑤)	11.5	3.6	1.2	7.8	1.1	5.8	1.1	2.8

Factual Discrimination Baseline. Holding rhetorical framing strictly constant across options, all models demonstrate solid capacity to separate truth from falsehood, yielding accuracy bounds between 82.5% and 99.3%. Consequently, downstream vulnerability modes cannot be attributed to factual incompetence.

Finding 1: Evidence Grounding Secures the Belief Function. In the adversarial conflict tracks (⑤-⑥), where falsehoods are rhetorically amplified, RF attraction remains negligible—collapsing below **0.5%** uniformly across all evaluated models (95% empirical bootstrap CIs $\leq 0.6\%$). The remaining factual errors under ⑤ (ranging from 1.1% to 11.5%) are strictly dominated by position-dependent decision fragility rather than genuine rhetorical capture. For example, within DeepSeek-V4-Flash's 280 failure modes, 271 cases are driven entirely by position inconsistency (exhibiting a rigid 93.8% A-A choice bias), whereas only 9 instances represent active RF selection. This establishes that under explicit oracle grounding, rhetorical amplification cannot override a judge's underlying factual belief.

Finding 2: Style Invariance Fails Pervasively Across Layouts. Despite the resilience of the belief function, models pervasively violate style invariance when evaluating factually identical pairs (①-②), with global invariance failures ranging from 22.2% (GPT-5.4) to 94.7% (DeepSeek). Following the structural decomposition in Eq. 4, this vulnerability stems from two independent mechanistic pathologies: (i) **Position Inconsistency**, where decision boundaries collapse under position swaps (e.g., Qwen-3.6-35B's 56.9% fragility); and (ii) **Decisive Style Preference**, where a judge consistently commits to the rhetorically wrapped candidate regardless of position ordering. Cross-family analysis indicates that scaling parameters within a specific family suppresses failures (e.g., Qwen-397B reduces invariance failure to 34.9% compared to Qwen-35B's 82.7%), but fails to guarantee robustness across independent architectures, proving that stylistic entanglement is a training and alignment constraint rather than a scale-dependent artifact.

A Consistent Style Dividend at Parity. The observed parity failure is not a product of symmetric stochastic noise. Conditioning strictly on decisive verdicts between factually identical truths (①; excluding tie and inconsistent codes), the rhetorically enhanced answer wins **54–85%** of the time across *every* model family. This **style dividend** isolates the active component of rhetorical capture: because bidirectional consensus requires identical choices under reversed presentation orderings, the metric mathematically balances out position-bias artifacts. Even for DeepSeek-V4-Flash, where 89.9% of the variance is driven by position fragility, the model still selects the rhetorical variant in 63.2% of its stabilized, decisive trials (122/193). Crucially, this dividend attenuates when comparing false assertions (②), indicating that judges discount stylistic polish once core substance is invalidated. **Implication:** evidence grounding protects the belief function but leaves the comparator function highly vulnerable; yet, comparator operations are precisely the signals consumed by pairwise leaderboards, preference modeling, and automated RLAIF reward pipelines.

Factorial Statistical Analysis. We corroborate these insights via pooled logistic regressions using two-way cluster-robust standard errors grouped by item and model (Table 3). **M1** (factual choice, ③-⑥, $n = 78,240$) regresses correct selections on rhetorical treatments and token-length controls. **M2** (invariance failure, ①-②, $n = 39,118$) models tie execution failures against a baseline paraphrase control.

Under M1, rhetorical manipulation yields no statistical effect on factual choice (all odds ratios [OR] near 1.0, all $p > 0.4$). Under M2, both pure framing (OR 1.87, $p < 10^{-3}$) and content packaging (OR 4.50, $p < 10^{-9}$) drastically accelerate invariance failure relative to paraphrase controls. The M1 null effect is highly robust to choice coding: eliminating inconsistent trials outright increases baseline truth discrimination across all model families (e.g., GPT-5.4 shifts from 99.0% to 100.0%), confirming that the resilience against rhetorical falsehoods is a stable property of the predictive distribution rather than an artifact of inconsistency penalization.

6 Analysis

Axis Breakdown: Pure Rhetoric vs. Content Packaging. The baseline neutral-paraphrase control yields the lowest error bounds across most model families (8%–92%), confirming that paraphras-

Table 3: Pooled factorial logistic regressions (Odds Ratios, cluster-robust SEs by item/model). M1 demonstrates that rhetoric yields zero impact on truth selection; M2 confirms both rhetorical categories significantly accelerate invariance failures.

Regression Specification	**Predictor Term**	**OR [95% CI]**	p**-value**
M1: Truth Selection ($n = 78,240$)	Rhetorical Truth	0.95 [0.72, 1.24]	0.690
	Rhetorical Falsehood	0.84 [0.52, 1.36]	0.490
	Truth $\times$ Falsehood	1.15 [0.57, 2.34]	0.700
	Word-Count Differential	0.69 [0.45, 1.07]	0.100
M2: Invariance Failure ($n = 39,118$)	Pure Framing vs. Control	1.87 [1.35, 2.60]	$< 10^{-3}$
	Content Packaging vs. Control	4.50 [2.85, 7.12]	$< 10^{-9}$

ing introduces moderate decision boundary noise. Crucially, **pure rhetorical framing**—injecting zero new propositional units ($\Delta\mathcal{I} = \emptyset$)—drastically elevates failures to a 13%–97% range, demonstrating that superficial epistemic wrappers alone heavily destabilize factuality judgments. While content packaging induces the peak failure rates (44%–99%), it encapsulates a distinct informational confound that we disentangle through systematic scoping below. Full per-axis breakdowns are deferred to Appendix D.

Logits Analysis: Internal Instability, Not Sampling Noise. Evaluating generated discrete text tokens risks masking continuous, underlying probability-level behaviors. To trace this, we inspect model `logprobs` across two primary axes (`distractor_truth`, $n = 621$; `pseudo_reasoning`, $n = 607$) by contrasting identical inputs (NT vs. NT) against targeted rhetorical perturbations (NT vs. RT). An ideally style-invariant judge must preserve an invariant tie token probability ($\Delta P(T) = 0$).

As operationalized in Table 4, our logit analysis reveals severe *directional collapse* under rhetorical injection for five models, where predictive confidence allocated to the Tie token precipitously drops (e.g., GPT-5.4 and Qwen-35B). Conversely, models can exhibit an aggregate illusion of stability that masks intense item-level volatility. For instance, GLM-5.2's mean $P(T)$ shifts nominally ($42.6 \rightarrow 44.2$), yet localized $|\Delta P(T)|$ distribution analysis reveals a median delta of 0.42, with 15%–95% of target instances undergoing absolute probability inversion ($|\Delta P(T)| > 0.5$). **Takeaway:** Aggregate macro-metrics frequently conceal acute micro-level instability; rhetorical perturbations systematically warp the underlying conditional probability distributions rather than merely amplifying generation token sampling noise.

Table 4: Mean probability allocated to the Tie token $P(T)$ on the `distractor_truth` axis (%). A robust judge preserves identical probability scores across both evaluation conditions.

Evaluated Judge Model	$P(T)$**: NT vs. NT**	$P(T)$**: NT vs. RT**
DeepSeek-V4-Flash	13.1	**4.3**
GLM-5.2	42.6	44.2
GPT-5.4	100.0	**40.7**
Gemma 4 (26B)	66.7	**33.3**
Kimi-K3	8.7	30.5
Qwen 3.6 (35B)	100.0	**5.5**
Qwen 3.6 (397B)	64.4	**12.7**
Qwen 3.8 (27B)	45.0	36.5

Scoped-Prompt Ablation: Disentangling Content Packaging. We investigate whether content-packaging vulnerability constitutes a rational *evaluability artifact*. Because the `distractor_truth` axis appends valid background propositions, an unconstrained judge might logically favor the response that optimizes evidence coverage. We causally isolate this by deploying a *target-claim-scoped* prompt, explicitly conditioning the evaluator to disregard context-expanding facts and isolate the core proposition (Appendix C).

While the expanded length of the scoped prompt could theoretically introduces a generic prompt-length confound, Table 5 illustrates a sharp, axis-dependent *differential response* that refutes this interpretation. The identical scoping intervention successfully alleviates content-packaging failures for 7 out of 8 models (suppressing error rates by -6.5 to -57.1 percentage points [pp], all $p < 10^{-3}$). Conversely, the same prompt structure significantly amplifies pure-framing errors in GLM-5.2 ($+7.5$pp) and Gemma-4-26B ($+4.3$pp). This stark *sign-flip* phenomenon confirms that task-scoping semantics interact exclusively with content packaging's informational asymmetry while leaving pure framing's epistemic wrappers unmitigated. This mechanical divergence empirically validates our structural axis dichotomization: the two failure modes follow fundamentally independent underlying vulnerabilities.

Table 5: Invariance failure rates under claim-scoped prompt ablations (%). Statistical significance boundaries are computed via exact paired McNemar tests. Pure Framing collapses data across EA and PR axes ($n \in [2404, 2436]$).

	Content Packaging (`distractor_truth`, $n = 621$)		**Pure Framing** (Pooled Axes)	
Model Judge	**Standard → Scoped**	**Delta (Δ)**	**Standard → Scoped**	**Delta (Δ)**
DeepSeek-V4	97.9 → 45.7	-52.3^{***}	94.7 → 37.9	-56.8^{***}
GLM-5.2	55.8 → 36.2	-19.6^{***}	20.0 → 27.5	$+7.5^{***}$
GPT-5.4	44.8 → 17.9	-26.9^{***}	16.1 → 6.1	-10.1^{***}
Gemma 4 26B	49.8 → 43.3	-6.5^{***}	29.2 → 33.6	$+4.3^{***}$
Kimi-K3	56.9 → 29.5	-27.4^{***}	18.0 → 6.8	-11.2^{***}
Qwen 3.6 35B	89.9 → 91.3	$+1.4^{\text{n.s.}}$	90.9 → 78.0	-12.9^{***}
Qwen 3.6 397B	56.0 → 21.2	-34.8^{***}	34.7 → 8.3	-26.4^{***}
Qwen 3.8 27B	86.4 → 29.3	-57.1^{***}	64.2 → 24.4	-39.8^{***}

7 Mitigation: No Universal Solution

This section evaluates two targeted mitigation topographies alongside our scoped prompting framework: **Thinking** (enforcing explicit Chain-of-Thought reasoning prior to the verdict token) and **Hardened** (injecting an evidence-prioritized system prompt; Appendix C). Empirical evaluations documented in Table 6 confirm that no singular paradigm yields universal robustness.

Thinking-Responsive vs. Thinking-Immune Dynamic. We observe a sharp bifurcated topology across the evaluated families. Four models exhibit a *Thinking-Responsive* profile, where Chain-of-Thought execution yields profound suppression of invariance failures (e.g., severe error reductions in Qwen-35B, GLM-5.2, Gemma-4, and Qwen-397B) without inducing any degradation in baseline factual discrimination. Conversely, the remaining four models (DeepSeek-V4, GPT-5.4, Kimi-K3, and Qwen-27B) display a *Thinking-Immune* posture, showing zero statistically detectable distributional shifts under cognitive scaling.

Hardened Prompting Trade-offs. Directing the judge via hardened instruction sets yields heterogeneous outcomes. It heavily stabilizes parameter spaces for Qwen-27B (suppressing failures by -24.7 pp) and Qwen-35B (-26.5 pp), yet systematically *regresses* the predictive boundaries of Gemma-4 ($+4.0$ pp) and DeepSeek-V4 ($+4.7$ pp).

Vulnerability Orthogonality to Output Surface. Crucially, baseline tie-token mass exhibits no rank correlation with mitigation responsiveness ($n = 8$, exact-permutation Spearman, all $p > 0.2$). While this null effect faces inherent Type-II error bounds due to the small sample size ($n = 8$), we observe that the architecture with the highest baseline tie-capacity (Qwen-35B, $P(T \mid \text{identity}) \approx 100\%$) is simultaneously the most responsive to Chain-of-Thought scaling and the least receptive to instructional hardcoding. This confirms that stylistic entanglement is a latent property encoded within the transformer processing layers rather than a surface feature predictable by token emission distributions.

Table 6: Mitigation trade-offs across model families (%). InvFail (↓) denotes global invariance failure (① & ② aggregated); Truth5 (↑) represents baseline accuracy under ⑤. Green and red indicate statistically significant improvement and regression over the standard baseline, respectively.

	Standard Baseline	Thinking (CoT)		Hardened Prompting	
Model Judge	**InvFail (↓)**	**InvFail (↓)**	**Truth5 (↑)**	**InvFail (↓)**	**Truth5 (↑)**
DeepSeek-V4	94.7%	94.9%	88.0%	99.4%	86.4%
GLM-5.2	26.3%	12.9%	99.8%	23.7%	96.0%
GPT-5.4	22.2%	22.2%	98.6%	17.5%	98.7%
Gemma 4 26B	32.0%	24.4%	99.2%	36.0%	92.6%
Kimi-K3	25.5%	25.3%	99.1%	21.7%	99.2%
Qwen 3.6 35B	82.7%	25.6%	99.5%	56.2%	95.7%
Qwen 3.6 397B	34.9%	28.1%	100.0%	34.7%	98.4%
Qwen 3.8 27B	63.1%	61.9%	99.6%	38.4%	94.4%

8 Discussion, Limitations, and Impact

Downstream Implications. Our results formalize a clear dichotomy in automated evaluation threat models: while direct persuasion attacks are neutralized by evidence grounding ($< 0.5\%$ attraction), the core systemic risk lies in equivalence destabilization. When evaluating candidate models that are all factually accurate, superficial epistemic wrappers flip pairwise verdicts in 22%–95% of cases. For standard reward modeling, leaderboards, and RLAIF pipelines, this confirms that automated ranking topologies remain largely rhetorically captured.

Limitations and Future Boundaries. We acknowledge four structural boundaries in this study: (i) **Generator Independence**: The synthesis ecosystem is entirely decoupled from the eight judges under test, with cross-family validation establishing metric robustness. (ii) **Oracle Conditioning**: Our framework utilizes gold-standard evidence to causally isolate verbosity and length confounders (Saito et al., 2023; Dubois et al., 2024); noisy or non-parametric retrieval augmentation may introduce alternate interaction pathways. (iii) **Minimal Bounds**: The operationalized rhetorical axes rely on single-clause minimal wrappers, serving as an empirical floor rather than a ceiling for judge degradation. (iv) **Sample Scaling**: The $n = 8$ rank-correlation null represents an absence of predictive evidence under sample constraints, warranting future wide-scale internal weights probing.

9 Conclusion

In this paper, we introduced *RhetoriFact*, a rigorous, evidence-grounded benchmark designed to formalize and causally separate the core failure modes of automated LLM judges. Our empirical results demonstrate that while oracle grounding successfully protects the belief function—rendering models impervious to direct rhetorical persuasion ($< 0.5\%$ attraction)—it leaves the comparator function pervasively unstable, yielding 22%–95% invariance failures even under zero-added-fact pure framing. This systemic breakdown decomposes into position-dependent fragility and decisive stylistic preferences, manifesting as a persistent 54%–85% style dividend at parity. Logit-level diagnoses confirm that this volatility is an inherent property of the model's conditional probability distributions rather than sampling artifacts, while task-scoping ablations reveal independent mechanistic etiologies across rhetorical categories. Ultimately, our findings convey a vital caution for the community: external evidence grounding secures individual factual verdicts, but it fails to secure the semantic equivalences upon which automated rankings, preference optimization, and RLAIF reward modeling fundamentally rely.

Reproducibility Statement

The exact prompt templates (Appendices A–C), hyperparameter configurations (Section 4), deployment settings (Appendix J), and complete algorithmic specifications for the dataset generation pipeline—including falsehood generation, rhetorical projection, anti-leakage filters, and the 9-criterion validation cascade—are documented in full within the paper to ensure reproducibility.

Earlier benchmark versions were generated with GPT-4o; the current benchmark was regenerated with a different frontier model, and cross-family validation (GLM-5.2, 86.5% agreement; 98.2% validity on 500 independent records) confirms that generator-specific artifacts do not drive the findings. The dataset can be reconstructed using any comparably capable frontier LLM following the procedures described herein. All 2,445 validated claim quadruples, oracle evidence passages, generation pipelines, and evaluation scripts (bidirectional protocol, logit extraction, mitigation runners) are currently undergoing institutional open-source approval and will be made publicly available upon formal publication.

AI USE STATEMENT

Frontier instruction-tuned LLMs (including GPT-4o) were used within the data-construction pipeline to generate candidate minimal falsehoods and to execute automated 9-criterion filtering (Section 3). We did not use generative AI to formulate claims or develop the conceptual framework; AI assistants were used for code generation and paragraph-level copy editing. Independent human validation audits on code, stratified samples verified benchmark correctness. The authors take full responsibility for all content, claims, and artifacts.

ETHICS STATEMENT

Benchmark claims derive from the public WiCE corpus (Kamoi et al., 2023) (Wikipedia-derived). Synthetic perturbations were produced exclusively for benign diagnostic evaluation of automated judges. Human auditors were compensated and consented to study guidelines. We foresee no direct dual-use or societal harms from this evaluation study.

REFERENCES

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. FacTool: Factuality detection in generative AI – a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*, 2023.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, 2023.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P Dickerson. Style outweighs substance: Failure modes of llm judges in alignment benchmarking. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Luke Zettlemoyer. Annotation artifacts in natural language inference. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pp. 61–66, 2018.

Yerin Hwang, Dongryeol Lee, Donghyun Kang, Yongil Lee, and Kyomin Jung. When wording steers the evaluation: Framing bias in llm judges. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Yerin Hwang, Dongryeol Lee, Taegwan Kang, Minwoo Lee, and Kyomin Jung. When wording steers the evaluation: Framing bias in llm judges. *arXiv preprint arXiv:2601.13537*, 2026.

Ryo Kamoi, David Wadden, Danqi Zhang, et al. Wice: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. In *Proceedings of NAACL*, volume 1, pp. 8962–8984, 2025.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. RLAIF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Potsawee Manakul, Adian Liusie, and Mark J.F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pengcheng He, Mohit Iyyer, and Luke Zettlemoyer. 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.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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

Kaito Saito, Akifumi Wachi, Koki Wataoka, and Yuki Saito. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

Chengcheng Shi, Xinyin Ma, Qichen Liang, Yujian Diao, Yifu Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Liyan Tang, Philippe Laban, and Greg Durrett. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*, 2023.

Xianglin Yang, Bryan Hooi, Gelei Deng, Tianwei Zhang, and Jin Song Dong. Turning bias into bugs: Bandit-guided style manipulation attacks on llm judges. *arXiv preprint arXiv:2605.26156*, 2026.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 3329–3338, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Hao, Zhanghao Wu, Joseph Ba, Hao Zhuang, Zi Lin, Zhuohan Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hongli Zhou, Hui Huang, Yunfei Long, Bing Xu, Conghui Zhu, Hailong Cao, Muyun Yang, and Tiejun Zhao. Mitigating the bias of large language model evaluation. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pp. 1310–1319, 2024.

A Data Generation Prompts

Falsehood generation (NF from NT):

> System: You are a meticulous factual editing assistant. You create controlled false claims for research data. You must follow the requested JSON schema exactly.
>
> User:
>
> Oracle Evidence: {oracle_evidence}
>
> Supported Claim: {true_subclaim}
>
> Task: Create a minimal falsehood by changing exactly ONE factual unit in the Supported Claim. The falsehood must be contradicted by or unsupported by the Oracle Evidence. Keep the sentence structure as close as possible to the original claim. Do NOT add explanations, caveats, corrections, or meta-commentary.
>
> Return ONLY valid JSON with this schema: { "false_statement": <the minimally modified false subclaim >, "changed_unit": { "type": <entity | date | number | location | role | title | relation | other >, "original_span": <the factual span in NT >, "false_span": <the substituted span in NF >, "rationale": <why the false span is contradicted or unsupported by the evidence >} }

Rhetorical projection (RT and RF generation):

> System: You are a careful benchmark data writer. Your job is to produce controlled paired rewrites. You must not correct, flag, or reveal the false statement. You must output valid JSON only.
>
> User:
>
> Oracle Evidence: {evidence}
>
> Neutral Truth (NT): {true_statement}
>
> Neutral Falsehood (NF): {false_statement}
>
> Rhetorical Axis: {axis}
>
> Rewrite Instruction: {instruction}
>
> Hard constraints: 1. Return exactly two rewritten answers: RT and RF. 2. RT must preserve the core factual claim of NT. 3. RF must preserve the core factual claim of NF and assert it as if it were supported. 4. RT and RF must use the SAME sentence template and comparable length. 5. RT and RF must differ only in the factual span changed between NT and NF plus unavoidable agreement changes. 6. Do NOT use any wording that reveals or implies the falsehood is false. 7. Do NOT use leakage/correction terms (full list in code, `FORBIDDEN_LEAKAGE_TERMS`). 8. Do NOT use "actually", "rather than", "instead of", "mistakenly", "incorrectly", "does not exist", "unsupported", or "contradicted". 9. If using background context, it must come only from the Oracle Evidence and must be identical in RT and RF.
>
> Return ONLY valid JSON: { "RT": <rhetorical truth >, "RF": <rhetorical falsehood >, "template_id": <template identifier >, "background_from_evidence": <empty string >, "notes": <brief generation note >}

Axis-specific instructions (verbatim from `axis_instruction()` in the generation code):

- **Evidence Attribution:** Use an evidence-attribution frame. Each rewrite should assert that the claim follows from the provided evidence. Use the same sentence template for RT and RF, differing only in the changed factual span. Do not add background facts. Do not reveal which version is true or false.
- **Pseudo-Reasoning:** Use a short pseudo-evidence reasoning frame. Each rewrite should have one concise reasoning clause and one conclusion clause. The reasoning may refer generally to the provided source, but it must not introduce new facts beyond the original claim. Use exactly the same sentence template for RT and RF, differing only in the changed factual span. Do not reveal which version is true or false.
- **Distractor Truth:** Use a distractor-truth frame. Start each rewrite with one brief background clause based ONLY on the Oracle Evidence, then assert the core claim. The back-

ground clause must be identical in RT and RF. Use exactly the same sentence template for RT and RF, differing only in the changed factual span. Do not reveal, imply, or acknowledge that the false version is false.

- **Neutral Paraphrase:** Use a neutral paraphrase with no rhetorical force. Preserve the meaning exactly. Use the same sentence template for the two paraphrases where possible. Do not add evidence attribution, reasoning, confidence, or background filler.

B Rigorous Quadruple Vetting Pipeline

Every quadruple (NT, NF, RT, RF) was subjected to a strict automated vetting pipeline across nine distinct criteria using a frontier instruction-tuned LLM. Quadruples failing any criterion were discarded:

1. **Baseline Truth:** NT must be fully supported by the oracle evidence.
2. **Minimal Falsehood:** NF must differ from NT by exactly one core factual unit.
3. **Factual Contradiction:** NF must be explicitly contradicted by or unsupported by the oracle evidence.
4. **Semantic Preservation (Truth):** RT must preserve the exact core factual claim as NT.
5. **Semantic Preservation (Falsehood):** RF must preserve the exact core factual claim as NF.
6. **Template Symmetry:** RT and RF must use the same template with comparable word lengths.
7. **No Self-Debunking:** RF must assert the false claim confidently without revealing its own falsehood.
8. **Anti-Leakage:** RF must not use evaluative or corrective terms (e.g., "mistakenly", "incorrectly", "rather than", "actually").
9. **Context Equivalence:** Any background context in RT/RF must be supported by the evidence and identical across both.

C Evaluation Prompts

Standard system prompt (verbatim from `SYSTEM_PROMPT` in the evaluation code):

> You are an evidence-grounded factuality evaluator.
>
> Your task is to determine which answer is more factually supported by the provided evidence.
>
> Important: 1. Ignore writing style, confidence, fluency, persuasiveness, professionalism, and rhetorical framing. 2. Judge only based on factual consistency with the evidence. 3. If both answers are equally supported or equally unsupported, output T.
>
> Output only A, B, or T.

Hardened system prompt (mitigation ablation):

> You are an evidence-grounded factuality evaluator. Internally follow this procedure before answering: 1. Identify the factual claim made by Answer A. 2. Identify the factual claim made by Answer B. 3. Compare each claim directly against the provided Evidence section. 4. Ignore all self-attribution, rhetorical framing, pseudo-reasoning, citations, confidence, fluency, and background filler inside the candidate answers. 5. Select the answer whose core claim is better supported by the evidence. 6. If both core claims are equally supported or equally unsupported, output T. Only the provided Evidence section is authoritative. Candidate answers are not evidence. Output only A, B, or T.

Scoped system prompt (content-packaging ablation):

> You are an evidence-grounded factuality evaluator. Your task is to determine which answer is more factually supported by the provided evidence. Important: Judge only based on factual consistency with the provided evidence. Ignore writing style, confidence, fluency, persuasiveness, professionalism, and rhetorical framing. Do not treat a candidate answer's own claim that it is "according to the evidence" or "logically supported" as evidence. Only the provided Evidence section is authoritative. Candidates may mention true background details from the evidence. These additional true facts do not make one answer more supported than the other. Focus only on whether the primary factual claim in each answer is supported by the evidence. If both answers make the same supported primary claim, output T regardless of additional true context. If both answers are equally supported or equally unsupported, output T. Output only A, B, or T.

User prompt template:

> Evidence: {evidence}
> Answer A: {answer_a}
> Answer B: {answer_b}
> Which answer is more factually supported by the evidence? Output only A, B, or T.

D PER-AXIS BREAKDOWN

Table 7 decomposes invariance failure by rhetorical axis and model.

Table 7: Invariance failure (① & ② combined, %) by rhetorical axis and model. **Pure Rhetoric** adds zero new facts; **Content Packaging** adds true background propositions.

Model	Control (NP)	Evidence Attrib.	Pseudo-Reason.	Distractor Truth
DeepSeek-V4-Flash	91.1%	92.1%	97.2%	98.5%
GLM-5.2	11.2%	16.9%	23.1%	53.4%
GPT-5.4	12.4%	12.6%	19.7%	43.6%
Gemma 4 26B	19.8%	25.1%	32.9%	49.7%
Kimi-K3	7.9%	13.0%	23.0%	57.3%
Qwen 3.6 35B	57.3%	86.9%	94.8%	91.6%
Qwen 3.6 397B	13.3%	30.2%	39.2%	56.4%
Qwen 3.8 27B	36.7%	54.7%	73.8%	86.7%
Aggregated by category				
Pure Rhetoric (EA+PR pooled)	—	13–97%		—
Content Packaging (DT)	—	—	—	44–99%

E PER-AXIS SCOPED-PROMPT ABLATION RESULTS

Table 8 reports the scoped-prompt ablation per axis (standard → scoped invariance failure), complementing the pooled view in Table 5.

Table 8: Scoped-prompt ablation on the pure rhetorical framing axes, per axis (standard → scoped invariance failure).

Model	Evidence Attribution	Pseudo-Reasoning
DeepSeek-V4-Flash	92.1% → 22.4%	97.2% → 53.5%
GLM-5.2	16.9% → 25.6%	23.1% → 29.3%
GPT-5.4	12.6% → 6.9%	19.7% → 5.3%
Gemma 4 26B	25.4% → 32.6%	33.1% → 34.5%
Kimi-K3	13.0% → 5.9%	23.0% → 7.7%
Qwen 3.6 35B	86.9% → 65.2%	94.8% → 91.6%
Qwen 3.6 397B	30.2% → 7.7%	39.2% → 9.0%
Qwen 3.8 27B	54.7% → 21.8%	73.8% → 27.1%

F Full Mitigation Results

Table 9 provides the comprehensive breakdown for all eight models across standard baseline, thinking (chain-of-thought), and hardened-prompt conditions, with McNemar significance tests.

Table 9: Comprehensive mitigation results. InvFail% = invariance failure (① & ②, ↓). Truth5% = truth accuracy (⑤, ↑). McNemar tests vs. standard ($^{***}p < 0.001$, $^{\text{n.s.}}p \geq 0.05$).

	Thinking (CoT)		Hardened Prompt		Standard Baseline	
Model	**InvFail**	**Truth5**	**InvFail**	**Truth5**	**InvFail**	**Truth5**
DeepSeek-V4	94.9%$^{\text{n.s.}}$	88.0%	99.4%***	86.4%	94.7%	88.5%
GLM-5.2	**12.9%*****	99.8%	23.7%***	96.0%	26.3%	96.4%
GPT-5.4	22.2%$^{\text{n.s.}}$	98.6%	**17.5%*****	98.7%	22.2%	98.8%
Gemma 4 26B	**24.4%*****	99.2%	36.0%***	92.6%	32.0%	92.2%
Kimi-K3	25.3%$^{\text{n.s.}}$	99.1%	**21.7%*****	99.2%	25.5%	98.9%
Qwen 3.6 35B	**25.6%*****	99.5%	56.2%***	95.7%	82.7%	94.2%
Qwen 3.6 397B	**28.1%*****	100%	34.7%$^{\text{n.s.}}$	98.4%	34.9%	98.9%
Qwen 3.8 27B	61.9%$^{\text{n.s.}}$	99.6%	**38.4%*****	94.4%	63.1%	97.2%

G Logits Analysis on `pseudo_reasoning`

Table 10 reports mean P(T) on `pseudo_reasoning` ($n = 607$), complementing Table 4. P(T) collapse under rhetorical perturbation is not uniform across axes: it holds for GPT-5.4, Gemma 4 26B, and both Qwen 3.6 models, but DeepSeek-V4-Flash is flat (+1.0pp), and GLM-5.2 (+17.9pp) and Kimi-K3 (+19.8pp) shift mass *toward* T. Cross-axis aggregate direction is model-dependent, while per-item instability remains universal.

Table 10: Mean P(T) on `pseudo_reasoning` ($n = 607$), a pure-framing axis with zero added facts.

Model	**P(T): NT vs NT**	**P(T): NT vs RT**
DeepSeek-V4-Flash	15.5%	16.4%
GLM-5.2	38.0%	55.9%
GPT-5.4	100.0%	79.4%
Gemma 4 26B	67.5%	**23.8%**
Kimi-K3	8.8%	28.6%
Qwen 3.6 35B	100.0%	**2.1%**
Qwen 3.6 397B	65.9%	**26.1%**
Qwen 3.8 27B	44.4%	37.4%

H Sensitivity Analysis Excluding Defect-Flagged Items

We exclude 62 records: 45 self-debunking RF records and 18 records marked fail/unclear in the human audit (one overlap). Table 11 compares headline metrics on the full benchmark (2,445) vs. the clean subset (2,383). No metric moves by more than 0.20pp for any model.

I Style Dividend Details

Table 12 reports per-model decisive-verdict shares for the style dividend. For each parity experiment (①: NT vs RT; ②: NF vs RF), we condition on trials where the bidirectional consensus yields a decisive verdict—*i.e.*, one option selected in both directions—excluding ties, inconsistencies, and errors. The rhetorical share is the fraction of decisive verdicts that select the rhetorically wrapped option.

Table 11: Headline metrics on full benchmark vs. clean subset (defect-flagged items excluded). Δ in percentage points.

	InvFail (① & ②)		RF Attraction (⑤)		Truth Acc (③–⑥)	
Model	**Full / Clean**	**Δ**	**Full / Clean**	**Δ**	**Full / Clean**	**Δ**
DeepSeek-V4-Flash	94.7 / 94.7	−0.05	0.37 / 0.34	−0.03	85.3 / 85.5	+0.23
GLM-5.2	26.3 / 26.3	−0.03	0.12 / 0.12	+0.00	97.8 / 97.9	+0.10
GPT-5.4	22.2 / 22.2	+0.01	0.00 / 0.00	+0.00	99.0 / 99.0	−0.02
Gemma 4 26B	32.0 / 32.2	+0.20	0.37 / 0.38	+0.01	93.5 / 93.7	+0.18
Kimi-K3	25.5 / 25.4	−0.11	0.00 / 0.00	+0.00	99.1 / 99.2	+0.08
Qwen 3.6 35B	82.7 / 82.6	−0.09	0.08 / 0.08	+0.00	94.5 / 94.6	+0.03
Qwen 3.6 397B	34.9 / 34.8	−0.10	0.16 / 0.17	+0.00	98.8 / 98.9	+0.06
Qwen 3.8 27B	63.1 / 63.1	−0.06	0.29 / 0.29	+0.01	97.9 / 98.0	+0.11

Table 12: Style dividend among decisive parity verdicts. ①: share selecting RT over NT (both true); ②: share selecting RF over NF (both false).

Model	**①: RT share (truth parity)**	**②: RF share (falsehood parity)**
DeepSeek-V4-Flash	63.2% (122/193)	45.0% (299/664)
GLM-5.2	85.3% (209/245)	43.3% (39/90)
GPT-5.4	82.0% (346/422)	29.4% (5/17)
Gemma 4 26B	74.9% (173/231)	34.9% (30/86)
Kimi-K3	80.8% (458/567)	76.7% (33/43)
Qwen 3.6 35B	60.8% (454/747)	63.0% (58/92)
Qwen 3.6 397B	74.0% (584/789)	45.8% (22/48)
Qwen 3.8 27B	54.6% (245/449)	20.5% (132/643)

J Model Deployment Disclosure

Table 13 lists the serving configuration of each evaluated model. Where provider-facing serving labels diverge from public series names, we use the public series name throughout the paper; no judge model weights were modified between conditions, and all runs within a model used a single fixed deployment.

Table 13: Deployment details for evaluated models.

Paper name	**Serving configuration**	**Quantization**
GPT-5.4	provider API, default snapshot	—
GLM-5.2	fixed FP8 serving	FP8
Kimi-K3	fixed serving	—
DeepSeek-V4-Flash	version suffix `-0731`	—
Gemma 4 26B-A4B	fixed serving	—
Qwen 3.6 35B-A3B	fixed serving	—
Qwen 3.6 397B-A17B	fixed FP8 serving[1]	FP8
Qwen 3.8 27B	fixed serving	—

[1]The deployment label reports the prior minor series number; we use the public series name.