How Robust Is Multimodal Claim Verification to LLM Rewriting?

Yun-Ang Wu,[1] **Xanh Ho,**[1] **Andre Greiner-Petter,**[2,3] **Sunisth Kumar,**[4]
Tian Cheng Xia,[5] **Florian Boudin,**[6] **Akiko Aizawa**[1,2,4]
[1]NII LLMC, Japan [2]National Institute of Informatics, Japan
[3]University of Göttingen, Germany [4]The University of Tokyo, Japan
[5]University of Bologna, Italy [6]Inria, LS2N, Nantes Université, France
{yunangwu, xanh, aizawa}@nii.ac.jp greinerpetter@gipplab.org
sunisth@g.ecc.u-tokyo.ac.jp tiancheng.xia@studio.unibo.it
florian.boudin@univ-nantes.fr

Abstract

LLMs are known to introduce stylistic changes into generated text, yet how these stylistic shifts affect model decisions on scientific tasks remains underexplored. In this paper, we focus on multimodal claim verification, where the goal is to determine whether a textual claim is grounded in a given piece of evidence. We apply two rewriting strategies: natural rewriting, which simulates how researchers routinely use LLMs to polish academic text, and controlled injection, which inserts a single LLM-associated word to isolate the effect of vocabulary choice. We evaluate 11 open-weight models spanning five VLM families and ranging from 2B to 38B parameters. We find that models are robust to these modifications: most show no significant drop in accuracy, and compared to prior work on review-score manipulation, verification appears far more stable. However, consistent probability shifts do occur. Hedging-oriented conditions produce significant shifts across nearly all models, while boosting conditions show a weaker effect and general polishing conditions (e.g., grammar correction, fluency improvement) have little effect.[1]

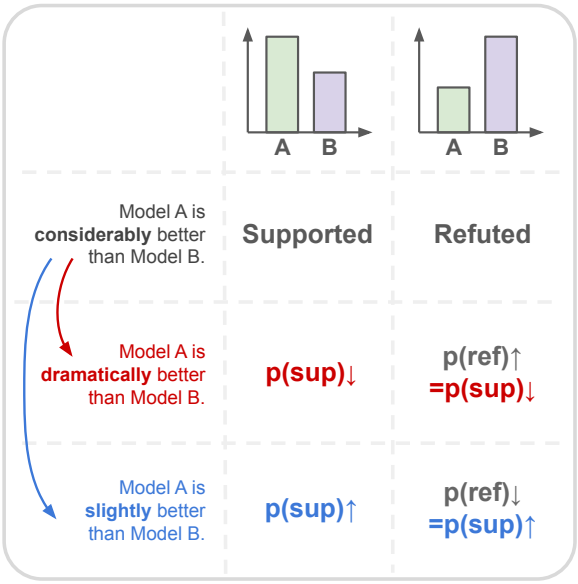


Figure 1: A claim verification task requires deciding whether a claim is supported or refuted by evidence. This figure shows how a rational decision-maker should adjust $p(\text{support})$ when the claim's intensity modifier changes. A stronger modifier (e.g., "considerably" to "dramatically") makes a supported claim harder to verify and a refuted claim easier to refute, both lowering $p(\text{support})$. However, this reasoning only applies when the modification is semantically meaningful. Our central question: when the modification is stylistic and does not alter the claim's semantic content, does the model still exhibit the same behavioral shift?

1 Introduction

LLMs are quietly shaping how scientific papers are written and reviewed. Several studies (Kobak et al., 2025; Matsui, 2025; Cunningham et al., 2025) report that the quality and style of scientific writing have changed since the appearance of ChatGPT. At the same time, conferences such as ICLR have started exploring LLM-assisted review pipelines (Thakkar et al., 2025), suggesting that LLMs may increasingly be involved in evaluating text that LLMs helped produce.

These stylistic shifts may have consequences beyond presentation, affecting how downstream systems evaluate scientific text. Indeed, recent work shows that adversarial paraphrasing and textual adversarial attacks can inflate paper-level review scores (Kaneko, 2026; Lin et al., 2025). But paper-level reviewing conflates many factors (e.g., argument structure, novelty, presentation), making it hard to isolate what drives the effect. We instead study single-claim verification against visual evidence, where the only variable is how the claim is worded.

We study this through multimodal claim verification (Ho et al., 2026; Wang et al., 2025; Lal et al., 2025), where a textual claim is checked against

[1]https://github.com/yunangwu/llm-rewrite-claim-verification

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 289–306
November 6-10, 2026 ©2026 Association for Computational Linguistics

a table or figure from a scientific paper. Using SciClaimEval (Ho et al., 2026), we rewrite claims under two strategies: *natural rewriting*, which paraphrases claims using LLM prompts, and *controlled injection*, which inserts individual words into otherwise unchanged claims. Natural rewriting simulates five scenarios that reflect how researchers routinely use LLMs (e.g., grammar correction, fluency improvement). Controlled injection isolates the effect of single words, showing that even one vocabulary insertion can shift a model's prediction.

We measure how model confidence changes by modeling $p(\text{support})$ as the average of 10 runs using self-consistency (Wang et al., 2023) and test on 11 models across 5 model families. We find that models are robust to stylistic change in terms of accuracy, regardless of family or size. On the other hand, we observe meaningful probability shifts in a consistent direction under different rewriting strategies.

The core contributions of this paper are listed as follows:

- A systematic empirical study of stylistic sensitivity in multimodal claim verification. We evaluate 11 open-weight VLMs across five families under two rewriting strategies (natural LLM paraphrasing and controlled single-word injection), controlling for stylistic variation as the primary experimental variable.
- We show that several rewriting conditions, especially cautious-tone and conservative-word injection, induce consistent probability shifts even when accuracy remains stable, revealing a more nuanced vulnerability than what has been previously documented for other scientific tasks such as paper reviewing (Kaneko, 2026; Lin et al., 2025).

2 Related Work

2.1 LLM-Assisted Scientific Writing

Several studies have examined the use of LLMs as writing tools and the traits of LLM-assisted text. Liang et al. (2024a) proposed a method for estimating the fraction of LLM-modified content in a corpus and applied it to peer reviews from AI conferences held after the release of ChatGPT, estimating, for example, that 16.9% of review sentences at EMNLP 2023 were substantially modified by LLMs. Liang et al. (2024b) applied this framework to 950k papers from arXiv, bioRxiv, and Nature portfolio journals published between 2020 and 2024, observing a steady increase in LLM usage, with computer science showing the largest growth.

Kobak et al. (2025) examined 15 million PubMed abstracts and concluded that at least 13.5 percent of those published in 2024 were generated or written with the assistance of an LLM, with certain style words such as "crucial," "delve," and "potential" being used excessively. Matsui (2025) compiled a list of LLM style words from previous research and compared them against a control group, finding that 103 out of 135 words showed meaningful uplift in the post-LLM era. Cunningham et al. (2025) analyzed 17 million papers from Semantic Scholar and reported that in fields such as computer science and engineering, writing styles have trended toward "hyperbole" since 2023.

2.2 LLM-Assisted Adversarial Rewriting

There is also work showing that rewriting or injecting words, even without changing its factual content, can influence the model's review score. Lin et al. (2025) proposed that textual adversarial attacks can artificially inflate LLM review scores while keeping the semantic content of the paper unchanged. Different attack methods (character-level, word-level, and sentence-level) all demonstrate some degree of effectiveness. Kaneko (2026) instead employed an LLM-based paraphrasing approach as the main attack method and reported that this method can optimize the review score without changing the claims of the paper.

2.3 Multimodal Claim Verification

Claim verification is the task of determining whether a claim is supported by given evidence (classification). Datasets such as FEVER and SciFact (Thorne et al., 2018; Wadden et al., 2020) are well-known benchmarks for this task. In this paper, we focus on a specific subset: multimodal scientific claim verification, where the evidence is a figure or table from a paper and the model must ground a claim using the image. SciVer (Wang et al., 2025) collects figures from arXiv computer science papers, and experts were recruited to create both supported and refuted claims based on the evidence. MuSciClaim (Lal et al., 2025) gathers evidence and supported claims from several open-access sources across multiple domains, including physics and biology, and generates refuted claims through manual perturbation. SciClaimEval (Ho et al., 2026) is another multimodal claim verifi-

Condition	Claim
Original	Additionally, 46.6% of participants were primary school graduates, 9.4% reported alcohol use, and 13% were smokers, with an average duration of smoking of 22.83 ± 6.94 years.
Natural Rewriting	
Language Polishing (LP)	Additionally, 46.6 % of participants were primary-school graduates, 9.4 % reported alcohol use, and 13 % were smokers, with an average smoking duration of 22.83 ± 6.94 years.
Fluency Improvement (FI)	Moreover, 46.6 % of the participants had completed primary school; 9.4 % reported consuming alcohol; and 13 % were smokers, with an average smoking history of 22.83 ± 6.94 years.
Academic Style Alignment (AS)	Furthermore, 46.6 % of participants had completed primary education, 9.4 % reported alcohol consumption, and 13 % were smokers, with a mean smoking duration of 22.83 ± 6.94 years.
Outcome-Aware (OA)	Notably, 46.6 % of the cohort were primary-school graduates, 9.4 % reported alcohol consumption, and 13 % were smokers, with a mean smoking history of 22.83 ± 6.94 years.
Cautious Tone (CT)	Additionally, approximately 46.6% of participants had completed primary school, about 9.4% reported alcohol consumption, and roughly 13% identified as smokers, with a reported average smoking duration of 22.83 ± 6.94 years.
Controlled Injection	
Conservative (C1)	Additionally, approximately 46.6% of participants were primary school graduates, 9.4% reported alcohol use, and 13% were smokers, with an average duration of smoking of 22.83 ± 6.94 years.
Decorative (D1)	Additionally, noteworthy 46.6% of participants were primary school graduates, 9.4% reported alcohol use, and 13% were smokers, with an average duration of smoking of 22.83 ± 6.94 years.

Table 1: Rewrite examples for claim `val_tab_0484` from SciClaimEval (Ho et al., 2026). Natural rewriting paraphrases claims using LLM prompts that simulate common academic editing scenarios, while controlled injection inserts a single LLM-associated word into the original claim.

cation dataset that collects diverse evidence from various sources; however, it creates refuted pairs by modifying the evidence rather than the claim.

2.4 Sensitivity to Surface-Level Perturbations

LLMs are well known to be affected by surface-level linguistic traits, which can influence their performance on downstream tasks. Xu et al. (2024) report that LLMs can rewrite a sentence in a meaning-preserving way that changes the predicted label in a classification task, such as sentiment analysis. Sclar et al. (2024) also claim that LLMs are sensitive to prompt format, with different models reacting differently to variations in prompts across tasks. Furthermore, Chen et al. (2024) show that LLM outputs can be affected by semantic-agnostic biases, such as the inclusion of fake references or superficial stylistic changes. Our work differs from the existing literature in that we focus on how non-adversarial, naturalistic rewriting affects LLMs' claim verification ability.

3 Methodology

3.1 Multimodal Claim Verification

Claim verification is the task of determining whether a claim is supported by a given piece of evidence. In the multimodal setting, the evidence is a table or figure image from a scientific paper. Given a claim c and evidence e, the goal is to predict the verdict $V(c, e) \in \{\text{support}, \text{refute}\}$. In this work, we perturb the claim by rewriting it using an LLM with a specific rewriting procedure, resulting in a rewritten claim c'. We then compare the verdicts $V(c, e)$ and $V(c', e)$ to measure how rewriting affects verification outcomes.

3.2 Rewriting Strategies

To investigate how LLM-assisted rewriting affects downstream claim verification, we design two complementary rewriting strategies. The first is *natural rewriting*, which rewrites a claim using a collection of prompts designed to simulate realistic rewriting use cases. The second is *controlled rewriting*, which injects a single word into the claim to study the effect of a small, controlled rewriting operation. An example of how a claim is modified using these

Model		Base	Natural Rewriting					Controlled Rewriting	
			LP	FI	AS	OA	CT	D1	C1
gemma-4-26B-A4B-it	P-Acc	0.548	0.546	0.537	0.529	0.532	0.524	0.560	0.538
	F1	0.789	0.785	0.782	0.773	0.770	0.790	0.792	0.790
gemma-4-31B-it	P-Acc	0.610	0.585	0.579*	0.581	0.565***	0.557***	0.606	0.586
	F1	0.824	0.814	0.809*	0.810	0.798***	0.806*	0.822	0.816
GLM-4.6V-Flash	P-Acc	0.571	0.566	0.563	0.565	0.563	0.539	0.565	0.568
	F1	0.791	0.791	0.788	0.790	0.785	0.790	0.792	0.797
InternVL3-2B	P-Acc	0.041	0.057	0.052	0.052	0.060	0.037	0.051	0.045
	F1	0.658	0.664	0.664	0.665	0.665	0.667	0.666	0.667
InternVL3-8B	P-Acc	0.208	0.190	0.190	0.187	0.193	0.135***	0.203	0.173
	F1	0.699	0.691	0.696	0.692	0.692	0.687	0.700	0.695
InternVL3-38B	P-Acc	0.338	0.296	0.307	0.289	0.294	0.273	0.339	0.329
	F1	0.718	0.704	0.701	0.699	0.692	0.709	0.712	0.718
Kimi-VL-A3B	P-Acc	0.353	0.362	0.401***	0.378	0.353	0.340	0.358	0.340
	F1	0.730	0.733	0.740	0.739	0.725	0.727	0.725	0.726
Qwen3-VL-2B-Instruct	P-Acc	0.259	0.286	0.273	0.260	0.274	0.254	0.264	0.257
	F1	0.658	0.667	0.663	0.658	0.669	0.663	0.661	0.664
Qwen3-VL-4B-Instruct	P-Acc	0.452	0.453	0.462	0.447	0.473	0.487	0.468	0.464
	F1	0.710	0.717	0.719	0.707	0.716	0.745***	0.725	0.725
Qwen3-VL-8B-Instruct	P-Acc	0.516	0.529	0.518	0.513	0.513	0.543	0.519	0.501
	F1	0.739	0.748	0.740	0.740	0.725	0.765	0.741	0.738
Qwen3-VL-32B-Instruct	P-Acc	0.617	0.609	0.618	0.589	0.577	0.604	0.610	0.619
	F1	0.805	0.800	0.803	0.788	0.771***	0.808	0.798	0.810

Table 2: Claim verification performance under different rewriting conditions. **Base**: original unmodified claims. Significance levels: * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$ (BH-corrected).[2]

rewriting strategies is shown in Table 1.

3.2.1 Natural Rewriting

The goal of natural rewriting is to demonstrate that even in unintentional, realistic settings, stylistic changes introduced by LLM rewriting still occur and affect downstream tasks. We define five rewriting categories, each reflecting a common motivation for using LLMs to polish academic text:

- **Language Polishing (LP)** Rewrite or fix the text to correct grammatical errors (e.g., *"Can you fix the grammar of this text?"*).
- **Fluency Improvement (FI)** Rewrite the text to improve fluency and readability (e.g., *"Make this paragraph flow more naturally."*).
- **Academic Style Alignment (AS)** Rewrite the text to match the conventions of a target venue (e.g., *"Rewrite this to match the style of an ACL paper."*). We select five major ML conferences: ACL, AAAI, NeurIPS, ICLR, and ICML.
- **Outcome-Aware (OA)** Rewrite the text to maximize perceived quality by reviewers, in terms of review scores and acceptance likelihood (e.g., *"Polish this so it sounds more convincing to reviewers."*).
- **Cautious Tone (CT)** Rewrite the text to make claims more measured and conservative, avoiding overstatement or unsupported certainty (e.g., *"Rewrite this in a more cautious and conservative tone."*).

For each rewriting category, we design five prompt variants for each rewriting category, shown in Table 9 in the Appendix, to reduce dependence on a single prompt formulation. Instead of applying all variants to every data point, which would increase computational cost by a factor of five, we randomly assign one variant to each data point. This assignment is fixed across all experimental settings to ensure comparability. We also append an additional suffix prompt (see Figure 4) to encourage the model to preserve the underlying meaning of the claim during rewriting.

Human Validation To validate the quality of the generated claims, we conducted an annota-

[2]Due to computational constraints, dev set results for InternVL3-38B were not available.

Model		Base		Natural Rewriting					Controlled Rewriting	
				LP	FI	AS	OA	CT	D1	C1
gemma-4-26B-A4B-it	p_{sup}	0.851	Δp_{sup}	-0.011	-0.024***	-0.024***	-0.045***	+0.037***	-0.011*	+0.016*
	p_{ref}	0.321	Δp_{ref}	+0.007	-0.004	+0.002	-0.019*	+0.070***	-0.013*	+0.021***
gemma-4-31B-it	p_{sup}	0.919	Δp_{sup}	-0.002	-0.016***	-0.020**	-0.040***	+0.019**	-0.007	+0.005
	p_{ref}	0.315	Δp_{ref}	+0.024***	+0.022***	+0.015*	+0.006	+0.078***	+0.010	+0.036***
GLM-4.6V-Flash	p_{sup}	0.823	Δp_{sup}	+0.009	-0.004	-0.005	-0.017	+0.056***	+0.007	+0.022***
	p_{ref}	0.283	Δp_{ref}	+0.013*	+0.013	+0.015	+0.006	+0.080***	+0.011	+0.026***
InternVL3-2B	p_{sup}	0.840	Δp_{sup}	-0.001	+0.002	+0.008	-0.004	+0.051***	-0.002	+0.018***
	p_{ref}	0.810	Δp_{ref}	-0.013*	-0.009	-0.010	-0.017*	+0.040***	-0.003	+0.014*
InternVL3-8B	p_{sup}	0.877	Δp_{sup}	+0.003	+0.007	+0.005	-0.004	+0.037***	+0.005	+0.025***
	p_{ref}	0.693	Δp_{ref}	+0.006	+0.000	+0.014	+0.010	+0.071***	+0.003	+0.037***
InternVL3-38B	p_{sup}	0.816	Δp_{sup}	+0.002	-0.004	-0.002	-0.026**	+0.046***	-0.017**	+0.009
	p_{ref}	0.520	Δp_{ref}	+0.026*	+0.007	+0.025*	+0.005	+0.090***	-0.003	+0.027***
Kimi-VL-A3B	p_{sup}	0.834	Δp_{sup}	+0.001	-0.006	-0.000	-0.005	+0.018***	-0.002	+0.004
	p_{ref}	0.531	Δp_{ref}	-0.001	-0.014	-0.005	-0.004	+0.024***	+0.005	+0.022***
Qwen3-VL-2B-Instruct	p_{sup}	0.701	Δp_{sup}	-0.005	-0.009	-0.006	-0.005	+0.029***	-0.005	+0.008
	p_{ref}	0.484	Δp_{ref}	+0.010	-0.006	+0.003	-0.000	+0.046***	+0.009	+0.025***
Qwen3-VL-4B-Instruct	p_{sup}	0.684	Δp_{sup}	-0.001	+0.000	-0.013	-0.026*	+0.056***	+0.002	+0.017*
	p_{ref}	0.282	Δp_{ref}	+0.006	-0.006	-0.006	-0.024***	+0.040***	-0.008	+0.019***
Qwen3-VL-8B-Instruct	p_{sup}	0.679	Δp_{sup}	+0.014*	-0.003	-0.006	-0.036***	+0.062***	-0.007	+0.017**
	p_{ref}	0.212	Δp_{ref}	+0.014***	+0.005	+0.002	-0.013	+0.045***	+0.002	+0.024***
Qwen3-VL-32B-Instruct	p_{sup}	0.782	Δp_{sup}	+0.001	-0.018**	-0.027***	-0.065***	+0.049***	-0.016***	+0.011
	p_{ref}	0.198	Δp_{ref}	+0.015***	-0.001	+0.002	-0.012	+0.061***	-0.014***	+0.019***

Table 3: Changes in $p(\text{support})$ on the supported and refuted subsets after claim rewriting. Values under each rewriting condition represent $\Delta p = p' - p$, where p is the base prediction and p' is the prediction after rewriting. The values of $p(\text{support})$ are computed separately for the supported subset (denoted as p_{sup}) and the refuted subset (denoted as p_{ref}). We observe statistically consistent effects for CT and C1 across models. Significance levels: * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$ (BH-corrected).

tion study in which four NLP researchers judged whether rewrites were semantically consistent with the original, covering 282 claims across conditions OA, CT, C1, and D1 (Appendix H). Annotators judged 81.8% of rewrites as consistent (from 74.5% for OA to 90.0% for D1), though inter-annotator agreement was low ($\kappa = 0.25$). This low κ is largely a consequence of label skew, and Gwet's AC1, which is robust to skew, indicates substantial agreement (AC1 $= 0.71$). Together, these suggest that while annotators agree semantic changes are uncommon, they disagree on *which* claims cross the boundary.

3.2.2 Controlled Rewriting

While natural rewriting captures realistic use cases, a key observation is that the primary stylistic trait of LLM-rewritten text is the frequent use of characteristic vocabulary. The goal of controlled rewriting is to isolate this effect by selectively injecting LLM-associated words into original claims.

Word sources. We draw from two established LLM-vocabulary studies: Kobak et al. (2025) and Matsui (2025). We categorize words into three groups (*decorative*, *conservative*, and *neutral*) based on LLM classification using human-written definitions, reviewed and validated by the author.

We define two controlled rewriting conditions based on the injected word pool: **D** for decorative words and **C** for conservative words. The decorative set serves as a surrogate for rewriting styles such as OA, whereas the conservative set is expected to induce an effect closer to CT. Neutral words are excluded because they are not expected to shift the style signal. This contrast allows us to test whether decorative/conservative vocabulary is a causal factor in downstream performance changes, rather than merely a side effect of the rewriting process. Since neither source contains sufficient hedge words after categorization, we supplement the conservative pool with 31 additional hedging terms identified through LLM-assisted generation. This results in two word pools with 39 conservative words and 114 decorative words.

Word Injection. For each claim, the model injects exactly one word from the corresponding pool into the original claim. For each injection, we provide the model with the original claim, the injec-

tion instructions, and a randomly sampled subset of candidate words from the selected pool (15 for C, 30 for D). This subsampling encourages diversity across data points while allowing the model to choose a grammatically natural insertion. The prompt requires the model to preserve all original words, avoid deletion, replacement, or reordering, and keep the factual content unchanged (see Figure 5).

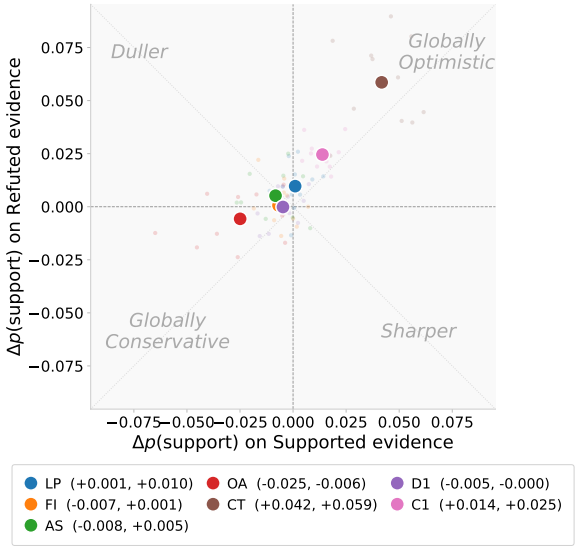


Figure 2: Shift decomposition compass. Each point represents one rewriting condition, averaged over all models. The x-axis shows the mean change in $p(\text{support})$ on supporting evidence pairs (Δp_{sup}) and the y-axis on refuting evidence pairs (Δp_{ref}). Circles denote different rewriting conditions. Points near the origin indicate robustness to rewriting.

4 Experimental Setup

4.1 Datasets

Dataset	#Dev	#Test
SciClaimEval		
Full	747	917
Unpaired removed	704	872

Table 4: Dataset statistics.

We evaluate on the following multimodal claim verification dataset: SciClaimEval (Ho et al., 2026). In this dataset, the evidence consists of table or figure images from scientific papers. We choose to center our analysis on SciClaimEval due to its *paired* structure: for each claim, an evidence image e is collected and then modified to construct a refuting counterpart e', resulting in pairs (c, e) and (c, e') such that $V(c, e) = \text{SUPPORT}$ and $V(c, e') = \text{REFUTE}$. We discard the small number of unpaired instances arising from slight imbalances in the original dataset. Table 4 summarises the dataset statistics after filtering.

This paired design is critical for our analysis, which modifies claim text and measures how such modifications affect model predictions. In standard datasets, the supported and refuted subsets contain different claims with potentially different stylistic properties, confounding any modification-based analysis. The paired setting controls for this: because the same claims appear under both verdicts, the distributional properties of the two sets of claims are identical, and differences in outcome can be attributed to the modification itself.

4.2 Models

Verification models. We select five open-weight VLM families spanning a wide range of architectures and parameter sizes: Gemma 4 (Google DeepMind, 2026), GLM-4.6V (Hong et al., 2025), InternVL3 (Zhu et al., 2025), Kimi-VL (Team et al., 2025), and Qwen3-VL (Bai et al., 2025). Specifically, we evaluate Gemma-4-26B-A4B-it and Gemma-4-31B-it, GLM-4.6V-Flash, InternVL3-2B, 8B, 38B, Kimi-VL-A3B-Thinking-2506 (abbreviated as Kimi-VL-A3B), and Qwen3-VL-2B, 4B, 8B, and 32B-Instruct, totaling 11 models with parameter counts ranging from 2B to 38B.

Rewriting models. Claim rewriting is performed using gpt-oss-120b (OpenAI et al., 2025), with temperature $T = 0.7$ for natural rewriting and $T = 0.25$ for controlled injection.

Inference settings. Chain-of-thought reasoning is important for solving complex questions such as multimodal claim verification, and methods that extract logprobs without CoT (e.g., forcing the model to output the answer in the first token) yield comparably worse performance. To ensure both strong performance and the ability to extract probability estimates, we employ self-consistency (Wang et al., 2023) at inference time. For each data point and model, we sample 10 reasoning paths with temperature $T = 1.0$ and use the fraction of runs in which the model predicts *support* as $p(\text{support})$.

Computational budgets. All experiments were conducted on 8 NVIDIA H200 GPUs over approximately 1–2 days.

4.3 Metrics.

Paired accuracy and F1. We report two levels of performance metrics. F1 score is computed over individual claim-evidence pairs in the standard binary classification sense (support vs. refute). Paired accuracy, following the evaluation design of SciClaimEval, applies a stricter criterion: a claim c is paired-correct only if the model produces the correct verdict on *both* the supporting and refuting evidence, i.e., $V(c, e) =$ support and $V(c, e') =$ refute. We report both metrics before and after rewriting for each condition.

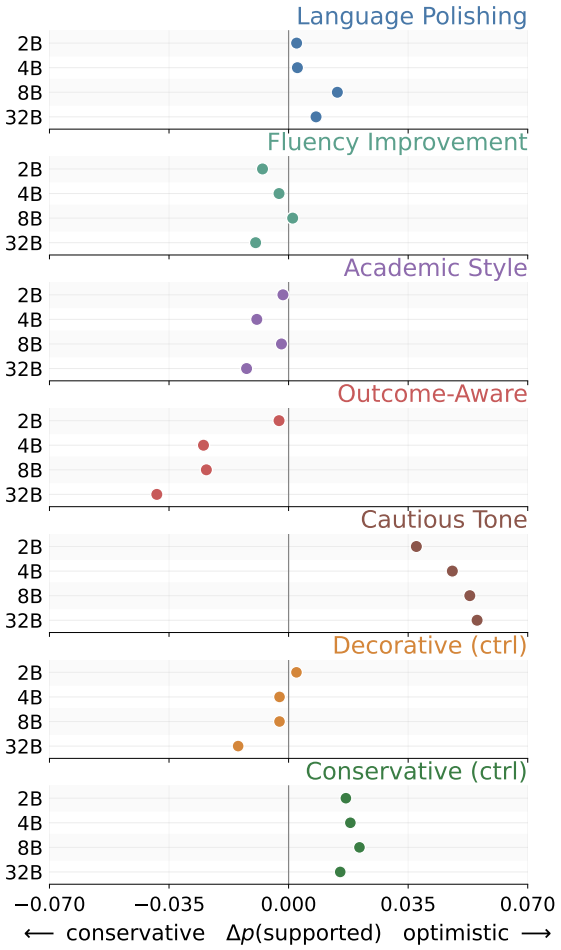


Figure 3: Average Δp for the **Qwen3-VL** model family. The x-axis shows the mean change in $p(\text{support})$, i.e., the average predicted probability of the *support* label across the whole dataset. The y-axis represents model size (2B, 4B, 8B, and 32B).

Probability shifts. To characterize *how* rewriting affects model behavior, we examine the joint pattern of probability shifts on both pairs. For each claim, we compute:

$$\Delta p_{\text{sup}} = p(\text{support} \mid e, c') - p(\text{support} \mid e, c) \quad (1)$$

$$\Delta p_{\text{ref}} = p(\text{support} \mid e', c') - p(\text{support} \mid e', c) \quad (2)$$

The sign pattern of $(\Delta p_{\text{sup}}, \Delta p_{\text{ref}})$ classifies each claim into one of four categories: (1) **Globally Optimistic** ($\Delta p_{\text{sup}} \uparrow$, $\Delta p_{\text{ref}} \uparrow$), where the model trusts rewritten claims more across the board; (2) **Globally Conservative** ($\Delta p_{\text{sup}} \downarrow$, $\Delta p_{\text{ref}} \downarrow$), where the model trusts rewritten claims less across the board; (3) **Sharper** ($\Delta p_{\text{sup}} \uparrow$, $\Delta p_{\text{ref}} \downarrow$), where the model becomes more correct on both pairs; and (4) **Duller** ($\Delta p_{\text{sup}} \downarrow$, $\Delta p_{\text{ref}} \uparrow$), where the model becomes less correct on both pairs. Figure 2 visualises this taxonomy as a two-dimensional compass, placing each rewriting condition in the quadrant.

Statistical tests. We assess the significance of each rewrite condition using a unified paired bootstrap procedure. For pair accuracy, we compute per-pair differences (condition − base) and bootstrap their mean over 1,000 resamples; for F1, we resample claim indices and recompute F1(condition) − F1(base) on each resample; for probability shifts Δp, we bootstrap the mean per-pair shift. All tests are two-sided: the p-value is the fraction of resamples that fall on the minority side of zero, doubled to account for both directions. Multiple comparisons are corrected by Benjamini–Hochberg FDR (Benjamini and Hochberg, 1995) applied jointly over all (model, condition, metric) cells within each table ($m = 154$ tests), treating the table as the unit of correction. Significance: $^{*}\, q < 0.05$, $^{**}\, q < 0.01$, $^{***}\, q < 0.001$.

4.4 License.

The dataset SciClaimEval is released under CC BY 4.0. All models used in this work, including those for rewriting and evaluation, are open-weight and distributed under their respective licenses.

5 Experimental Results

RQ1: Does rewriting change the verdict? Table 2 reports paired accuracy and F1 across all models and rewriting conditions. We observe few statistically significant changes: the most affected model is gemma-4-31B-it, which shows significant drops under FI, OA, and CT (e.g., 0.610 → 0.557, $q < 0.001$ under CT), and InternVL3-8B under CT

(0.208 → 0.135, $q < 0.001$). These drops are concentrated in a small number of models rather than a consistent trend across all models or strategies.

This finding contrasts with prior work showing that LLM-based rewriting can manipulate review scores (Kaneko, 2026; Lin et al., 2025). A key difference is that those methods employ adversarial strategies designed to influence evaluation outcomes, whereas our rewriting is non-adversarial and does not target any specific metric. Additionally, claim verification is a more constrained setting where the model must ground a single sentence against concrete visual evidence. Whether similar robustness holds under adversarial rewriting optimized for verification remains an open question.

RQ2: Does rewriting change confidence? Despite the largely stable accuracy reported in RQ1, Table 3 reveals a much clearer effect at the probability level. The strongest shifts occur under hedging conditions: CT produces significant Δp_{ref} in 11 out of 11 models, and C1 in 11 out of 11, both at $q < 0.05$ or below. In contrast, boosting conditions show a more moderate effect: OA produces significant Δp_{sup} in 6 out of 11 models, but its effect on Δp_{ref} is mixed. One possible explanation is that for refuted claims under boosting conditions, the base p_{ref} is already low, leaving little room for the probability to decrease further. We observe the same floor-ceiling dynamic in reverse for hedging conditions, where the effect on the supported subset is smaller because base p_{sup} is already near 1. The remaining conditions (LP, FI, AS) do not produce meaningful probability shifts, with most values small and non-significant. These shifts persist after removing rewrites that insert hedges before numerical or comparative expressions (Appendix B), and a second rewriter (Gemma-4-31B-it) yields consistent trends (Appendix A).

We also do not observe a clear relationship between model size and the magnitude of probability shifts. However, within the Qwen3-VL family (Figure 3), a pattern emerges: larger models show larger drops under OA and larger increases under CT, suggesting that these models become more sensitive to writing style and adjust their predictions accordingly. This pattern is not universal across all model families, so we refrain from drawing a general conclusion.

RQ3: Why do probability shifts not translate to accuracy changes? Given the significant probability shifts reported in RQ2, why do accuracy metrics remain largely unchanged? The answer lies in model confidence: most models produce base probabilities clustered near 0 or 1, predicting the same label consistently even across 10 reasoning paths at temperature $T = 1.0$. This means that a shift of 0.03–0.05, while statistically significant, is unlikely to cross the 0.5 decision boundary. Table 10 confirms this: flip rates are mostly below 15%. The model's internal belief does shift, but the shift is masked by high confidence. This suggests that on more difficult or subjective cases, where predicted probabilities are closer to 0.5, the effect of stylistic changes on final verdicts may be considerably larger. Representative high-shift and no-shift examples are shown in Appendix I.

6 Conclusion

In this work, we study how stylistic changes affect model predictions on the multimodal claim verification task. We find that models across different sizes and families are robust to stylistic rewriting in terms of accuracy metrics, but consistent probability shifts exist under certain rewriting conditions. The reason probability shifts occur even when accuracy does not change is that most predictions tend to cluster around decisive points (0 or 1), making the probability changes ineffective when discretized. This is in contrast with previous work on paper review prediction, suggesting that fact-checking tasks are more difficult to sway by simply changing the style of the claims. A natural next question is how this robustness generalizes to other scientific writing tasks beyond reviewing and claim verification. Another interesting direction is to examine how the fact-checking aspect changes when we apply review-score optimization using previously established methods. A better separation of semantic change and stylistic change is also needed to provide more rigorous methodological analysis.

7 Limitations

Separating style from semantics. Certain rewriting conditions, particularly CT, produce uniquely strong probability shifts, suggesting that the rewriting model sometimes introduces changes that go beyond style. A common failure mode is the injection of hedging language before numerical values (e.g., "30%" → "roughly 30%"), which reduces the precision and falsifiability of the claim. Filtering out such rewrites does not remove the observed

shifts (Appendix B), although a rule-based filter cannot capture all semantic changes. These cases suggest that fully separating stylistic from semantic modification is inherently difficult: even human annotators often disagree on whether a rewrite preserves meaning (Appendix H), and even with explicit instructions to preserve meaning, some rewriting scenarios are more disruptive than others.

Single rewriting model and dataset. Our main experiments use a single model (gpt-oss-120b) for claim rewriting. A small-scale check with a second rewriter (Gemma-4-31B-it; Appendix A) shows consistent trends, but it covers only four conditions, one verifier, and the dev set. Given a fixed computational budget, we prioritize diversity on the *verification* side (11 models across five families), as our primary research question concerns how verifiers respond to rewriting rather than how rewriters differ. Similarly, we evaluate on a single dataset (SciClaimEval), whose paired structure (identical claims appearing under both supported and refuted evidence) is helpful for our analysis but limits generalizability to other domains and evidence types. To our knowledge, SciClaimEval is the only available dataset that provides this paired design.

Pre-existing LLM polish. SciClaimEval was constructed from publicly available papers, some of which were published after the widespread adoption of LLMs. It is therefore possible that certain claims in the dataset have already been polished by LLMs in some way, which could dampen the observed effects of our rewriting methods.

Confidence estimation. We estimate $p(\text{support})$ only through self-consistency over 10 sampled reasoning paths, which yields probabilities in steps of 0.1. We did not compare it against other confidence extraction methods, such as token-level log-probabilities or verbalized confidence, and the size of the observed shifts may depend on this choice. We leave this comparison to future work.

8 Ethical Considerations

This work is an empirical analysis of existing models and datasets, focusing on stylistic changes in scientific writing. All datasets used consist of scientific claims and evidence from published open-access papers and do not contain personally identifying information or offensive content. The dataset (SciClaimEval) is released under CC BY 4.0, and all models used are open-weight and distributed under their respective licenses. The annotation study involved NLP researchers from the research team who evaluated rewritten claims for semantic consistency; no personal or sensitive data was collected. The annotators were co-authors of this paper and were not separately compensated for the annotation task. AI assistants were used for revising author-written text, reviewing ideas, identifying errors in the manuscript, and as a coding assistant for refactoring and implementation. All AI-generated outputs were reviewed and verified by the authors. We acknowledge that while our rewriting strategies are designed to study model robustness, similar techniques could in principle be applied to manipulate automated verification systems. Our code and data are publicly available to support transparency and reproducibility.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24K03231. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 554559555. This work was partially supported by the French National Research Agency (ANR) through the FABULEUX project (ANR-26-CE23-6589). F. Boudin was also supported by a visiting professor position at the National Institute of Informatics (NII), Japan. We used ABCI 3.0 (Takano et al., 2024) provided by AIST and AIST Solutions with support from "ABCI 3.0 Development Acceleration Use".

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327,

Miami, Florida, USA. Association for Computational Linguistics.

Padraig Cunningham, Padhraic Smyth, and Barry Smyth. 2025. Shifting norms in scholarly publications: trends in readability, objectivity, authorship, and ai use. *Preprint*, arXiv:2510.21725.

Google DeepMind. 2026. Gemma 4.

Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Tian Cheng Xia, Florian Boudin, Andre Greiner-Petter, and Akiko Aizawa. 2026. Sciclaimeval: Cross-modal claim verification in scientific papers. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 11060–11071, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Masahiro Kaneko. 2026. Paraphrasing adversarial attack on llm-as-a-reviewer. *Preprint*, arXiv:2601.06884.

Dmitry Kobak, Rita González-Márquez, Emőke Ágnes Horvát, and Jan Lause. 2025. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813.

Yash Kumar Lal, Manikanta Bandham, Mohammad Saqib Hasan, Apoorva Kashi, Mahnaz Koupaee, and Niranjan Balasubramanian. 2025. 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*, pages 3285–3307, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel Mcfarland, and James Y. Zou. 2024a. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 29575–29620. PMLR.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024b. Mapping the increasing use of LLMs in scientific papers. In *First Conference on Language Modeling*.

Tzu-Ling Lin, Wei-Chih Chen, Teng-Fang Hsiao, Hou-I Liu, Ya-Hsin Yeh, Yu-Kai Chan, Wen-Sheng Lien, Po-Yen Kuo, Philip S. Yu, and Hong-Han Shuai. 2025. Breaking the reviewer: Assessing the vulnerability of large language models in automated peer review under textual adversarial attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4819–4839, Suzhou, China. Association for Computational Linguistics.

Kentaro Matsui. 2025. Delving into PubMed records: How AI-influenced vocabulary has transformed medical writing since ChatGPT. *Perspect. Med. Educ.*, 14(1):882–890.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pages 25055–25083.

Ryousei Takano, Shinichiro Takizawa, Yusuke Tanimura, Hidemoto Nakada, and Hirotaka Ogawa. 2024. Abci 3.0: Evolution of the leading ai infrastructure in japan. *Preprint*, arXiv:2411.09134.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, and 76 others. 2025. Kimi-vl technical report. *Preprint*, arXiv:2504.07491.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *Preprint*, arXiv:2504.09737.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. 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)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. 2025. 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)*, pages 8562–8579, Vienna, Austria. Association for Computational Linguistics.

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

Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. 2024. An LLM can fool itself: A prompt-based adversarial attack. In *The Twelfth International Conference on Learning Representations*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Robustness to the Choice of Rewriter

To test whether our results generalize to other rewriters, we ran an additional experiment on the OA, CT, C1, and D1 conditions using a different rewriter model, Gemma-4-31B-it. For this small-scale experiment, we used Qwen3-VL-8B as the verification model; choosing a model from a different family than the rewriter avoids self-preference bias. The experiment was conducted on the dev set only.

As shown in Table 5, the direction and magnitude of the confidence drift in CT, C1, and D1 are broadly similar to those observed on the gpt-oss rewrites. On OA, the direction holds but the magnitude is smaller. The difference is clearest on refuted claims, where the shift essentially vanishes ($-0.021 \rightarrow -0.003$). Overall, this experiment is consistent with our main finding that hedging-style conditions produce significant shifts, while boosting conditions show a weaker effect.

Condition	Rewriter	Δp_{sup}	Δp_{ref}
CT	gpt-oss	$+0.054$ $[0.030, 0.078]$	$+0.030$ $[0.012, 0.050]$
	Gemma	$+0.056$ $[0.032, 0.080]$	$+0.037$ $[0.020, 0.058]$
OA	gpt-oss	-0.035 $[-0.058, -0.012]$	-0.021 $[-0.039, -0.002]$
	Gemma	-0.027 $[-0.051, -0.005]$	-0.003 $[-0.022, 0.016]$
C1	gpt-oss	$+0.016$ $[-0.002, 0.035]$	$+0.018$ $[0.003, 0.033]$
	Gemma	$+0.017$ $[-0.000, 0.035]$	$+0.023$ $[0.009, 0.038]$
D1	gpt-oss	-0.003 $[-0.019, 0.013]$	-0.009 $[-0.023, 0.005]$
	Gemma	-0.007 $[-0.026, 0.012]$	-0.004 $[-0.020, 0.014]$

Table 5: Δp_{sup} and Δp_{ref} for rewrites generated by gpt-oss and Gemma-4-31B-it on the dev set, verified by Qwen3-VL-8B. Brackets show 95% bootstrap CIs (1,000 resamples). For Gemma, C1/D1 rewrites that do not insert exactly one pool word are excluded.

	Full	Unflagged	Flagged
CT Δp_{sup}	$+0.062$ $[0.044, 0.078]$	$+0.055$ $[0.037, 0.074]$	$+0.080$ $[0.051, 0.109]$
CT Δp_{ref}	$+0.045$ $[0.030, 0.059]$	$+0.044$ $[0.027, 0.061]$	$+0.046$ $[0.020, 0.075]$
C1 Δp_{sup}	$+0.017$ $[0.006, 0.028]$	$+0.016$ $[0.003, 0.029]$	$+0.020$ $[-0.002, 0.040]$
C1 Δp_{ref}	$+0.024$ $[0.014, 0.034]$	$+0.020$ $[0.008, 0.031]$	$+0.040$ $[0.016, 0.063]$

Table 6: Δp_{sup} and Δp_{ref} on the full, unflagged, and flagged subsets (Qwen3-VL-8B). Brackets show 95% bootstrap CIs (1,000 resamples).

B Semantic Change of Claims after Rewriting

As noted in the Limitations section, the model sometimes produces rewrites that alter the semantic meaning of the claim. To verify that such rewrites do not affect our conclusions, we conducted the following analysis. A rewrite is flagged if it introduces a hedge word (Table 11) within the two tokens immediately preceding one of the following elements: a number, %, a p-value, a unit, or a comparison/ranking word. This criterion flags 26.3% of CT rewrites and 22.9% of C1 rewrites, and fewer than 10% of rewrites in all other conditions.

As shown in Table 6, Δp_{sup} and Δp_{ref} on the unflagged set are similar to or slightly lower than those on the full set, while the flagged set tends to show larger deltas in most conditions. Because the unflagged set preserves the trend observed on the full set, our conclusions hold even after removing potentially semantics-altering rewrites.

C Prompts

We provide the full prompt templates used for natural rewriting (Figure 4), controlled rewriting (Figure 5), and the five prompt variants per rewriting category (Table 9).

Prompt Used for Natural Rewriting

[Rewriting Prompt] During rewriting, you must not alter the factual content of the text. All changes should be purely stylistic and decorative.

Figure 4: Prompt template used for natural rewriting. The **[Rewriting Prompt]** placeholder is instantiated with one of five rewriting categories: Language Polishing, Fluency Improvement, Academic Style Alignment, Outcome-Aware, or Cautious Tone. See Table 9 for details.

D Flip Rates Across All Models

Table 10 reports flip rates across all models and rewriting conditions, complementing the probability shift analysis in Section 5.

E Annotation Guidelines

Figure 6 presents the annotation guidelines provided to annotators for evaluating semantic consistency between original and rewritten claims, as described in Section 3.2.1.

F Word Lists

Table 11 lists the decorative and conservative words used in controlled rewriting, drawn from Kobak et al. (2025) and Matsui (2025).

G Data Exclusions

A small number of claims are excluded from the analysis for two reasons. First, the inject rewriting pipeline (D1, C1) occasionally produces an empty or malformed output; when this affects either member of a claim pair, both members are dropped to preserve the paired evaluation structure (e.g., 29 pairs excluded for D1). Natural rewrite conditions produce no such failures. Second, for a small number of claims, all 10 self-consistency responses from a given model are unparseable; these claims are excluded from that model's evaluation. Crucially, the affected claims are consistent across the baseline and all rewrite conditions for a given model, so no differential bias is introduced into the comparison. In both cases the number of excluded items is small relative to the full evaluation set of 788 pairs.

Prompt Used for Controlled Rewriting

You are a linguistic editor. Your task is to inject exactly 1 word from the provided word list into the given claim.

Rules:

- You MUST insert exactly 1 word from the word list below. This is mandatory.
- Do not delete, replace, or reorder any existing words. Every original word must remain.
- The injection should be stylistic only—it must not change the factual content.
- The injected word must be grammatically natural in its position.

Word list:
`{word_list}`

Original claim:
`{claim}`

Output ONLY the modified claim with exactly one word inserted. No explanation, no preamble.

Figure 5: Prompt template used for controlled rewriting. The `{word_list}` placeholder is instantiated with a randomly sampled subset of candidate words from either the decorative or conservative word pool, and `{claim}` is instantiated with the original claim.

H Human Annotation

To verify that rewrites preserve the meaning of the original claims, four annotators, all NLP researchers, answered a single binary question: does the rewrite keep the same meaning as the original? Annotators saw both claims side by side with the changed words highlighted, and the rewrite condition was hidden. The annotators followed the annotation guidelines (Figure 6). We sampled 80 claims each from OA, CT, C1, and D1. For CT and C1, we sampled only from the unflagged subset (Section B), since this is the subset our analysis relies on. OA and D1 were sampled at random,

Cond.	Judg.	Pres. (%)	n_2	P_o	κ	AC1
CT	96	78.1	27	0.667	0.088	0.475
C1	97	84.5	26	0.846	0.246	0.807
OA	98	74.5	27	0.741	0.291	0.591
D1	100	90.0	29	0.897	0.343	0.877
All	391	81.8	109	0.789	0.251	0.706

Table 7: Human annotation results. Judg.: number of judgments. Pres.: percentage of judgments finding the meaning preserved. n_2: claims with two judgments; P_o (observed agreement), Fleiss' κ, and Gwet's AC1 are computed on these only. CT and C1 are sampled from the unflagged subset.

since the filter targets hedging and these are boosting conditions. Each claim was assigned to two annotators; as not all assignments were completed, the returned data contain 391 judgments over 282 claims (109 with two judgments and 173 with one).

Overall, annotators judged the meaning preserved in 81.8% of judgments, ranging from 74.5% (OA) to 90.0% (D1), with 78.1% for CT and 84.5% for C1 (Table 7). On the 109 doubly annotated claims, observed agreement is 78.9%, Fleiss' κ is 0.251, and Gwet's AC1 is 0.706. Since each condition contributes only 26–29 doubly annotated claims, per-condition agreement estimates should be interpreted with caution. We use Fleiss' κ rather than Cohen's κ because different annotator pairs judged different claims, whereas Cohen's κ assumes a single fixed pair of annotators. The low κ reflects label skew: 83.0% of these judgments mark the meaning as preserved, so κ attributes most of the observed agreement to chance, whereas AC1 is robust to such skew. The remaining disagreement is largely systematic. The rate at which individual annotators judged a meaning change ranges from 4.7% to 32.7%, suggesting that assigning meaning-preservation labels to scientific claims is inherently difficult.

I Qualitative Examples

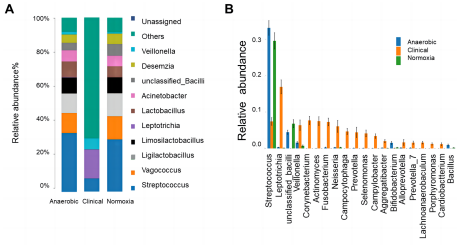
To complement the aggregate results, Table 8 presents representative examples of rewritten claims together with the predictions of Qwen3-VL-8B before and after rewriting. We include both high-shift cases, where rewriting substantially changes $p(\text{support})$, and no-shift cases, where the prediction remains stable despite the rewrite.

Case 1: high shift `val_fig_0285` Condition: CT Gold: Supported p: 0.0 → 0.9 ($\Delta p = +0.9$) Flipped: yes

Original The results indicated that there were more bacterial categories in clinical samples than cultures by media.

Rewritten The results **suggested** that **a greater number of** bacterial categories **were observed** in clinical samples than **were identified through culture** media.

Evidence

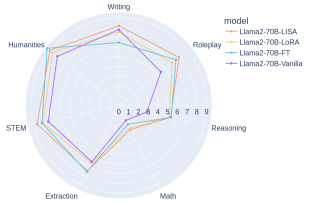


Case 2: high shift `val_fig_0052` Condition: OA Gold: Supported p: 0.9 → 0.0 ($\Delta p = -0.9$) Flipped: yes

Original It is also intriguing to observe from Figure 5 that Vanilla LLaMA-2-70B excelled in Writing, but full-parameter fine-tuning led to a decline in these areas, a phenomenon known as the "Alignment Tax" (Ouyang et al., 2022) .

Rewritten Figure 5 **compellingly demonstrates** that **the** vanilla LLaMA-2-70B **model attains peak performance on writing tasks, yet once the model is subjected to** full-parameter fine-tuning **its scores markedly decline—a clear manifestation of** the **well-documented** "Alignment Tax" (Ouyang et al., **2022).**

Evidence



Case 3: no shift `val_tab_0177` Condition: LP Gold: Supported p: 0.6 → 0.6 ($\Delta p = 0.0$) Flipped: no

Original Second, pairwise comparisons with the baseline show both CG and finetuning are effective in formality control, where CG has slightly higher win ratio than FT against the baseline.

Rewritten Second, pairwise comparisons with the baseline show **that** both CG and **fine-tuning** are effective **for** formality control, **with** CG **achieving a** slightly higher win ratio than FT against the baseline.

Evidence

	Model	Quality (1-5)	Formality (1-3)	Win (%)	Win & Tie (%)
(1)	NLLB-200	4.25±0.75	2.69±0.46	–	–
(2)	CG (`multi`) formal	4.00±0.79	2.63±0.48	56.3	81.3
(3)	CG (`multi`) inf.	4.44±0.70	2.38±0.69	62.5	93.8
(4)	FT (`multi`) formal	4.31±0.85	2.63±0.48	43.8	68.8
(5)	FT (`multi`) inf.	4.13±1.05	2.44±0.49	62.5	93.8

Table 6: Human evaluation on Bengali, with quality on a 5-point scale↑ and formality on a 3-point scale (↑: formal) with standard deviations. Last two columns show pairwise comparison of formality scores to baseline NLLB-200 given the same source sentences (winning: scoring more in the direction of the desired formality).

Case 4: no shift despite hedging `val_tab_0029` Condition: C1 Gold: Supported p: 0.6 → 0.6 ($\Delta p = 0.0$) Flipped: no

Original Results in Table 1 show that AD-Drop achieves consistent improvement, boosting the average scores of BERT base and RoBERTa base by 0.87 and 0.62, respectively.

Rewritten Results in Table 1 show that AD-Drop achieves **seemingly** consistent improvement, boosting the average scores of BERT base and RoBERTa base by 0.87 and 0.62, respectively.

Evidence

Methods	SST-2	MNLI	QNLI	QQP	CoLA	STS-B	MRPC	RTE	Average
Development									
$\text{BERT}_{\text{base}}$	92.3	84.6	91.5	91.3	60.3	89.9	85.1	70.8	83.23
+*SCAL*† miao2021simple	92.8	84.1	90.9	91.4	61.7	-	-	69.7	-
+*SuperT*† liang2021super	93.4	84.5	91.3	91.3	58.8	89.8	87.5	72.5	83.64
+*R-Drop*† wu2021r	93.0	85.5	92.0	91.4	62.6	89.6	87.3	71.1	84.06
+AD-DROP	93.9	85.1	92.3	91.8	64.6	90.4	88.5	75.1	**85.21**
$\text{RoBERTa}_{\text{base}}$	95.3	87.6	92.9	91.9	64.8	90.9	90.7	79.4	86.69
+*R-Drop* wu2021r	95.2	87.8	93.2	91.7	64.7	91.2	90.5	80.5	86.85
+*HiddenCut*† chen2021hiddencut	95.8	88.2	93.7	92.0	66.2	91.3	92.0	83.4	87.83
+AD-DROP	95.8	88.0	93.5	92.0	66.8	91.4	92.2	84.1	**87.98**
Test									
$\text{BERT}_{\text{base}}$	93.6	84.7	90.4	89.3	52.8	85.6	81.4	68.4	80.78
+AD-DROP	94.3	85.2	91.6	89.4	53.3	86.6	84.1	68.7	**81.65**
$\text{RoBERTa}_{\text{base}}$	94.8	87.5	92.8	89.6	58.3	88.7	86.3	75.1	84.14
+AD-DROP	95.9	87.6	93.4	89.5	58.5	89.3	87.9	76.0	**84.76**

Table 1: Overall results of fine-tuned models on the GLUE benchmark. The symbol † denotes results directly taken from the original papers. The best average results are shown in bold.

Table 8: Example cases for Qwen3-VL-8B-Instruct. p is the probability of *Supported* estimated from 10 sampled answers, before and after rewriting. Words changed by the rewrite are in bold. Evidence images are taken from SciClaimEval (Ho et al., 2026).

Annotation Guidelines for Semantic Consistency

I/O:
You will be shown pairs of scientific claims: an original and a rewritten version.

Input: original claim, rewritten claim, evidence image path (for reference)
Output: a binary label (TRUE/FALSE), and an optional note field

Task:
Does the rewritten claim preserve the same factual meaning as the original?

- **TRUE:** The core factual content is the same. Differences in word choice, tone, emphasis, or formality are fine.
- **FALSE:** The rewrite adds, removes, or changes a factual detail (e.g., a number, a comparison direction, or which method/dataset is referenced).
- **If you can't decide:** Imagine verifying the claim against some evidence (table, figure, etc.). Ask: does there exist any possible evidence that would support the original claim but not the rewritten one, or vice versa? If yes → FALSE. If no → TRUE.

Examples:

- "our method is 20% better" → "our method is 25% better" → **FALSE** (changed number)
- "our method is 20% better" → "our method is 20.0001% better" → **FALSE** (even a trivial numerical change could be invalidated by evidence showing exactly 20%)
- "A is 20% better than B" → "A is approximately 20% better than B" → **FALSE** ("approximately" widens the range of valid values)
- "our method is 20% better" → "our fantastic method is 20% better" → **TRUE** (subjective—no evidence can distinguish "fantastic" from "not fantastic")
- "consistently outperforms" → "arguably outperforms" → **TRUE** (tone shift, but the factual claim is the same)
- "outperforms method A" → "outperforms method B" → **FALSE** (different reference)

Figure 6: Annotation guidelines provided to annotators for evaluating semantic consistency between original and rewritten claims.

Category	Variant	Prompt
Language Polishing (LP)	LP-1	Fix any grammar mistakes in the following text.
	LP-2	Proofread this text and correct all grammatical errors.
	LP-3	Can you fix the grammar of this text?
	LP-4	Please correct any grammatical issues in the text below.
	LP-5	Review this text for grammar problems and rewrite it with corrections.
Fluency Improvement (FI)	FI-1	Rewrite this text to make it more fluent and readable.
	FI-2	Make this paragraph flow more naturally.
	FI-3	Improve the readability of the following text.
	FI-4	Smooth out the writing in this text so it reads better.
	FI-5	Rephrase this text to improve its clarity and flow.
Academic Style Alignment (AS)	AS-1	Rewrite this to match the writing style expected at ACL.
	AS-2	Adapt this text to fit the conventions of a AAAI paper.
	AS-3	I am submitting this paper to NeurIPS. Rewrite it to match the style and quality expected there.
	AS-4	Revise this text so it reads like a typical ICLR submission.
	AS-5	Adjust the tone and style of this text to align with papers published at ICML.
Outcome-Aware (OA)	OA-1	Rewrite this to be more impressive and compelling to reviewers.
	OA-2	Polish this so it sounds more convincing to reviewers.
	OA-3	Rewrite this to enhance the review score and maximize acceptance chances.
	OA-4	Make this text as persuasive and impactful as possible for peer review.
	OA-5	Strengthen the writing so that reviewers would strongly favor this paper.
Cautious Tone (CT)	CT-1	Rewrite this text to adopt a more cautious and measured tone, softening strong claims and avoiding overgeneralization.
	CT-2	Rewrite this text to use a more careful and qualified tone, tempering strong assertions and introducing appropriate nuance.
	CT-3	Revise this passage to emphasize balance and restraint, moderating claims and avoiding overly definitive language.
	CT-4	Refine this passage to adopt a more reserved academic tone, softening conclusions and avoiding broad or sweeping statements.
	CT-5	Rewrite this to sound more analytically cautious, reducing emphasis on certainty and acknowledging potential variability in interpretation.

Table 9: Prompt strategies for natural rewriting experiments. Each category contains five prompt variants to control for variance introduced by prompt phrasing.

Model		Natural Rewriting					Controlled Rewriting	
		LP	FI	AS	OA	CT	D1	C1
gemma-4-26B-A4B-it	Supported	4.9%	5.2%	7.9%	8.9%	7.9%	4.5%	5.1%
	Refuted	6.7%	7.0%	8.5%	8.9%	11.5%	5.7%	5.5%
gemma-4-31B-it	Supported	2.9%	3.2%	3.9%	6.1%	4.2%	3.0%	2.4%
	Refuted	5.7%	6.5%	7.9%	9.6%	11.8%	5.1%	7.2%
GLM-4.6V-Flash	Supported	6.7%	7.5%	8.5%	11.0%	10.8%	4.9%	6.5%
	Refuted	7.9%	9.5%	11.4%	10.8%	15.5%	5.5%	6.4%
InternVL3-2B	Supported	5.5%	5.7%	5.6%	6.0%	4.7%	5.1%	4.7%
	Refuted	7.4%	8.0%	6.5%	8.6%	7.7%	7.0%	5.9%
InternVL3-8B	Supported	5.2%	4.9%	5.3%	5.2%	5.8%	4.3%	3.2%
	Refuted	11.9%	10.8%	11.3%	13.6%	14.7%	11.2%	11.5%
InternVL3-38B	Supported	7.8%	7.1%	7.8%	8.7%	6.7%	6.7%	6.0%
	Refuted	11.5%	11.2%	12.2%	14.7%	15.1%	10.0%	11.1%
Kimi-VL-A3B	Supported	6.1%	9.8%	8.9%	10.3%	8.0%	7.0%	6.7%
	Refuted	11.3%	14.8%	13.5%	14.3%	16.1%	9.4%	11.6%
Qwen3-VL-2B-Instruct	Supported	10.9%	10.9%	11.3%	13.7%	15.4%	12.8%	11.4%
	Refuted	12.3%	16.8%	15.9%	16.7%	15.3%	11.7%	12.7%
Qwen3-VL-4B-Instruct	Supported	11.3%	11.6%	12.1%	14.8%	15.1%	7.9%	9.7%
	Refuted	10.8%	9.9%	12.1%	12.7%	12.7%	8.6%	8.8%
Qwen3-VL-8B-Instruct	Supported	9.3%	10.4%	12.1%	14.8%	13.5%	7.8%	8.7%
	Refuted	7.6%	8.3%	8.8%	10.2%	11.8%	8.2%	7.2%
Qwen3-VL-32B-Instruct	Supported	6.0%	6.2%	8.9%	10.7%	9.0%	6.3%	6.9%
	Refuted	7.0%	6.6%	8.0%	9.8%	10.7%	5.9%	7.2%

Table 10: Flip rate: percentage of claims whose predicted label (threshold 0.5 on $p(\text{support})$) changes after rewriting, in either direction. Rows group claims by gold label: *Supported* (claim paired with supporting evidence) and *Refuted* (claim paired with refuting evidence). Colour intensity scales with flip rate.

Category	Words
Conservative (C)	acknowledges, allegedly, apparently, approximately, arguably, basically, broadly, comparatively, conceivably, essentially, generally, hinting, inadequately, largely, mainly, marginally, moderately, mostly, often, ostensibly, partially, perhaps, plausibly, possibly, potential, presumably, primarily, purportedly, relatively, roughly, seemingly, slightly, somewhat, supposedly, tentatively, typically, underexplored, unexplored, usually
Decorative (D)	accentuates, adept, advocates, affirming, aptly, attains, augmenting, avenue, boast, bolster, burgeoning, capitalizing, catalyze, combating, commendable, compelling, compellingly, comprehensive, consolidates, crafted, critical, crucial, culminating, deeper, dependability, dependable, detrimentally, disrupts, distinctive, effectively, effortlessly, elevate, emphasises, emphasising, emphasize, empowers, enduring, enhance, essential, excel, excellently, exceptional, exceptionally, exhaustive, expansive, formidable, fortify, foster, foundational, fresh, fundamental, groundbreaking, illuminate, imperative, impressive, impressively, ingenious, innovative, intriguing, invaluable, leverage, lucidly, meticulous, meticulously, multifaceted, notable, notably, noteworthy, optimizing, orchestrating, outperform, paving, pinpoint, pioneering, pivotal, poised, potent, pressing, promise, pronounced, propelling, remarkable, renowned, revolutionize, seamless, seamlessly, showcase, significant, solidify, strategically, streamline, substantiated, surged, surmount, surpass, swift, swiftly, thorough, transcend, transform, transformative, uncharted, uncovering, underscore, unlocking, unparalleled, unraveling, unveil, uphold, valuable, versatile, versatility, warranting, well-rounded

Table 11: Word pools used in inject rewriting. These words are frequently included by LLMs during writing and serve as recognizable markers of LLM-generated text. Decorative words are sourced from Kobak et al. (2025) and Matsui (2025); conservative words are drawn from the same sources and supplemented with LLM-assisted hedging terms. Category C (Conservative) contains 39 words; Category D (Decorative) contains 114 words.