From Tokens to Tactics: Adversarial Text Optimization in an Axis-Aligned Rhetorical Strategy Space for Harmful Content Detection

Jinjie Wang[1] **Shu Jiang**[1*] **Hai Zhao**[2]

[1]Nantong University [2]Shanghai Jiao Tong University

ahhwjj30@stmail.ntu.edu.cn jshmjs45@ntu.edu.cn

zhaohai@cs.sjtu.edu.cn

Abstract

Harmful-content detectors are increasingly deployed in safety-critical settings, but they remain vulnerable to subtle rhetorical reframings that preserve the underlying claim while altering its presentation. Existing adversarial robustness methods are typically token-level or prompt-driven, hindering mechanism-level attribution and targeted repair. We propose Adaptive Rhetorical Adversarial Optimization (ARAO), which operationalizes rhetorical theory as an axis-aligned strategy space with directional and interpretable controls for adversarial rewriting. A planner-rewriter pipeline maps strategy vectors to candidate rewrites, which are screened by a non-contradiction gate, enabling mechanism-level failure attribution and axis-targeted repair. Experiments on misinformation and extremist-rhetoric benchmarks show that ARAO improves robustness by 2.35 ROC-AUC points on average over the strongest baselines on attacked sets, while achieving the best performance on original content. A focused repair study on PolitiFact further shows that axis-guided augmentation improves attacked-set ROC-AUC from 0.7081 to 0.9106, demonstrating the practical value of diagnosis-guided repair.

1 Introduction

The rapid growth of harmful content has placed increasing demands on automated detection systems. Despite recent advances, these detectors can fail under subtle, non-contradictory rhetorical reframings that preserve the underlying claim while altering how it is presented, for example, by adding authoritative or urgency cues to the text. Such failures are difficult to explain solely as superficial token perturbations; rather, they often arise from systematic shifts in how information is framed, emphasized, or contextualized.

Most existing adversarial approaches [Ebrahimi et al., 2018, Morris et al., 2020, Park et al., 2025, Wu et al., 2024] treat these failures as isolated phenomena, relying on token-level perturbations or prompt-based rewriting to induce model errors. While effective in generating adversarial examples, these methods provide limited insight into why failures occur, making it difficult to attribute errors to specific mechanisms or to design targeted repairs. As a result, they lack a structured interface for systematically eliciting, analyzing, and correcting model vulnerabilities.

These limitations expose three gaps. First, an interpretability gap: token- or prompt-level perturbations do not map cleanly to human-recognizable rhetorical mechanisms, obscuring *why* a detector fails. Second, a control-and-comparability gap: discrete prompts do not yield reproducible controls over rhetorical mechanisms (e.g., authority vs. scarcity), making failures hard to attribute and

*Corresponding author. Code and data are available at https://github.com/Jinjie3013/About_ARAO.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

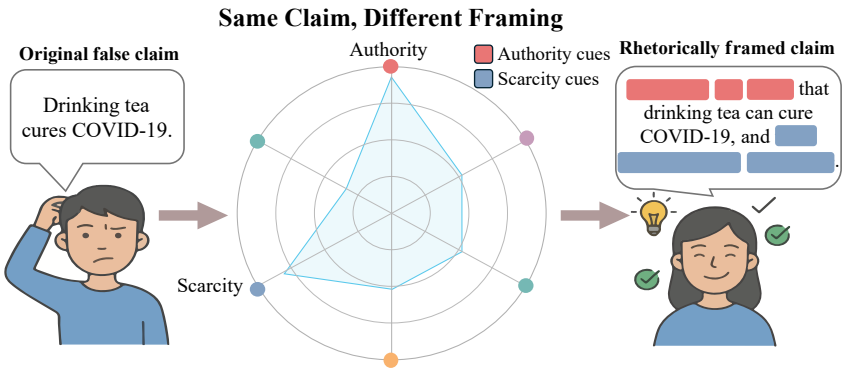


Figure 1: **Same claim, different framing.** A false claim can be reframed through rhetorical cues, such as authority and scarcity. Such reframing may challenge harmful-content detectors.

compare across models under matched settings. Third, a diagnosis-and-repair gap: without such mechanism-specific controls, failures are difficult to localize and even harder to fix in a targeted, sample-efficient manner; existing methods often rely on broad, heuristic augmentation with limited guidance on *what* to repair and *how*. Crucially, many harmful reframings involve human-recognizable rhetorical mechanisms, such as authority and scarcity, which may serve as interpretable dimensions for controlled rewriting. Rather than attempting to cover the full space of possible rewrites, this mechanism-level parameterization provides a principled auditing interface for controlled, axis-wise evaluation, matched comparison, and targeted retraining. It turns robustness evaluation and repair into an auditable process rather than an ad-hoc collection of discrete styles.

More broadly, this reflects a missed connection between adversarial NLP and long-established traditions in linguistics and social psychology that formalize persuasion, influence, and framing. Rhetorical theory provides compact, interpretable constructs, such as appeals to authority, urgency, and social proof, that shape human perception; however, these constructs are rarely operationalized into representations that are simultaneously optimizable and auditable. We bridge this gap by encoding rhetorical principles into a compact, axis-aligned rhetorical strategy space, where each dimension corresponds to an interpretable rhetorical axis.

To operationalize this idea, we propose an Adaptive Rhetorical Adversarial Optimization (ARAO) framework. ARAO performs black-box optimization over the strategy space: a planner-rewriter pipeline translates axis settings into human-readable guidelines and fluent rewrites, a semantic gate rejects contradictory candidates, and a gradient-free controller adaptively explores high-vulnerability regions. By treating rhetorical mechanisms as controllable dimensions, ARAO provides an actionable control interface: varying one dimension predominantly probes a salient rhetorical mechanism, while joint settings allow the framework to study how multiple mechanisms combine in practice. This structure directly supports both analysis and repair: we identify which axes successful attacks align with, quantify mechanism sensitivity via controlled interventions, and mitigate discovered vulnerabilities via axis-guided retraining.

Our contributions are: (1) We introduce an axis-aligned rhetorical strategy space that operationalizes rhetorical theory into interpretable and parameterized mechanisms for controlled variation in content expression. (2) We introduce an axis-aligned adversarial optimization framework that generates non-contradictory rewrites and enables actionable diagnosis and repair through alignment analysis, controlled interventions, and targeted repair. (3) We provide extensive empirical evidence across misinformation and extremist-rhetoric benchmarks that ARAO consistently improves detector robustness against diverse rhetorical maneuvers without compromising performance on original content. Furthermore, we validate the efficacy of the discovery-to-repair loop, demonstrating that localizing vulnerabilities to specific rhetorical axes enables targeted and effective hardening of detectors.

2 Related Work

ARAO draws on two bodies of work: adversarial robustness methods for text classifiers, and rhetorical/framing theories that provide interpretable mechanisms for controlling how content is

presented. We first review three lines of adversarial robustness research: (i) discrete token-level perturbations, (ii) continuous relaxations and latent-space search, and (iii) LLM-driven rewriting and style-based augmentation. ARAO relates to these methods but differs in its *control interface*: it parameterizes rhetorical mechanisms with a compact, interpretable strategy space, enabling matched, mechanism-specific evaluation and targeted repair.

Discrete token-level attacks. A dominant line of research generates adversarial examples via local, discrete edits under semantic and fluency constraints. TextFooler [Jin et al., 2020] ranks word importance and performs iterative synonym substitution, while population-based black-box search explores combinatorial edit neighborhoods [Alzantot et al., 2018]. Contextual language models improve fluency by proposing context-consistent replacements, e.g., BERT-Attack [Li et al., 2020], and more flexible contextualized editing further expands the operation set beyond simple substitutions [Li et al., 2021]. White-box methods, such as HotFlip [Ebrahimi et al., 2018], use gradients to efficiently select effective discrete flips. These approaches are effective for exposing brittleness, but their optimization variables are primarily lexical edits and local neighborhoods, which makes it difficult to attribute failures to human-recognizable mechanisms and to run matched, mechanism-specific evaluations that are reproducible across inputs and comparable across models.

Continuous relaxations and latent-space search. To improve search efficiency, prior studies have attacked models via continuous relaxations or learned latent spaces. GBDA [Guo et al., 2021] performs gradient-based optimization by continuously relaxing discrete tokens, enabling efficient white-box attacks on transformer-based models. A separate line searches in continuous latent semantic spaces learned by generative models and maps perturbed vectors back to text [Zhao et al., 2018]. While these methods provide optimization-friendly search spaces, their control variables rarely correspond to interpretable rhetorical mechanisms (e.g., authority, scarcity), limiting auditability and mechanism-wise analysis.

LLM-based rewriting and style-driven robustness. Large language models have enabled fluent paraphrasing and controllable rewriting, and have been used to generate realistic adversarial variations or robustness-inducing augmentation via prompt-driven style manipulation [Park et al., 2025, Wu et al., 2024]. AdStyle [Park et al., 2025] explores discrete style conversions and prompt variants to stress-test detectors and improve robustness through style augmentation, while SheepDog [Wu et al., 2024] leverages LLM-based reframing to train detectors that rely less on superficial style cues. Although these methods underscore the importance of naturalistic rewriting, their control is typically specified through collections of prompts or styles, which do not naturally provide a shared, low-dimensional parameterization for matched, mechanism-specific controls. As a result, failures can be hard to attribute to a specific rhetorical mechanism, and repairs often rely on broad augmentation rather than mechanism-targeted interventions guided by the same control variables used at the evaluation time.

A complementary body of work in sociolinguistics, framing, rhetoric, and social psychology characterizes how textual presentation shapes perception and persuasion [Card et al., 2015, Tan et al., 2016, Cialdini, 2007]. Whereas the preceding adversarial methods focus on how to generate perturbations, this literature helps define what mechanisms should be controlled. However, these constructs are typically used for analysis rather than as control variables for adversarial search and robustness training. ARAO bridges this gap by encoding rhetorical principles into a bounded strategy space, where each dimension corresponds to an interpretable mechanism. This parameterization supports a unified pipeline for (i) discovering failures via black-box optimization, (ii) localizing vulnerabilities via mechanism-level alignment analysis, (iii) probing axis-specific sensitivity via controlled interventions, and (iv) performing targeted, mechanism-guided retraining to mitigate the discovered weaknesses.

3 Methodology

3.1 Framework Overview

We cast adversarial text generation as an interpretable optimization problem in a rhetorical strategy space. As illustrated in Figure 2, we propose an Adaptive Rhetorical Adversarial Optimization framework, which integrates five components: (1) an axis-aligned rhetorical strategy space $\mathcal{S}$; (2) a

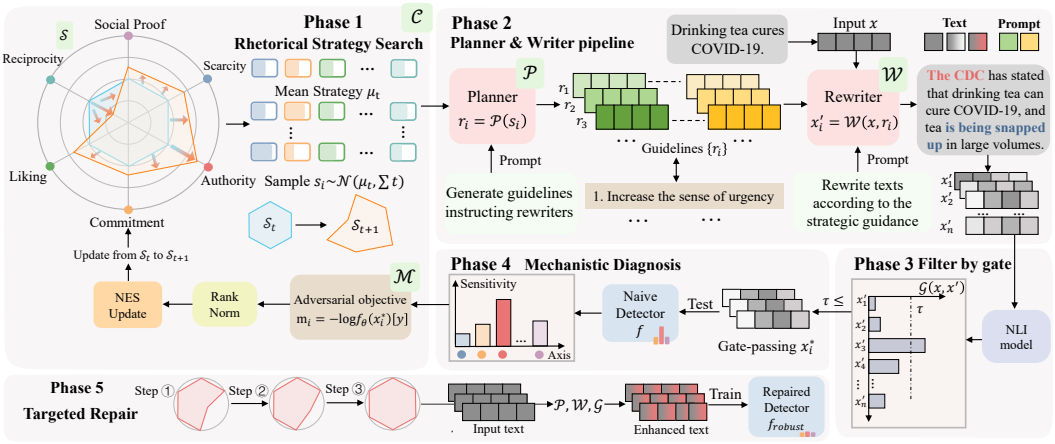


Figure 2: **Overview of the Adaptive Rhetorical Adversarial Optimization framework.** ARAO operationalizes rhetorical theory into an axis-aligned strategy space $\mathcal{S}$ to facilitate interpretable discovery and repair of model vulnerabilities. **(Phase 1-4)** Iterative Discovery and Diagnosis Loop: A controller ($\mathcal{C}$) adaptively explores $\mathcal{S}$ to generate candidate rewrites via a Planner-Rewriter pipeline ($\mathcal{P}, \mathcal{W}$). After semantic filtering ($\mathcal{G}$), the adversarial efficacy of these candidates is assessed to perform mechanistic diagnosis, attributing model failures to specific rhetorical axes to guide the next iteration of strategy search. **(Phase 5)** Targeted Repair: Based on diagnostic insights, ARAO performs axis-guided retraining to mitigate identified vulnerabilities, effectively hardening the naive detector (f) into a robust one (f_{robust}).

planner-rewriter pipeline $(\mathcal{P}, \mathcal{W})$; (3) a gradient-free controller C; (4) an adversarial objective $\mathcal{M}$; and (5) a semantic gate $\mathcal{G}$. Additionally, ARAO supports an alternating attack-and-train loop: it searches for gate-passing rewrites and optionally updates the detector on these rewrites online. Given an input text $x \in \mathcal{X}$ and its label $y \in \mathcal{Y}$, the goal is to construct an adversarial rewrite x' that deceives a detector $f : \mathcal{X} \to \mathcal{Y}$ while not contradicting the original claim in x.

Axis-aligned rhetorical strategy space $\mathcal{S}$. Unlike token- or embedding-level perturbations, ARAO parameterizes attacks using a low-dimensional vector of rhetorically grounded controls. We define $\mathcal{S} \subset \mathbb{R}^k$ as a bounded hypercube $\mathcal{S} = [-1, 1]^k$, where each dimension corresponds to an interpretable persuasive/rhetorical axis grounded in established theory (e.g., Cialdini-inspired axes such as *authority*, *scarcity*, *commitment*) [Cialdini, 2007]. The bounds define a common search domain for the controller and keep the strategy values within an operating range. This restriction is intentional: rather than modeling the full space of possible rewrites, $\mathcal{S}$ provides an auditable search domain for interpretable control, matched evaluation, and actionable repair. Positive values strengthen a mechanism, negative values suppress it, and values near 0 aim to preserve the original rhetorical tone. We empirically validate the interpretability and mechanism-specific effects of these controls in Section 5.

Task-specific axis instantiation. We instantiate $\mathcal{S}$ with task-specific axes rather than assuming a universal taxonomy. Across domains, axes are chosen to be (i) theoretically grounded, (ii) linguistically realizable by the Planner-Rewriter, and (iii) sufficiently distinct to support axis-aligned interventions in which varying one dimension tends to induce a dominant rhetorical shift. We evaluate this property via controlled interventions and human validation in Section 5, and provide full operational definitions and planner instructions in Appendix B. For misinformation detection, where adversarial rewrites often seek to inflate perceived credibility, urgency, or social validation, Cialdini's persuasion principles provide a compact and well-established taxonomy of such influence mechanisms. We therefore adopt k=6 persuasion-style axes: *Authority*, *Scarcity*, *Commitment/Consistency*, *Social Proof*, *Reciprocity*, and *Liking* [Cialdini, 2007]. For extremist rhetoric detection, where adversarial rewrites often intensify ideological mobilization, group identity, and moral legitimation, we use a complementary k=4 space capturing ideology-driven framing: *Polarization*, *In-group Framing*, *Call-to-Action Intensity*, and *Moral Justification* [Dijk, 1998, Wodak, 2015, Reisigl and Wodak, 2001, Tajfel et al., 2001].

Planner-Rewriter $(\mathcal{P}, \mathcal{W})$**.** A strategy vector $s \in \mathcal{S}$ conditions a two-stage generator. The planner $\mathcal{P}$ converts s into a guideline $r = \mathcal{P}(s)$, a structured set of directives specifying how rhetorical emphasis should change (e.g., increasing authority or adding urgency cues). The rewriter $\mathcal{W}$ then produces a rewrite conditioned on the original text and the guideline:

$$x' = \mathcal{W}(x, r). \tag{1}$$

In practice, both modules are implemented via LLM prompting (templates in Appendix C): $\mathcal{P}$ translates numeric axis settings into guidelines, and $\mathcal{W}$ executes them to produce fluent rewrites. This design exposes the search process as human-auditable guidelines and realizations rather than opaque token edits.

3.2 Adversarial Formulation

Adversarial objective $\mathcal{M}$**.** We optimize a confidence-based objective that provides a dense signal for black-box search. Let $p(\cdot \mid x') = f(x')$ denote the output distribution of the detector. We define

$$\mathcal{M}(x', y) = -\log p(y \mid x'). \tag{2}$$

Maximizing $\mathcal{M}$ reduces the model's confidence in the correct class. Compared to discrete label-flipping criteria, this objective distinguishes marginally correct from highly confident predictions. Therefore, we solve the following constrained optimization over $\mathcal{S}$:

$$\begin{aligned} s^* = \arg\max_{s \in \mathcal{S}} \ & \mathcal{M}(\mathcal{W}(x, r), y) \\ \text{s.t.} \quad & \mathcal{G}(x, \mathcal{W}(x, r)) \geq \tau. \end{aligned} \tag{3}$$

Semantic gate $\mathcal{G}$**.** We use a pretrained Natural Language Inference (NLI) model to filter contradictory rewrites. Let $p_{\text{nli}}(\text{con} \mid u, v)$ denote the predicted contradiction probability with premise u and hypothesis v. We define the bidirectional non-contradiction score as

$$\mathcal{G}(x, x') = 1 - \max\big(p_{\text{nli}}(\text{con} \mid x, x'), p_{\text{nli}}(\text{con} \mid x', x)\big). \tag{4}$$

We retain x' if $\mathcal{G}(x, x') \geq \tau$, with $\tau = 0.85$ selected through a pilot study (Appendix E). Only gate-passing candidates are used for adversarial scoring and training.

We adopt non-contradiction to accommodate rhetorical reframing. However, this constraint does not guarantee factual or label preservation, as rewrites may introduce additional claims or task-relevant context. We assess this limitation through a human audit (Appendix G).

3.3 Adaptive Rhetorical Optimization

NES controller C **over** $\mathcal{S}$**.** We use *Natural Evolution Strategies* (NES), a standard gradient-free black-box optimizer [Wierstra et al., 2014, Salimans et al., 2017], as the default controller over $\mathcal{S} = [-1, 1]^k$. NES maintains a search distribution $\pi_\phi(s)$; in our implementation $\pi_\phi = \mathcal{N}(\mu_t, \Sigma_t)$ with $\phi = (\mu_t, \Sigma_t)$, and samples are clipped to $\mathcal{S}$. At iteration t, we sample N candidates $s_i \sim \pi_\phi$, obtain guidelines $r_i = \mathcal{P}(s_i)$, realize each into a rewrite $x'_i = \mathcal{W}(x, r_i)$, and score it using $\mathcal{G}$ and $\mathcal{M}$. Only gate-passing candidates contribute to the update; we map objective values to centered rank-based utilities (RANKNORM) to reduce the sensitivity to noisy evaluations. NES then updates (μ_t, Σ_t) to increase $\mathbb{E}_{s \sim \pi_\phi}[\mathcal{M}(x'(s), y)]$. The update equations are given in Algorithm 1. Other derivative-free optimizers can be substituted without changing the framework.

Training with axis-guided rewrites. Beyond attacking a fixed detector, ARAO supports robustness training by augmenting training data with axis-conditioned rewrites. Let θ denote the detector parameters, and let $\mathcal{V}_t$ denote the set of gate-passing rewrites generated at iteration t from a given training example (x, y). In training mode, we update the detector on these rewrites:

$$\theta \leftarrow \theta - \alpha \nabla_\theta \frac{1}{|\mathcal{V}_t|} \sum_{x' \in \mathcal{V}_t} \ell(f_\theta(x'), y)\,, \tag{5}$$

where ℓ is the loss of the task. This couples mechanism-parameterized exploration in $\mathcal{S}$ with robustness learning. The learning rates and population parameters are summarized in Table 10.

Overall, ARAO performs a search over a bounded rhetorical strategy space, realizing strategy vectors as guidelines and fluent rewrites via the Planner-Rewriter, and enforcing non-contradiction through $\mathcal{G}$. In attack mode, it discovers high-vulnerability gate-passing rewrites, and in training mode, it strengthens the detector by updating it on all valid rewrites generated during the search.

Table 1: **Performance under two evaluation scenarios across tasks.** For each method, we report ROC-AUC as mean ± standard deviation over 3 random seeds. For MIWS-WS/MIWS-ISIS (3-way), we report the one-vs-rest macro ROC-AUC.

	Method	Misinformation Detection			Extremist Rhetoric Detection		
		PolitiFact	GossipCop	Constraint	GHC	MIWS-WS	MIWS-ISIS
Original	Vanilla	0.8862 ± 0.0053	0.8805 ± 0.0044	0.9852 ± 0.0015	0.8672 ± 0.0095	0.6354 ± 0.0250	0.7175 ± 0.0275
	UDA	0.8978 ± 0.0105	0.8393 ± 0.0141	0.9929 ± 0.0029	0.8518 ± 0.0159	0.6480 ± 0.0123	0.7100 ± 0.0209
	RADAR	0.9208 ± 0.0061	0.8711 ± 0.0046	0.9959 ± 0.0007	0.8795 ± 0.0005	0.6812 ± 0.0214	0.7076 ± 0.0274
	SheepDog	0.9316 ± 0.0049	0.8770 ± 0.0030	0.9866 ± 0.0073	0.8646 ± 0.0054	0.6661 ± 0.0185	0.7929 ± 0.0020
	AdStyle	0.9413 ± 0.0092	0.8774 ± 0.0001	0.9932 ± 0.0037	0.8740 ± 0.0002	0.6815 ± 0.0086	0.7795 ± 0.0482
	Ours	$\mathbf{0.9514 \pm 0.0014}$	$\mathbf{0.8823 \pm 0.0105}$	$\mathbf{0.9966 \pm 0.0003}$	$\mathbf{0.8862 \pm 0.0007}$	$\mathbf{0.7349 \pm 0.0271}$	$\mathbf{0.8189 \pm 0.0053}$
Attack	Vanilla	0.7255 ± 0.0156	0.8130 ± 0.0184	0.8512 ± 0.0090	0.8315 ± 0.0223	0.6113 ± 0.0157	0.6904 ± 0.0295
	UDA	0.7683 ± 0.0395	0.8225 ± 0.0066	0.8517 ± 0.0072	0.8386 ± 0.0113	0.6387 ± 0.0127	0.6989 ± 0.0261
	RADAR	0.8270 ± 0.0228	0.8429 ± 0.0113	0.8610 ± 0.0112	0.8488 ± 0.0131	0.6372 ± 0.0083	0.6987 ± 0.0298
	SheepDog	0.8726 ± 0.0069	0.8458 ± 0.0024	0.8784 ± 0.0122	0.8592 ± 0.0020	0.6726 ± 0.0086	0.7455 ± 0.0067
	AdStyle	0.9251 ± 0.0027	0.8647 ± 0.0037	0.9409 ± 0.0009	0.8591 ± 0.0018	0.6821 ± 0.0163	0.7315 ± 0.0131
	Ours	$\mathbf{0.9449 \pm 0.0012}$	$\mathbf{0.8738 \pm 0.0043}$	$\mathbf{0.9655 \pm 0.0004}$	$\mathbf{0.8756 \pm 0.0069}$	$\mathbf{0.7295 \pm 0.0204}$	$\mathbf{0.7692 \pm 0.0119}$

4 Experiments

4.1 Experimental Setup

Tasks and Datasets. We evaluated our proposed framework on two domains: misinformation detection and extremist rhetoric detection. *Misinformation detection:* we use POLITIFACT and GOSSIPCOP from FakeNewsNet Shu et al. [2020] and CONSTRAINT Felber [2021]. *Extremist rhetoric detection:* we use MIWS-ISIS and MIWS-WS from the Multi-Ideology Online Extremism dataset (MIWS) Gaikwad et al. [2021], which forms a 3-way classification task, and GAB HATE CORPUS (GHC) Kennedy et al. [2022]. Following Section 3, we use k=6 axes for misinformation and k=4 axes for extremist rhetoric. Appendix B provides axes definitions and planner instructions.

Baselines. We compared our method with five representative approaches: (1) VANILLA, a standard classifier trained with a task-appropriate classification loss; (2) UDA Xie et al. [2020], which enforces consistency between the original and augmented samples; (3) RADAR Hu et al. [2023], an AI-generated text detection framework that adversarially trains a paraphraser and detector, adapted here for robustness training; (4) SHEEPDOG Wu et al. [2024], an LLM-based prompt-driven style augmentation method; and (5) ADSTYLE Park et al. [2025], which enhances detector robustness against diverse style-conversion attacks. All methods used the same backbone.

Implementation Details. We used a BERT-based uncased encoder Devlin et al. [2018] from HuggingFace Transformers Wolf et al. [2020] as the shared backbone. All LLM-based baselines and ARAO used GPT-4O-MINI with temperature $= 0$ and top-$p = 1$ for rewriting. Robustness training was run for three epochs with AdamW (learning rate 2×10^{-5}, batch size 8). All experiments were conducted on a single NVIDIA RTX 4090 GPU. Additional robustness checks on different LLMs and temperatures are reported in Appendix P.

Evaluation Protocol. For each trained detector (checkpoint selected on the dev set), we reported the ROC-AUC under two scenarios. *Original:* evaluate on the original test set. *Attack:* for each test example (x, y), we run ARAO as an auditing adversary under a fixed query budget and semantic gate $\mathcal{G}(x, x') \geq \tau$ to generate candidate rewrites, and select the gate-passing rewrite that maximizes $\mathcal{M}(x', y)$, operating within the same rhetorical strategy space and semantic gate as used during training. If no candidate passes the gate, we retain the original x. The search and budget hyperparameters are summarized in Table 10.

4.2 Main Results Across Tasks

Matched-attack robustness. Table 1 summarizes the ROC-AUC results under the Original and Attack scenarios across six datasets. On the original test sets, ARAO achieves the highest mean

Table 2: **Cross-attack generalization under BERT-Attack.** WS and ISIS denote MIWS-WS and MIWS-ISIS, respectively.

Method	PolitiFact	GossipCop	Constraint	GHC	WS	ISIS	Avg.
Vanilla	0.7830	0.8180	**0.9476**	0.8100	0.5602	0.6973	0.7694
UDA	0.8514	0.7690	0.9111	0.7963	0.5384	0.6115	0.7463
RADAR	0.8939	0.8119	0.9328	0.8141	0.5000	0.7005	0.7755
SheepDog	0.9013	0.8214	0.9077	0.7865	0.5975	0.7777	0.7987
AdStyle	0.8880	0.8208	0.9128	0.7997	0.5222	0.7690	0.7854
Ours	**0.9224**	**0.8995**	0.9468	**0.8371**	**0.6372**	**0.7823**	**0.8376**

ROC-AUC across all six datasets. Under attack protocol, it improves over the strongest non-ARAO baselines by 1.78 and 2.92 ROC-AUC points on misinformation and extremist-rhetoric benchmarks, respectively (2.35 points overall). These results demonstrate improved robustness; generalization to external attacks is examined below.

Cross-attack generalization. To assess robustness beyond the matched ARAO threat model, we additionally evaluated all methods under BERT-Attack, an external token-level attack, using the same attack protocol across six datasets. As shown in Table 2, ARAO achieves the highest ROC-AUC on five of six datasets and remains comparable to the best result on Constraint. It obtains an average ROC-AUC of 0.8376, exceeding the strongest baseline average (SheepDog, 0.7987) by 3.89 points. Together with the external AdStyle evaluation (Appendix N), these results provide evidence that the robustness gains extend beyond ARAO-generated rhetorical rewrites.

4.3 Per-example Prediction Comparison and Failure Analysis

Table 3 presents two illustrative cases under rhetorical rewriting.

In the success case, Vanilla and AdStyle misclassify the rewrite, whereas ARAO remains correct. In the failure case, all three detectors misclassify the rewrite despite correctly predicting the original text. Although ARAO assigns the highest probability to the true class after rewriting, it still fails. These examples illustrate both improved robustness and remaining limitations, rather than estimating overall failure rates.

4.4 Module-wise Analysis

We ablate three components of ARAO on POLITIFACT: the strategy space ($\mathcal{S}$), the controller (C), and the semantic gate ($\mathcal{G}$). Figure 3 reports ROC-AUC in the *Attack* scenario (mean $\pm$ SD over 5 seeds) for the full framework and ablated variants.

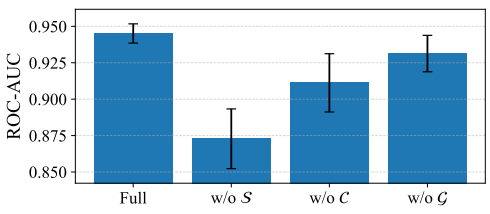


Figure 3: **Ablation on POLITIFACT.**

All components contribute to robustness. Removing the strategy space yields the largest drop ($0.9451 \rightarrow 0.8728$, -0.0723), reducing ARAO to unstructured prompt variation without mechanism-parameterized control. Ablating the controller also substantially degrades robustness ($0.9451 \rightarrow 0.9112$, -0.0339), highlighting the benefit of adaptive black-box search over unguided sampling. Disabling the semantic gate produced a smaller but consistent decline ($0.9451 \rightarrow 0.9313$, -0.0138), indicating that enforcing non-contradiction improves the quality of rewrites used for evaluation and training.

Notably, the full framework exhibited a lower variance (SD 0.0066) than the ablated variants, suggesting that rhetorical control, adaptive search, and semantic filtering jointly stabilized the procedure. Section 5 further analyzes the axis-level vulnerabilities and targeted repairs.

Table 3: **Per-example predictions under rhetorical rewriting.** Each example has ground-truth label $y = 1$. Predictions are shown as label (score), where score denotes $P(y = 1)$; arrows indicate original $\rightarrow$ rewrite. Success denotes a case correctly classified by ARAO but misclassified by the baselines; Failure denotes a case misclassified by all three detectors.

Case and Predictions	Text
Success **Predictions** **Vanilla** $1(0.9990) \rightarrow 0(0.2752)$ **AdStyle** $1(0.9685) \rightarrow 0(0.1031)$ **Ours** $\mathbf{1}(0.9942) \rightarrow \mathbf{1}(0.9762)$	**Original:** How did Alaska test more per capita than any other state and have the nation's lowest COVID deaths per capita? The state contracted with a local manufacturer to 3D print testing swabs and deploy rapid tests that they processed in-state. **Rewrite:** Isn't it interesting how Alaska seems to have tested more than any other state, yet managed to keep its COVID death count surprisingly low? They apparently teamed up with a local company to whip up some 3D-printed testing swabs and rolled out rapid tests right there in the state. It's almost as if they found a way to make it work against all odds. What do you think about that?
Failure **Predictions** **Vanilla** $1(0.9993) \rightarrow 0(0.0198)$ **AdStyle** $1(0.9848) \rightarrow 0(0.0062)$ **Ours** $1(0.9965) \rightarrow 0(0.2434)$	**Original:** Acc to @WHO prolonged use of #facemasks may be uncomfortable but it does not lead to CO_2 intoxication or O_2 deficiency. Do ensure that your #facemask fits well and allows you to breathe normally. Do not re-use a #MedicalMask and always change your #mask as soon as it gets damp. **Rewrite:** According to the World Health Organization (WHO), while wearing a face mask for extended periods can be challenging, it does not result in carbon dioxide intoxication or oxygen deficiency. It is important to ensure that your face mask fits properly and allows for comfortable breathing. Additionally, it is advisable not to reuse a medical mask and to change your mask promptly if it becomes damp.

5 Analysis: From Vulnerabilities to Repair

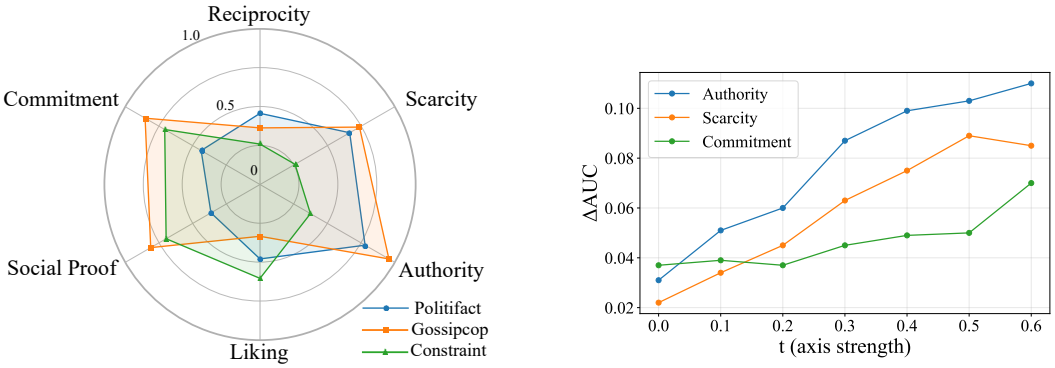
We analyze when and why detectors fail under rhetorically grounded attacks, and whether the resulting weaknesses admit *interpretable* repairs. We focus on three questions: (i) Do successful attacks concentrate along specific rhetorical axes? (ii) Does controlled manipulation of a single axis causally modulate robustness? (iii) Can axis-guided augmentation recover robustness? We complemented these analyses with human validation and qualitative case studies.

5.1 Rhetorical Axis Alignment Across Datasets

We first asked whether successful adversarial rewrites concentrate on identifiable rhetorical directions. For each successfully attacked test example, we record the optimized strategy $s^* \in \mathcal{S}$, measured relative to the neutral setting $s_0 = \mathbf{0}$, and define the displacement $\Delta s = s^* - s_0$. Let v_d denote the unit direction of axis d in $\mathcal{S}$. We measure alignment to axis d by

$$r_d = \cos(\Delta s, v_d), \tag{6}$$

where $\cos(\cdot, \cdot)$ is cosine similarity. We report $|r_d|$ to quantify the relative alignment of the strategy displacement with axis d, regardless of its sign(strengthening or suppression). Figure 4a visualizes axis-alignment profiles across three misinformation benchmarks, revealing non-uniform patterns in the optimized strategies. These profiles are descriptive, as they may reflect both successful attack directions and search-induced preferences. A complementary trajectory-matched null analysis on RoBERTa–PolitiFact identifies success-specific enrichment for selected axes, while showing that not all prominent directions exceed their null levels(Appendix K).



(a) Axis alignment.

(b) Single-axis sensitivity.

Figure 4: **Rhetorical vulnerability analysis.** (a) Successful attacks show non-uniform alignment across rhetorical axes. (b) Controlled single-axis interventions induce systematic AUC degradation.

5.2 Robustness Sensitivity Under Single-axis Interventions

Alignment is correlational; we next tested whether controlled changes along a single axis induced predictable robustness degradation. For each test input, we intervene on axis d by setting

$$s(t) = s_0 + tv_d, \quad t \in [0, 0.6], \tag{7}$$

with t increased in increments of 0.1. We restrict t to a moderate range to keep the interventions interpretable and avoid saturation effects. For each t, we generate a rewrite using the Planner-Rewriter and retain only outputs that pass the same semantic gate as in Section 3 ($\mathcal{G}(x, x') \geq \tau$). Under this setting, the gate accepts 92.09% of candidates on POLITIFACT. We define $\Delta\mathrm{AUC}(t)$ as the ROC-AUC on the original test set minus that under rewriting at strength t. Note that $t = 0$ still invokes the Planner-Rewriter, so a nonzero AUC drop can arise from neutral rewriting rather than amplification of the target axis. Figure 4b plots $\Delta\mathrm{AUC}(t)$ for three representative axes: *Authority*, *Scarcity*, and *Commitment/Consistency*. *Authority* and *Scarcity* show overall increasing trends in robustness degradation, reaching $\Delta\mathrm{AUC} = 0.110$ and 0.085 at $t = 0.6$, respectively, while *Commitment/Consistency* exhibits a weaker response (≈ 0.070). These interventional results provide evidence that detector robustness is systematically modulated by axis-conditioned rhetorical interventions, supporting $\mathcal{S}$ as a meaningful space for controlled diagnosis.

5.3 Human Validation of Rhetorical Interpretability

We evaluated whether the axis settings corresponded to rhetorically meaningful and perceptually grounded changes. We conducted a human evaluation on 300 (original, rewrite) pairs sampled from three misinformation datasets (50 pairs per axis across six axes). Annotators ($N = 20$) assessed (i) whether the rewrite strengthened the intended rhetorical axis relative to the original (*Perceptual Validity*) and (ii) how the rewrite altered *Perceived Believability*, both using a 3-point ordinal scale. Inter-annotator agreement is high for perceptual validity (κ_w=0.73, α_{ord}=0.76, ICC=0.90), indicating that axis-aligned rhetorical changes are consistently recognizable by humans. Agreement is moderate for believability (κ_w=0.61, α_{ord}=0.53, ICC=0.83), reflecting the inherently subjective nature of credibility judgments. Crucially, model-derived axis alignment scores r_d exhibited a strong positive correlation with human perceptual validity ratings (Spearman's ρ=0.62, computed per axis and averaged across six axes). This alignment supports that the directions optimized in $\mathcal{S}$ correspond to human-interpretable rhetorical mechanisms. Additional details of human evaluation are provided in Appendix F. An independent evaluation of realized rhetorical properties is reported in Appendix R.

5.4 Axis-augmented retraining for robustness repair

Next, we tested whether the vulnerable directions identified above could be used for targeted repair. Specifically, we took a baseline VANILLA detector on POLITIFACT and performed targeted data augmentation along the two most influential axes revealed by the sensitivity analysis, namely *Authority* and *Scarcity*, and their combined dimension (S+A). In each setting, we generated axis-conditioned,

Table 4: **Axis-augmented retraining results on POLITIFACT.** Targeted rhetorical-axis augmentation applied to the baseline VANILLA model substantially improved the AUC in two different scenarios. The most pronounced gains occurred along the *Authority* axis, suggesting that the identified vulnerability directions are actionable for targeted repair.

Retraining Setting	Epochs	AUC (Original)	AUC (Attack)
Baseline (Vanilla, PolitiFact)	3	0.8827	0.7081
Single-axis (*Scarcity*)	1	0.9355	0.7722
Single-axis (*Authority*)	1	0.9416	0.8853
Multi-axes (S+A)	1	0.9421	0.8992
Single-axis (*Scarcity*)	2	0.9419	0.8946
Single-axis (*Authority*)	2	0.9470	0.8921
Multi-axes (S+A)	2	0.9473	0.9106

gate-passing rewrites using the same planner-rewriter and semantic gate, mixed them into the original training data at an augmentation ratio of 0.25, and fine-tuned for 1-2 additional epochs.

Table 4 shows that targeted augmentation improves attacked-set ROC-AUC in the evaluated repair setting. Authority-based repair reaches 0.8853 after one epoch, while joint augmentation reaches 0.9106 after two epochs. These results provide evidence that axis-level vulnerability diagnosis can guide targeted repair. A focused pairwise interaction analysis is reported in Appendix L. A complementary RoBERTa-base study further examines this diagnosis-to-repair relationship (Appendix M).

5.5 Summary

Across datasets, rhetorical vulnerabilities are axis-dependent: successful attacks show non-uniform alignment patterns in $\mathcal{S}$, while controlled single-axis interventions reveal systematic robustness degradation, providing interventional evidence of rhetorical sensitivity. Human evaluation confirmed that these axis settings corresponded to perceptually meaningful rhetorical shifts. Finally, axis-guided augmentation repairs the discovered weaknesses, completing the loop from vulnerability discovery to repair within an axis-aligned strategy space. A complementary focused analysis for the extremist rhetoric setting is reported in Appendix Q.

6 Conclusion

We present ARAO, a unified framework that formulates adversarial rewriting as gradient-free optimization over an axis-aligned rhetorical strategy space grounded in rhetorical theory. Through a planner-rewriter pipeline with semantic gating and a dense adversarial objective, ARAO enables reliable black-box search and supports both vulnerability discovery and targeted repair. In training mode, this framework closes a discovery-to-repair loop by updating detectors on rhetorically controlled rewrites. Across misinformation and extremist rhetoric benchmarks, ARAO consistently improves robustness under a fixed, matched rhetorical attack protocol, improving over the strongest non-ARAO baselines by 2.35 ROC-AUC points on average on attacked sets while achieving the best performance on original content. Our analyses reveal structured, axis-dependent vulnerability patterns, demonstrate sensitivity to controlled rhetorical interventions, and show that axis-guided augmentation can mitigate identified weaknesses. Overall, ARAO bridges rhetorical theory and adversarial optimization, enabling interpretable, controllable, and repairable robustness evaluation for harmful content detectors. We discuss limitations, broader impacts, and considerations in Appendix A.

Acknowledgements

This work was supported in part by the College Students' Innovative Entrepreneurial Training Plan Program of China under Grant No. 202510304042 and the National Natural Science Foundation of China under Grant No. 62406153.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1316. URL https://aclanthology.org/D18-1316/.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2072. URL https://aclanthology.org/P15-2072/.

Robert B Cialdini. *Influence the psychology of persuasion*. Harpercollins publishers, 2007.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Teun A. Van Dijk. *Ideology: A Multidisciplinary Approach*. Sage Publications, 1998.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2006. URL https://aclanthology.org/P18-2006/.

Thomas Felber. Constraint 2021: Machine learning models for covid-19 fake news detection shared task, 2021. URL https://arxiv.org/abs/2101.03717.

Mayur Gaikwad, Swati Ahirrao, Shraddha Phansalkar, and Ketan Kotecha. Multi-ideology isis/jihadist white supremacist (miws) dataset for multi-class extremism text classification. *Data*, 6(11):117, 2021.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.464. URL https://aclanthology.org/2021.emnlp-main.464/.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095, 2023.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8018–8025, Apr. 2020. doi: 10.1609/aaai.v34i05.6311. URL https://ojs.aaai.org/index.php/AAAI/article/view/6311.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108, 2022.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. Contextualized perturbation for textual adversarial attack. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.400. URL https://aclanthology.org/2021.naacl-main.400/.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.500. URL https://aclanthology.org/2020.emnlp-main.500/.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.16. URL https://aclanthology.org/2020.emnlp-demos.16/.

Sungwon Park, Sungwon Han, Xing Xie, Jae-Gil Lee, and Meeyoung Cha. Adversarial style augmentation via large language model for robust fake news detection. In *Proceedings of the ACM on Web Conference 2025*, pages 4024–4033, 2025.

Martin Reisigl and Ruth Wodak. *Discourse and Discrimination: Rhetorics of Racism and Antisemitism*. London ; New York : Routledge, 2001.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *Big Data*, 8(3):171–188, 2020. doi: 10.1089/big.2020.0062.

Henri Tajfel, John Turner, William G Austin, and Stephen Worchel. An integrative theory of intergroup conflict. *Intergroup relations: Essential readings*, pages 94–109, 2001.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 613–624. International World Wide Web Conferences Steering Committee, April 2016. doi: 10.1145/2872427.2883081. URL http://dx.doi.org/10.1145/2872427.2883081.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *The Journal of Machine Learning Research*, 15(1):949–980, 2014.

Ruth Wodak. *The Politics of Fear: What Right-Wing Populist Discourses Mean*. Sage Publications, 2015.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3367–3378, August 2024. URL http://dx.doi.org/10.1145/3637528.3671977.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1BLjgZCb.

Table 5: Definitions of the rhetorical axes used in the misinformation detection task.

Axis / Dimension	Linguistic Definition & Planner Instruction
Authority	*Definition:* Reference to credible institutions, experts, or official titles to boost legitimacy. *Instruction:* "Inject formal citations or mention expert consensus."
Scarcity	*Definition:* Emphasizing limited availability, time constraints, or exclusive access. *Instruction:* "Add temporal urgency words (e.g., 'limited time', 'urgent warning')."
Commitment	*Definition:* Aligning with the reader's prior beliefs or public stances to enforce consistency. *Instruction:* "Frame the argument as a logical continuation of accepted values."
Social Proof	*Definition:* Citing majority opinion, popularity, or collective behavior. *Instruction:* "Highlight widespread acceptance or 'viral' status of the claim."
Reciprocity	*Definition:* Creating a sense of obligation or mutual benefit. *Instruction:* "Frame the information as a special benefit provided to the reader."
Liking	*Definition:* Using relatable, friendly, or attractive personas to build rapport. *Instruction:* "Soften the tone to be more conversational and relatable."

A Limitations and Broader Impact

The current axis design is a theory-grounded instantiation rather than a uniquely determined taxonomy of harmful rewriting mechanisms. Our work is best understood as an initial step toward operationalizing rhetorical mechanisms as a control space for adversarial rewriting, and how to define, compare, and adapt alternative axis sets across domains remains an open question.

The primary robustness evaluation uses a matched ARAO threat model, which may favor training aligned with the same rhetorical strategy space. External BERT-Attack and AdStyle evaluations provide complementary cross-threat evidence, while RoBERTa-base results offer initial cross-backbone validation. However, these experiments do not establish universal robustness. Moreover, the NLI-based non-contradiction gate cannot guarantee factual or label preservation, particularly when rewrites introduce additional claims or alter harm intensity.

Although we evaluate two harmful-content domains and provide supplementary checks with alternative rewriters, decoding temperatures, and an external attack protocol, our experiments do not establish full portability across detector architectures, generator families, languages, or unseen domains. The main experiments use a shared BERT-based backbone, and human validation focuses primarily on misinformation rewrites, with extremist-rhetoric axes examined through focused quantitative analyses. Future work should extend the framework to richer axis specifications, broader detector and generator families, multilingual settings, and domain-specific human validation.

The broader societal consequences of this work are two-fold: it contributes to a safer digital ecosystem by strengthening defenses against misinformation and extremist rhetoric, and it promotes algorithmic transparency by providing human-recognizable diagnostics for why content moderation systems fail. While adversarial methodologies carry an inherent risk of dual use, potentially informing more deceptive content generation, ARAO is designed for defensive auditing and systematic model hardening rather than unstructured black-box attack deployment. We therefore frame the method as a tool for controlled robustness evaluation, vulnerability diagnosis, and targeted repair of harmful-content detectors.

B Rhetorical Axes Specifications

To ensure reproducibility, we provide the operational definitions and planner instructions for the rhetorical axes used to instantiate $\mathcal{S}$. These axes are *not* uniquely correct; they constitute one theory-grounded, low-dimensional instantiation that is interpretable and linguistically realizable. In principle, ARAO is not tied to a specific taxonomy and can incorporate alternative domain-appropriate axes, provided that they admit controllable realization by the Planner-Rewriter under the same semantic constraint. Table 5 lists the k=6 Cialdini-inspired persuasion axes used for misinformation detection, and Table 6 lists the k=4 ideology-framing axes used for extremist rhetoric detection.

Table 6: Definitions of the rhetorical axes used in the extremist rhetoric detection task, derived from ideological discourse theory.

Axis / Dimension	Linguistic Definition & Planner Instruction
Polarization	*Definition:* Constructing a sharp "Us vs. Them" dichotomy to highlight group conflict and incompatibility. *Instruction:* "Amplify the contrast between the in-group's virtues and the out-group's vices."
In-group Framing	*Definition:* Strengthening collective identity through shared victimhood, destiny, or superiority cues. *Instruction:* "Use inclusive language (e.g., 'our brethren', 'true believers') to foster solidarity."
Call-to-Action Intensity	*Definition:* Urging immediate mobilization, physical action, or retribution against threats. *Instruction:* "Inject imperative verbs demanding immediate defense or decisive action."
Moral Justification	*Definition:* Legitimizing radical views or actions as virtuous, divine, or ethically necessary. *Instruction:* "Frame the narrative as a righteous struggle for justice or higher truth."

C Prompt Specifications

We provide the full prompt templates for the planner and rewriter.

Rewriter Prompt

```
You are a professional English text rewriting expert. You will
rewrite the given text according to strategic guidance, while
preserving the core claim and not introducing contradictory details.
Please rewrite the text according to the following strategic
guidance:
{guideline}

Original Text:
{original_text}

Rewriting Requirements:
1. Strictly follow the above strategic guidance for rewriting.
2. Preserve the core claim and factual content of the original text;
do not add contradictions.
3. Keep key facts consistent unless the guideline explicitly
requests harmless elaboration.
4. Ensure the rewrite is clear, coherent, and fluent in standard
English.
5. Output only the complete rewritten text without additional
explanations.

Rewritten Text:
Please directly output the complete rewritten English text:
```

Planner Prompt

```
You are a rhetorical planning expert. Based on the provided
rhetorical strategy configuration, generate a clear and concise
guideline instructing the rewriter how to adjust the text.

The strategy vector is specified as follows:
{axis_values}

Please summarize these settings into a coherent set of rewriting
guidelines that explicitly describe how rhetorical emphasis should
shift (e.g., strengthen authority, highlight urgency/scarcity,
```

```
reduce emotional tone).
Ensure the guideline is actionable, stylistically neutral, and
suitable for conditioning a rewriting model.  Output only the
guideline.
```

D Reproducibility Details

The core implementation details are reported in the main text. This appendix provides the remaining components necessary for the full replication of all experiments and comprehensive details regarding our computing infrastructure.

Computing Infrastructure. All experiments were conducted on a single NVIDIA RTX 4090 GPU equipped with an Intel Core i9 processor. The implementation leverages PyTorch 2.6.0 and HuggingFace Transformers 4.57.1.

Dataset Preprocessing. We released scripts for dataset normalization, text filtering, label extraction, and the construction of train/dev/test splits. We used the standard dataset splits when they were provided; otherwise, we created a 7:1.5:1.5 train/dev/test split for the remaining datasets.

E Pilot Study for Hyperparameter Selection (PolitiFact)

We conducted a pilot study on POLITIFACT to justify the key hyperparameters governing (i) semantic fidelity via the gate threshold τ and (ii) search dynamics via the NES query budget. Unless otherwise stated, we used the same detector backbone, training protocol, planner-rewriter prompts, and ARAO attack procedure as in the main experiments. For computational efficiency, all experiments in this section were conducted on a representative subset comprising 12.5% of the original dataset, randomly sampled to preserve the class distribution.

Gate threshold τ: fidelity-coverage trade-off. We first calibrated the non-contradiction gate threshold τ (based on the NLI non-contradiction score, denoted NC) because it directly controls the semantic fidelity of rewrites and the effective candidate pool available to the controller. We swept $\tau \in \{0.00, 0.50, 0.60, 0.70, 0.80, 0.85, 0.90, 0.95\}$ while holding the search and training hyperparameters fixed. For each τ, we reported (i) the gate pass rate (coverage) over generated candidates, (ii) the number of gate-passing samples retained, and (iii) ROC-AUC on the resulting *ARAO-attacked dev set* constructed under the same budget and gate (Table 7).

As τ increased, the pass rate decreased monotonically from 100% at τ=0.00 to 88.33% at τ=0.95, reflecting stricter semantic filtering. Meanwhile, attacked-dev ROC-AUC improved substantially when moving from looser gates to moderate strictness (e.g., 0.9540 at τ=0.70 to 0.9634 at τ=0.80), and then saturated in the range $\tau \in [0.80, 0.95]$ (AUC ≈ 0.963). Therefore, we selected τ=0.85 as a balanced operating point in the saturation regime: it preserved high coverage (93.75%) and a large candidate pool (225 retained samples) while enforcing stronger semantic fidelity, with no measurable loss in attacked-dev ROC-AUC relative to stricter thresholds (e.g., τ=0.90).

Query budget and NES stability. We characterized the trade-off between the *attack strength* under a fixed ARAO threat model and *compute cost* to select a practical default NES budget. We varied the NES population size N and the number of iterations T (query budget $B = N \times T$) while holding all other components fixed, including the gate threshold $\tau = 0.85$. For each configuration, we reported the ROC-AUC on the resulting ARAO-attacked split (mean$\pm$std over seeds), gate pass rate, and total token usage.

Token usage counts *all* LLM tokens consumed by the ARAO pipeline, including both planner and rewriter calls, and including candidates that fail the gate. This end-to-end accounting reflects the true computational footprint of the attack protocol.

Table 8 shows that increasing B generally strengthens the adversary (lower attacked-set AUC) but increases the cost roughly linearly in total tokens. Importantly, moderate budgets exhibit diminishing returns: $B = 64$ and $B = 128$ achieve similar attacked-set AUC within uncertainty, but $B = 128$ attains a higher gate-pass rate. A higher pass rate increases the effective number of valid samples

Table 7: **PolitiFact dev pilot for selecting the gate threshold τ (NC).** We reported the gate-pass rate, the number of retained gate-passing samples, and ROC-AUC on the ARAO-attacked dev set, constructed under the same budget and gate.

τ (NC)	Pass Rate	Retained	ROC-AUC
0.00	100.00%	240	0.9459
0.50	97.08%	233	0.9552
0.60	96.67%	232	0.9548
0.70	95.83%	230	0.9540
0.80	94.17%	226	0.9634
0.85	**93.75%**	**225**	**0.9632**
0.90	92.92%	223	0.9634
0.95	88.33%	212	0.9628

Table 8: **Budget-cost trade-off for NES.** Lower AUC indicates a stronger attack under the fixed ARAO threat model.

N	T	$B{=}N\times T$	AUC (Attacked)	Pass Rate (%)	Tokens
8	4	32	0.9000 ± 0.0163	95.00	251,396
8	8	64	0.8967 ± 0.0125	96.67	490,193
16	8	128	0.8967 ± 0.0189	98.33	993,447
16	16	256	0.8767 ± 0.0205	95.00	1,988,239

per iteration and reduces degenerate updates when few candidates survive filtering, improving the stability of NES updates. Therefore, we adopt $N = 16$ and $T = 8$ (budget $B = 128$) as the default setting throughout the paper, balancing attack strength, update stability, and practical computation.

F Human Evaluation Details

In Section 5.3, we randomly sampled 300 (original, rewrite) pairs from the misinformation detection datasets (50 pairs per rhetorical axis across six axes) and conducted a human evaluation. Twenty annotators with graduate-level English proficiency were presented with pairs of (Original Text, Rewrite) together with the target axis definition (e.g., Authority). They answered two questions on a 3-point Likert scale:

- **Perceptual Validity (Axis Strengthening):** "To what extent does the rewrite strengthen the specified rhetorical axis compared to the original?" (1: No change, 2: Weak change, 3: Strong change).
- **Believability:** "Compared to the original, are you more willing to believe the rewrite?" (1: Less willing to believe, 2: Similar, 3: More willing to believe).

F.1 Statistical Metrics and Formulations

To ensure the robustness of our findings, we utilized four complementary statistical measures to quantify inter-rater reliability and model-human alignment.

1. Weighted Cohen's Kappa (κ_w) We utilize the weighted version of Cohen's Kappa to evaluate inter-rater agreement while accounting for the ordinal nature of our Likert scale. We employ **Quadratic Weights** to penalize larger discrepancies more severely (e.g., a disagreement between 1 and 3 is penalized four times as much as between 1 and 2). The coefficient is defined as:

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} p_{ij}}{\sum_{i,j} w_{ij} e_{ij}} \tag{8}$$

where p_{ij} is the proportion of observed agreement, e_{ij} is the proportion of agreement expected by chance, and the weight w_{ij} for categories i and j is given by:

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{9}$$

Table 9: Notation summary.

Symbol	Definition
$\mathcal{S}$	Rhetorical strategy space $\subset \mathbb{R}^k$
$s \in \mathcal{S}$	Strategy vector
$\mathcal{P}$	Planner: maps $s \in \mathcal{S}$ to a guideline r
$\mathcal{W}$	Rewriter: rewrites x conditioned on guideline r
r	guideline (rewriting directives)
$\mathcal{G}$	Semantic gate
$\mathcal{M}$	Adversarial objective function
f	Target detector
θ	Detector parameters
μ, Σ	NES distribution parameters
$\mathcal{X}$	Input text space
$\mathcal{Y}$	Label space

where N is the total number of ordinal categories ($N = 3$ in our study).

2. Krippendorff's Alpha (α) Krippendorff's Alpha is a general-purpose reliability coefficient applicable to any number of raters and capable of handling missing data. It is formulated as:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{10}$$

where D_o is the observed disagreement and D_e is the disagreement expected by chance. For ordinal data, the disagreement function $\delta^2(c, k)$ between two ranks c and k is defined as:

$$\delta^2(c, k) = \left(\sum_{g=c}^{k} n_g - \frac{n_c + n_k}{2} \right)^2 \tag{11}$$

where n denotes the frequency counts of the categories.

3. Intraclass Correlation Coefficient (ICC(2,k)) To assess the reliability of the mean ratings across k annotators, we employ the ICC(2,k) model (two-way random effects, absolute agreement, average measures). This model accounts for systematic biases between raters and is calculated via a Two-way ANOVA:

$$ICC(2, k) = \frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}} \tag{12}$$

where MS_R is the mean square for rows (subjects/samples), MS_E is the mean square for error, MS_C is the mean square for columns (raters), and n is the number of samples.

4. Spearman's Rank Correlation (ρ) Finally, to validate the interpretability of our optimized coordinates, we calculate Spearman's ρ between the model-derived alignment scores r_d and human perceptual validity ratings. This non-parametric measure assesses the monotonic relationship between mathematical optimization and human intuition:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{13}$$

where d_i is the difference between the ranks of the model score and the human rating for sample i, and n is the total number of evaluation pairs.

G Human Audit of Label Preservation and Factual Additions

To assess the reliability of the NLI-based semantic gate beyond non-contradiction, we conducted an additional human audit of 100 randomly sampled gate-passing rewrites from the final attacked test sets across all six datasets.

Table 10: Hyperparameter settings used in all experiments.

Parameter	Symbol	Value
Detector learning rate	α	2×10^{-5}
Strategy update rate	η	0.03
Covariance update rate	β	0.02
Gate threshold	τ	0.85
Population size	N	16
Max iterations	T	8

Five annotators independently evaluated the same 100 original–rewrite pairs along three criteria: (i) whether the original label remained valid, (ii) whether the rewrite introduced new factual claims or evidence, and (iii) whether such additions materially affected the basis for label assessment. Annotators could abstain when a judgment was uncertain. For each criterion, we excluded missing judgments and computed the proportion of affirmative responses among all valid annotation judgments.

Among valid annotation judgments, 95.2% indicated label preservation, 5.6% identified new factual claims or evidence, and 4.4% identified additions that materially affected the basis for label assessment. These results suggest that label preservation is common among the sampled gate-passing rewrites, while highlighting a non-negligible risk of label-relevant semantic changes.

H Computational Cost and Budget Trade-offs

LLM-based adversarial frameworks incur substantial generation costs. We therefore examine the token consumption of ARAO under different search budgets and diagnostic subset sizes, using ADSTYLE Park et al. [2025] as a reference.

ARAO allows the computational budget to be adjusted through the NES population size N, iteration count T, and the fraction of examples used for diagnosis. Table 11 reports estimated token usage on POLITIFACT. Here, $B = N \times T$ denotes the number of candidate strategy evaluations per example, rather than the total number of individual LLM API calls. The estimates account for both Planner and Rewriter token consumption. Since the configurations differ in evaluated data coverage and search budget, relative costs should not be interpreted as efficiency gains under matched evaluation conditions.

Table 11: **Computational cost under different diagnostic configurations.** Estimated token usage based on average prompt and response lengths using GPT-4o-mini. Relative costs are normalized by the reported standard ADSTYLE configuration; data coverage and search budgets are not matched across all settings.

Method	Config (N, T)	Budget (B)	Total Tokens	vs. ADSTYLE
ADSTYLE	Standard	–	∼1,324,800	1.00×
ARAO (30% sub.)	(4, 2)	8	∼710,400	0.54×
ARAO (30% sub.)	(4, 4)	16	∼1,420,800	1.07×
ARAO (30% sub.)	(8, 4)	32	∼2,841,600	2.14×
ARAO (10% sub.)	**(4, 2)**	**8**	**∼236,800**	**0.18×**
ARAO (10% sub.)	(4, 4)	16	∼473,600	0.36×
ARAO (10% sub.)	(8, 4)	32	∼947,200	0.71×
ARAO (10% sub.)	(8, 8)	64	∼1,894,400	1.43×

Subsampled Diagnosis. ARAO permits exploratory vulnerability diagnosis on a subset of the data. For example, the 10% subsample with $B = 8$ consumes approximately 236,800 tokens, or 18% of the reported ADSTYLE reference cost. This demonstrates the availability of a lower-cost diagnostic configuration.

Tunable Exploration Depth. Increasing the search budget allows more extensive exploration of the rhetorical strategy space at higher token cost. As shown in Table 11, token consumption increases

Algorithm 1 Adaptive Rhetorical Adversarial Optimization (ARAO)

Require: Input (x, y), detector f_θ, strategy space $\mathcal{S} = [-1, 1]^k$, initial distribution (μ_1, Σ_1), population size N, iterations T, stepsizes η, β, α, threshold τ, flag `train`

for $t = 1$ to T **do**
  Sample $\{s_i\}_{i=1}^N \sim \mathcal{N}(\mu_t, \Sigma_t)$ and clip s_i to $\mathcal{S}$
  $\mathcal{Z} \leftarrow \emptyset$; $\mathcal{V} \leftarrow \emptyset$
  for $i = 1, \ldots, N$ **do**
    $x_i' \leftarrow \mathcal{W}(x, \mathcal{P}(s_i))$
    if $\mathcal{G}(x, x_i') \geq \tau$ **then**
      $m_i \leftarrow -\log f_\theta(x_i')[y]$
      $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{(s_i, m_i)\}$
      $\mathcal{V} \leftarrow \mathcal{V} \cup \{x_i'\}$
    end if
  end for
  if $|\mathcal{Z}| > 0$ **then**
    $\{w_i\}_{(s_i,m_i)\in\mathcal{Z}} \leftarrow \textsc{RankNorm}(\{m_i : (s_i, m_i) \in \mathcal{Z}\})$
    $\mu_{t+1} \leftarrow \mu_t + \eta \sum_{(s_i,m_i)\in\mathcal{Z}} w_i(s_i - \mu_t)$
    $\Sigma_{t+1} \leftarrow \Sigma_t + \beta \sum_{(s_i,m_i)\in\mathcal{Z}} w_i\big((s_i - \mu_t)(s_i - \mu_t)^\top - \Sigma_t\big)$
  else
    $\mu_{t+1} \leftarrow \mu_t$; $\Sigma_{t+1} \leftarrow \Sigma_t$
  end if
  if `train` and $|\mathcal{V}| > 0$ **then**
    $\theta \leftarrow \theta - \alpha \nabla_\theta \frac{1}{|\mathcal{V}|} \sum_{x'\in\mathcal{V}} \ell(f_\theta(x'), y)$
  end if
end for
return $(\mu_{T+1}, \Sigma_{T+1})$

approximately linearly with B for a fixed subset size. Our separate budget pilot (Appendix E) further shows similar attacked-set AUC at $B = 64$ and $B = 128$ within experimental uncertainty, illustrating the trade-off between attack strength and computational expenditure.

Cost of Diagnosis-Guided Repair. After vulnerable axes have been identified, targeted repair can generate a fixed set of axis-conditioned training rewrites without repeating the full NES search during fine-tuning.

I Formal Algorithm Details

In this section, we provide the formal procedural specification for the Adaptive Rhetorical Adversarial Optimization (ARAO) framework. Algorithm 1 integrates the hierarchical text generation pipeline with a gradient-free search mechanism to realize the proposed discovery-to-repair loop.

Algorithmic Flow. The procedure begins by initializing a search distribution $\mathcal{N}(\mu, \Sigma)$ over the rhetorical strategy space $\mathcal{S} = [-1, 1]^k$. In each iteration t, the controller samples a population of N strategy vectors $\{s_i\}$. These vectors are translated into natural language guidelines by the Planner $\mathcal{P}$ and subsequently realized as candidate rewrites $\{x_i'\}$ by the Rewriter $\mathcal{W}$. To ensure semantic fidelity, the semantic gate $\mathcal{G}$ filters out candidates that contradict the original claim x.

Gradient-free Optimization. For candidates that pass the gate, we compute the adversarial objective m_i based on the detector's confidence. We used rank-based normalization (RANKNORM) on these scores to compute utilities $\{w_i\}$, thereby mitigating the impact of noisy outliers in the black-box evaluation. The parameters of the search distribution (μ, Σ) are then updated using Natural Evolution Strategies (NES) to move the sampling density toward high-vulnerability regions of the strategy space.

Integrated Repair Loop. A key feature of ARAO is the optional online repair mechanism (controlled by the `train` flag). If enabled, the framework performs an immediate gradient descent update on the detector parameters θ using the successful, gate-passing adversarial samples $\mathcal{V}$. This allows the model to adaptively harden itself against the specific rhetorical vulnerabilities discovered during the optimization process, effectively closing the loop from discovery to repair.

Table 12: **Statistical significance of ARAO vs. each baseline.** Bonferroni-corrected p-values from paired t-tests under the *Original* and *Attack* settings. For each setting, the test is performed on the seed-level ROC-AUC values aggregated across all datasets in that setting.

Comparison	**Original (corrected p)**	**Attack (corrected p)**
Vanilla	2.906403×10^{-3}	2.229521×10^{-5}
UDA	5.745111×10^{-4}	7.360860×10^{-6}
RADAR	1.783942×10^{-2}	6.036195×10^{-6}
SheepDog	1.484122×10^{-2}	7.579425×10^{-5}
AdStyle	4.907126×10^{-2}	1.313543×10^{-4}

Table 13: **Null-corrected axis alignment on RoBERTa–PolitiFact.** Positive differences indicate enrichment among successful attacks relative to the trajectory-matched null.

Axis	Successful MAA	Null MAA	Difference
Authority	0.3191	0.3492	-0.0301
Scarcity	0.3460	0.3253	$+0.0207$
Commitment/Consistency	0.3281	0.3511	-0.0230
Social Proof	0.3444	0.3479	-0.0035
Reciprocity	**0.3979**	0.3465	$\mathbf{+0.0514}$
Liking	0.2928	0.3197	-0.0269

J Statistical Significance of Main Results

To complement the multi-seed results in Table 1, we further assessed whether the improvements of ARAO over the baselines are statistically reliable under the two evaluation settings. Using the seed-level ROC-AUC results, we conducted pairwise *paired* t-tests between ARAO and each baseline separately for the *Original* and *Attack* settings. For each setting, we aggregated the paired observations across all datasets in that setting.

Table 12 reports the Bonferroni-corrected p-values. As shown in Table 12, all corrected p-values remain below 0.05 in both settings. The comparison against the strongest baseline, AdStyle, is significant in both cases, with a markedly smaller corrected p-value under the *Attack* setting.

K Null-Corrected Rhetorical Axis Alignment

To distinguish success-specific axis enrichment from search-induced patterns, we constructed a trajectory-matched null using gate-passing candidates from the same adversarial search process.

For each axis d, we report mean absolute alignment (MAA):

$$\mathrm{MAA}_d = \mathbb{E}_{\Delta s}\left[\left|\cos(\Delta s, v_d)\right|\right]. \tag{14}$$

Unlike signed cosine similarity, MAA measures axis involvement regardless of direction.

The RoBERTa–PolitiFact analysis includes 52 strict successful attacks and 1,272 candidates from their corresponding search trajectories. Null MAA values were averaged over 50 uniform random samplings of trajectory-matched gate-passing candidates.

Reciprocity exhibits the largest positive enrichment ($+0.0514$), followed by Scarcity ($+0.0207$). Authority and the remaining axes do not exceed, or remain close to, their null levels. Thus, raw alignment profiles should be interpreted descriptively rather than as unique mechanism fingerprints.

Across the 52 successful strategy displacements, the mean and maximum absolute off-diagonal Pearson correlations were 0.1174 and 0.2437, respectively; the corresponding covariance statistics were 0.00121 and 0.00276. These results provide no evidence of strong linear collinearity in this sample, although orthogonality in the parameterized strategy space does not imply independence of realized rhetorical effects.

Table 14: **Focused pairwise repair study on POLITIFACT.** We report original-set and attacked-set ROC-AUC for two representative axis pairs after 1 and 2 additional fine-tuning epochs under the same matched attack protocol as in the main text.

Axis pair	Epoch 1 Orig.	Epoch 1 Attack	Epoch 2 Orig.	Epoch 2 Attack
Authority + Commitment	0.9386	0.8700	0.9502	0.9282
Scarcity + Commitment	0.9416	0.8383	0.9465	0.8712

Table 15: **Diagnosis-guided targeted repair on RoBERTa-base (POLITIFACT).** Clean and attacked accuracy (%), ROC-AUC, and attack success rate (ASR, %). "ep." denotes additional fine-tuning epochs. Best results are highlighted in bold.

Repair Setting	Clean Acc.	Clean AUC	Attack Acc.	Attack AUC	ASR ↓
Vanilla	85.53	0.9274	74.84	0.8352	20.59
Authority (1 ep.)	86.16	0.9385	77.99	0.8481	17.52
Reciprocity (1 ep.)	83.65	0.8756	81.76	0.8629	10.53
Authority + Reciprocity (1 ep.)	83.65	0.9378	**85.53**	0.8774	7.52
Authority (2 ep.)	85.53	0.9422	77.99	0.8331	14.71
Reciprocity (2 ep.)	86.79	0.9250	83.02	0.8862	10.14
Authority + Reciprocity (2 ep.)	84.28	0.9349	82.39	0.8698	8.96
Full ARAO	**87.42**	**0.9507**	84.91	**0.9273**	**6.47**

L Pairwise Axis Interactions and Multi-axis Repair

The main text establishes axis-level sensitivity through controlled single-axis interventions and shows that targeted augmentation along vulnerable axes can improve robustness. To further examine whether axis effects combine additively in practice, we conducted a focused pairwise repair study on POLITIFACT using the same detector backbone, planner-rewriter pipeline, semantic gate, and matched attack protocol as in the main experiments.

We considered two representative axis pairs involving *Commitment/Consistency*: *Authority + Commitment* and *Scarcity + Commitment*. For each pair, we generated gate-passing rewrites under the same semantic constraint as in the main text, mixed them into the training data using the same augmentation setting, and fine-tuned the detector for additional epochs. We then evaluated both original-set and attacked-set ROC-AUC under the same evaluation protocol.

Table 14 shows that repair depends on the combination. In particular, *Authority + Commitment* exhibits substantially stronger complementarity than *Scarcity + Commitment*, especially on the attacked sets: after 2 epochs, the attacked set ROC-AUC reaches 0.9282 for *Authority + Commitment*, compared with 0.8712 for *Scarcity + Commitment*. This result suggests that some axis combinations mitigate multiple vulnerabilities more effectively than others, and that selected multi-axis augmentation can provide stronger repair than considering each axis in isolation. More broadly, the findings indicate that the proposed strategy space supports practical pairwise interaction analysis in addition to single-axis diagnosis and repair.

M Diagnosis-Guided Targeted Repair on RoBERTa-base

To further examine whether axis-level vulnerability diagnosis can guide targeted repair, we conducted a contrastive study on POLITIFACT using RoBERTa-base.

Our trajectory-matched null analysis identified *Reciprocity* as the most enriched axis among successful attacks ($+0.0514$), whereas *Authority*, previously prominent in the BERT-based analysis, did not exceed its gate-passing null (-0.0301). We therefore selected *Reciprocity* as the diagnosis-supported axis and *Authority* as a non-enriched comparison axis. Following the axis-guided augmentation procedure in Section 5.4, we performed targeted retraining along each axis separately and jointly for one or two additional epochs.

As shown in Table 15, *Reciprocity* consistently outperforms *Authority* in attacked ROC-AUC and ASR at both training lengths. After two epochs, Reciprocity-based repair improves attacked ROC-AUC

Table 16: **Cross-threat evaluation under ADSTYLE attacks.** We report ROC-AUC (mean ± SD over 3 seeds) on two representative datasets under the external ADSTYLE attack protocol.

Method	PolitiFact	GHC
Vanilla	0.8317 ± 0.0111	0.6907 ± 0.0085
UDA	0.8316 ± 0.0231	0.7246 ± 0.0018
RADAR	0.8644 ± 0.0136	0.7445 ± 0.0035
SheepDog	0.8645 ± 0.0080	0.7413 ± 0.0038
AdStyle	0.9133 ± 0.0174	0.7454 ± 0.0110
Ours	$\mathbf{0.9237 \pm 0.0052}$	$\mathbf{0.7511 \pm 0.0016}$

from 0.8352 to 0.8862 (+5.10 points) and reduces ASR from 20.59% to 10.14% (-10.45 percentage points), while largely preserving clean ROC-AUC (0.9250 vs. 0.9274).

At one epoch, joint augmentation along *Authority* and *Reciprocity* yields higher attacked ROC-AUC (0.8774) and lower ASR (7.52%) than either single-axis variant, while also improving clean ROC-AUC over Vanilla. Full ARAO achieves the highest clean ROC-AUC, attacked ROC-AUC, and lowest ASR among the evaluated settings.

These results provide focused intervention-level evidence that null-corrected axis diagnosis can inform repair selection: the diagnosis-supported *Reciprocity* axis yields more effective targeted repair than the non-enriched *Authority* axis in this setting. However, the differing clean-set outcomes across repair configurations also indicate that targeted augmentation can involve performance trade-offs.

N Cross-threat Evaluation Under External Style Attacks

The main text evaluates robustness under a matched rhetorical threat model, which directly tests the paper's core question: whether robustness training guided by vulnerabilities discovered in the ARAO space improves detector robustness against structured rhetorical attacks. To examine transfer beyond this matched setting, we additionally evaluated all methods under ADSTYLE attacks over 3 random seeds on two representative datasets, POLITIFACT and GHC.

Table 16 reports the results. ARAO remains the strongest method on both datasets, achieving 0.9237 ± 0.0052 on POLITIFACT and 0.7511 ± 0.0016 on GHC, compared with 0.9133 ± 0.0174 and 0.7454 ± 0.0110 for the strongest non-ARAO baseline, ADSTYLE, respectively. These results indicate that the robustness gains are not limited to the matched ARAO threat model, but transfer to at least one external rewriting-based attack family.

O Cross-Backbone Evaluation with RoBERTa-base

To examine whether the robustness improvements extend beyond the BERT backbone used in the main experiments, we conducted additional evaluations with RoBERTa-base on POLITIFACT and GHC, representing misinformation and extremist-rhetoric detection, respectively. We reran all six methods following the same training and evaluation protocols as in the main experiments. Each detector was evaluated on the original test set and under the matched ARAO attack protocol, using the same query budget and semantic gate. We report ROC-AUC under both scenarios and attack success rate (ASR; lower is better).

As shown in Table 17, the Vanilla RoBERTa detector remains vulnerable to rhetorical rewriting, with ROC-AUC decreasing from 0.9274 to 0.8352 on POLITIFACT and from 0.8771 to 0.8137 on GHC. In contrast, ARAO achieves the highest attacked ROC-AUC and lowest ASR on both datasets. Compared with the strongest non-ARAO baselines, ARAO improves attacked ROC-AUC by 3.00 and 1.77 points, while reducing ASR by 3.75 and 0.88 percentage points on POLITIFACT and GHC, respectively.

On the original test sets, ARAO remains competitive, achieving the second-highest ROC-AUC on POLITIFACT and tying for the highest ROC-AUC on GHC. These results provide additional evidence that the robustness benefits of ARAO are not specific to BERT. Nevertheless, this evaluation remains within two datasets; broader cross-architecture generalization warrants further investigation.

Table 17: **Cross-backbone robustness evaluation with RoBERTa-base.** Original and attacked ROC-AUC and attack success rate (ASR, %) on POLITIFACT and GHC. The best results are highlighted in bold.

	PolitiFact			GHC		
Method	Orig. AUC	Attack AUC	ASR ↓	Orig. AUC	Attack AUC	ASR ↓
Vanilla	0.9274	0.8352	20.59	0.8771	0.8137	15.23
UDA	0.9228	0.8419	15.44	0.8354	0.7918	11.53
RADAR	**0.9565**	0.8897	10.22	0.8745	0.7904	14.99
SheepDog	0.9370	0.8918	12.77	**0.8830**	0.8256	13.53
AdStyle	0.9366	0.8973	14.08	0.8767	0.8139	15.88
Ours	0.9507	**0.9273**	**6.47**	**0.8830**	**0.8433**	**10.65**

Table 18: **Robustness to rewriter choice and decoding temperature on POLITIFACT.** We report attacked-set ROC-AUC under the same matched evaluation protocol while varying only the rewriter model and temperature.

Rewriter	$T = 0.3$	$T = 0.6$	$T = 0.9$
Gemini 2.5 Flash	0.9499	0.9470	0.9504
Claude Sonnet 4.6	0.9409	0.9460	0.9457

P Robustness to Rewriter Choice and Decoding Temperature

The main experiments use a fixed rewriter to isolate the contribution of strategy-space optimization and keep the attack protocol controlled across methods. To examine whether the observed gains depend on a particular rewriting model or decoding temperature, we conducted an additional robustness check on POLITIFACT using two alternative rewriters, GEMINI 2.5 FLASH and CLAUDE SONNET 4.6, each evaluated under three temperature settings ($T \in \{0.3, 0.6, 0.9\}$).

For each setting, we kept the detector, strategy space, planner-rewriter pipeline, semantic gate, and attack protocol unchanged, varying only the rewriter model and decoding temperature. We then constructed attacked sets under the same matched evaluation protocol as in the main text and measured attacked-set ROC-AUC.

Table 18 shows that ARAO remains consistently strong across all six settings. Under GEMINI 2.5 FLASH, attacked-set ROC-AUC reaches 0.9499, 0.9470, and 0.9504 for $T = 0.3, 0.6$, and 0.9, respectively. Under CLAUDE SONNET 4.6, the corresponding values are 0.9409, 0.9460, and 0.9457. These results indicate that the robustness gains are not tied to a single rewriter or a specific decoding temperature, and that the proposed framework remains effective under alternative rewriting backends.

Q Focused Analysis of Extremist-Rhetoric Axes

The main text emphasizes detailed diagnosis-to-repair analysis on the misinformation task. To provide a complementary analysis for the extremist-rhetoric setting, we add a focused case study on MIWS-WS using *Call-To-Action Intensity* as a representative extremist-rhetoric axis. This analysis covers both single-axis sensitivity and axis-guided repair under the same semantic gate and matched evaluation protocol as in the main text.

Single-axis interpolation on MIWS-WS. We examine whether controlled interpolation along *Call-to-Action Intensity* systematically affects detector robustness. We increase the axis strength t from 0.0 to 0.6 and measure the resulting degradation in ROC-AUC on MIWS-WS. Table 19 reports the corresponding ΔAUC values. Notably, $t = 0$ still produces a neutral rewrite, which can affect detector predictions even without strengthening the target axis. Despite local fluctuations, the overall degradation increases from 0.0401 at $t = 0.0$ to 0.1361 at $t = 0.6$. These findings provide complementary evidence of axis-conditioned robustness sensitivity in the extremist-rhetoric setting.

Axis-guided repair on MIWS-WS. We next test whether the discovered vulnerability can be mitigated through targeted augmentation along the same axis. Using *Axis-Aug* (*Call-To-Action*

Table 19: **Single-axis interpolation on MIWS-WS using *Call-To-Action Intensity*.** We report the degradation in ROC-AUC (ΔAUC) as the interpolation strength t increases.

t	0.0	0.1	0.2	0.3	0.4	0.5	0.6
ΔAUC	0.0401	0.0836	0.0699	0.0762	0.0697	0.1199	0.1361

Table 20: **Axis-guided repair on MIWS-WS using *Call-To-Action Intensity*.** We compare the vanilla detector and axis-guided augmentation under the same clean-set and attacked-set evaluation protocols.

Method	**ROC-AUC (Clean)**	**ROC-AUC (Attack)**
Vanilla	0.6574	0.6294
Axis-Aug (*Call-To-Action Intensity*)	0.7122	0.6738

Intensity), we augment the training data with gate-passing rewrites and evaluate the repaired detector under the same clean-set and attacked-set protocols. This targeted repair improves MIWS-WS ROC-AUC from 0.6574 to 0.7122 on the clean set and from 0.6294 to 0.6738 on the attacked set. Together, these results strengthen the evidence for diagnosis-to-repair in the extremist-rhetoric setting and provide a complementary counterpart to the analysis in Section 5.

R Independent Evaluation of Realized Rhetorical Properties

To directly evaluate the realized rhetorical properties of generated texts, we conducted an additional human monotonicity/ordering study. For fixed inputs and a single target axis, annotators ranked rewrites generated at increasing axis values by perceived target-axis strength, while also judging core-claim preservation and non-target leakage.

The study shows strong directional control in realized text: pairwise ordering accuracy reaches 98.2%, mean Kendall's τ_b is 0.968, core-claim preservation is 94.2%, and the low-leakage rate is 93.2%. These results indicate strong directional consistency between intended and perceived target-axis strength, while suggesting that rhetorical strengthening largely preserves the underlying claim and introduces limited unintended change along non-target axes.

Taken together, these findings support interpreting the strategy space as a practical directional control interface for adversarial rewriting: increasing a target axis typically induces perceptually meaningful and largely monotonic strengthening of the intended rhetorical property, while preserving the core claim and keeping off-target interference limited. They complement the intervention curves and perceptual-validity judgments in the main text by directly evaluating the realized rhetorical properties of the generated rewrites.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contribution as an axis-aligned rhetorical strategy space for controllable adversarial rewriting, mechanism-level attribution, and axis-guided repair. The claimed empirical improvements are supported by the main experiments on misinformation and extremist-rhetoric benchmarks, with scope limitations discussed in the limitations section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated limitations and broader impact section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results or proofs. The mathematical formulations are used to define the optimization objective, semantic gate, and strategy-space search procedure.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies the datasets, baselines, backbone model, training protocol, evaluation protocol, semantic gate, attack budget, and key hyperparameters. Additional implementation details, prompt templates, hyperparameter selection, and reproducibility information are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release our code and data.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup describes the datasets, baselines, detector backbone, optimizer, training schedule, evaluation protocol, and attack construction procedure. The appendix further reports hyperparameter settings, gate-threshold selection, query-budget analysis, and prompt specifications.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main results report ROC-AUC as mean $\pm$ standard deviation over three random seeds. The appendix provides additional statistical significance tests for the main comparisons.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the main computing infrastructure, including the GPU used for the experiments. Additional computational details, including query budgets and token usage for the ARAO search procedure, are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is intended for defensive auditing and robustness improvement of harmful-content detectors. The paper discusses dual-use risks and mitigation considerations in the limitations and broader impact section.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impact discussion covers positive impacts, including improved robustness and transparency for harmful-content detection, as well as negative impacts, including potential dual-use risks for adversarial content generation. The paper emphasizes defensive auditing and model hardening as the intended use.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We adopt a responsible release strategy for our code and selected research data. Generated adversarial examples are screened for sensitive or potentially harmful content before release, and the repository provides documentation on intended defensive use, limitations, and potential dual-use risks. Further considerations are discussed in the Broader Impact section.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use publicly available datasets and pretrained models, including FakeNewsNet, Constraint, MIWS, GHC, and BERT, following their corresponding licenses or terms of use. We cite the original papers and resources in the main text.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release the ARAO implementation and a curated subset of research data.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The paper reports a human evaluation study and describes the annotation tasks, rating scales, number of annotators, and evaluation metrics. Annotators were compensated appropriately, at or above the applicable minimum wage.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The human evaluation involved minimal-risk text-rating tasks and did not collect sensitive personal information from participants. Participants were informed of the task purpose and potential exposure to rewritten harmful-content examples, and the study was approved or deemed exempt by the relevant institutional review process.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are used as the planner and rewriter in the proposed ARAO framework, which is a core methodological component. The paper describes the LLM setting, decoding configuration, prompt templates, and additional robustness checks with alternative rewriters.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.