GAINMASQUE: AUDITING MASKED RETRIEVAL GAINS AND REPAIRING EVIDENCE INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Incorrect information can undermine content quality, veracity, and integrity, and these challenges are amplified when such information is shared within multimodal information streams. Retrieval-augmented detection methods usually incorporate external evidence to improve predictions, but retrieval gains can mask whether model predictions depend on the retrieved evidence. To solve this issue, we propose GainMasque, an auditing and repair framework that disentangles retrieval gain from genuine evidence gain. First, GainMasque audits retrieval by comparing predictions with original and controlled evidence swaps, uncovering when retrieval gains remain without claim-conditioned evidence use. Then, we delineate systematic sources of masked retrieval gain, including retrieval and fusion mechanisms that rely on dataset, source, or label-correlated signals rather than evidence. To address these failures, we further introduce Decoupled Residual Evidence Fusion (DREF), which freezes the content model, corrects test-time shifts in the evidence score through bias alignment, and controls the evidence contribution using relevance- and agreement-based trust. GainMasque is backbone-agnostic and can be applied across multiple multimodal encoders and retrieval settings. Experimental results across diverse datasets and vision-language models show that retrieval gain often does not imply evidence gain, while DREF yields evidence gain with claim-conditioned scorers and bounds losses under noisy or corrupted retrieval. Together, GainMasque provides a systematic framework for auditing what drives retrieval gains and repairing how retrieval evidence shapes predictions.

1 INTRODUCTION

Incorrect information is increasingly presented in multimodal forms that combine text, images, and other media, motivating automatic methods for content verification and detection (Shu et al., 2017; Zhou & Zafarani, 2020; Abdali et al., 2024; Zhong & Baghel, 2024; Huang et al., 2025). Recent systems increasingly augment content representations with external knowledge or retrieved evidence (Lewis et al., 2020; Zhang et al., 2019; Abdelnabi et al., 2022). However, retrieval gain alone does not show that the model actually uses claim-specific evidence. A detector may instead improve because retrieval introduces label-correlated, source-specific, availability, or fusion effects that remain predictive even when the retrieved content is not relevant to the claim.

Prior evaluations do not clearly separate these possibilities. Knowledge-augmented detectors typically measure the accuracy gained after external evidence is added, while robustness studies replace retrieved evidence with irrelevant or random context (Shi et al., 2023b; Yoran et al., 2024). Such interventions remove both claim-specific information and label-level signals, making it difficult to determine which part of the gain actually depends on claim–evidence alignment. To distinguish this distinction, the comparison must preserve label-level properties while breaking the alignment between a claim and its evidence.

We introduce GAINMASQUE, a framework for auditing and repairing retrieval-augmented multimodal incorrect information detection. GAINMASQUE distinguishes *retrieval gain* (RG), the improvement obtained after adding a claim's evidence, from *evidence gain* (EG), the portion of that improvement removed when the evidence is replaced by same-label evidence from another claim. Their difference is the *masked retrieval gain*: the gain that remains after claim–evidence alignment is broken. The resulting same-label swap preserves label, source, and evidence availability while

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

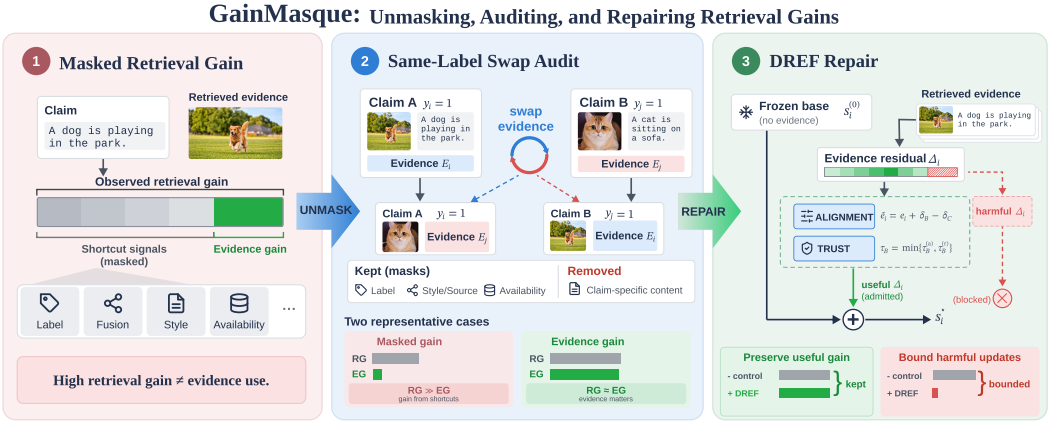


Figure 1: GAINMASQUE framework. **(1)** RG may arise from shortcuts rather than claim-specific evidence. **(2)** Same-label swaps preserve label-level cues while breaking claim–evidence alignment; EG measures the claim-specific component of RG. **(3)** DREF aligns evidence residuals to the frozen base and applies dual trust, preserving useful gains while limiting harmful updates.

removing claim-specific evidence content, allowing retrieval gain to be decomposed into evidence-dependent and masked components.

Across Twitter (Boididou et al., 2015), Weibo (Jin et al., 2017), NewsCLIPpings (Luo et al., 2021), VERITE (Papadopoulos et al., 2024), MOCHEG (Yao et al., 2023), and AVeriTeC (Schlichtkrull et al., 2023), the audit reveals four recurring sources of masked retrieval gain: **label transfer**, **fusion shift**, **source style**, and **evidence availability**. These findings show that retrieval gain can remain even when claim-specific evidence is removed. We therefore introduce Decoupled Residual Evidence Fusion (DREF), which keeps the base predictor frozen, aligns the evidence residual relative to the base prediction, and controls its contribution through relevance- and agreement-based trust. Rather than treating evidence as an independent predictor, DREF regulates when and how evidence may modify a frozen decision, preserving useful evidence gains while limiting harmful updates under unreliable retrieval.

Our contributions are threefold:

- We introduce GAINMASQUE, which separates retrieval gain from claim-specific evidence gain through a same-label evidence swap that preserves label-level signals while breaking claim–evidence alignment.
- We identify four recurring sources of masked retrieval gain—label transfer, fusion shift, source style, and evidence availability—showing that retrieval improvements can persist without claim-specific evidence use.
- We introduce DREF, a decision-level repair that combines base-referenced alignment with relevance- and agreement-based trust to preserve useful evidence contributions while limiting harmful prediction shifts, without retraining the base predictor.

Across six datasets, multiple retrieval and fusion designs, diverse vision-language backbones, and external retrieval and reading systems, retrieval gain frequently exceeds evidence gain, while DREF substantially reduces the harm caused by corrupted and adversarial evidence.

2 RELATED WORK

2.1 EVALUATING WHETHER THE EVIDENCE IS USED

Evidence use is commonly tested by removing or perturbing inputs. Partial-input baselines show that evidence alone can rival claim–evidence input in fact-checking (Hansen et al., 2021), while such failures do not by themselves validate a dataset (Feng et al., 2019). Multimodal controls permute or

remove cross-modal information (Gat et al., 2021; Hessel & Lee, 2020); retrieval-augmented models replace passages with irrelevant or random ones (Shi et al., 2023b; Yoran et al., 2024; Cuconasu et al., 2024); and dataset audits remove leaked evidence (Glockner et al., 2022). Similarity-based incorrect-information detectors can likewise perform strongly through shortcuts rather than factuality (Papadopoulos et al., 2025). These controls typically remove both claim-specific content and label-correlated signals, making their effects difficult to separate. Our same-label swap instead preserves label, source, and evidence availability while breaking claim–evidence alignment. It is related to within-class permutation (Ojala & Garriga, 2010) and constrained multimodal redundancy analysis (Liang et al., 2023), but uses this intervention to test whether retrieval gain depends on claim-specific evidence. To our knowledge, such a label-preserving swap has not been used to audit retrieval gain in fact-checking or retrieval-augmented evaluation.

2.2 Calibrating and trusting an auxiliary signal

Prior work improves auxiliary predictions through score combination and calibration (Wolpert, 1992; Platt, 1999; Steyerberg et al., 2004; Zhou et al., 2024), held-out safety constraints (Thomas et al., 2019), relevance-aware retrieval control (Yan et al., 2024), and agreement-based reliability estimation (Baek et al., 2022). DREF instead treats evidence as a residual update to a frozen predictor, aligning it relative to the base prediction and controlling its influence through complementary relevance and agreement trust. More broadly, multimodal detectors often fuse textual, visual, and knowledge features into a single representation, obscuring how auxiliary evidence changes predictions and exposing fusion to heterogeneous or unreliable signals. More importantly, retrieval gains are rarely tested against evidence that is wrong for the claim. GainMasque addresses this gap with an explicit evidence-use audit, using a modular detector and fusion mechanism to expose when noisy evidence produces masked rather than claim-specific gain.

3 Methodology

We introduce the components of GainMasque, as illustrated in Figure 1. Our objective is to measure how much of a retrieval gain is evidence gain, and then to repair how evidence shapes a frozen prediction. We first define retrieval gain, evidence gain, and masked gain together with the audit used to estimate them, and then introduce DREF, our decision-level repair. The expanded audit formulation and its modular detector instantiation are provided in Appendix A.

3.1 Problem definition: retrieval and evidence gain

Retrieval-augmented prediction is typically evaluated by comparing a model with and without retrieved evidence. Given an information unit x_i with label y_i, let E_i denote its evidence and $f(x_i, E_i)$ the model output when evidence E_i is provided. We consider three evidence conditions: the unit's related evidence E_i, no evidence $\varnothing$, and swapped evidence $E_{\pi(i)}$ from another unit with the same label. Let $\mathcal{A}_{E_i}$, $\mathcal{A}_{\varnothing}$, and $\mathcal{A}_{E_j}$ denote the corresponding accuracies; their expanded definitions are given in Appendix A.1. We define

$$\mathrm{RG} = \mathcal{A}_{E_i} - \mathcal{A}_{\varnothing}, \qquad \mathrm{EG} = \mathcal{A}_{E_i} - \mathcal{A}_{E_j}. \tag{1}$$

RG measures the change in prediction accuracy after evidence is introduced. For the swap, $E_{\pi(i)}$ comes from another unit with the same label while matching the evidence source and availability. EG measures the accuracy change induced by breaking this claim–evidence alignment. Together, RG − EG is the residual gain after the swap, which we refer to as a *masked retrieval gain*: an observed retrieval improvement that masks the absence of evidence use. Our objective is therefore not only to measure whether retrieval improves prediction, but to determine how much of that improvement actually depends on claim-specific evidence, identify why the remaining gain survives, and control how the evidence signal modifies a frozen prediction. Evidence-source conventions, including retrieved and gold evidence, are detailed in Appendix A.1.

3.2 The GainMasque audit

Same-label evidence swap. Let $\mathcal{D} = \{(x_i, y_i, E_i, S_i)\}_{i=1}^{N}$ denote an evaluation set, where S_i denotes the evidence source and $h_i = \mathbb{I}[E_i \neq \varnothing]$ indicates evidence availability. The candidate

swap partners for x_i are

$$\Omega_i = \{j \in [N] \setminus \{i\} : y_j = y_i,\ S_j = S_i,\ h_j = h_i\}. \tag{2}$$

We select $\pi(i) \in \Omega_i$ and define the swapped evidence as $E_{\pi(i)}$. For each x_i, the audit evaluates the same fixed system under E_i, $E_{\pi(i)}$, and $\varnothing$, corresponding to matched evidence, same-label swapped evidence, and no evidence, respectively. The intervention changes only the evidence while keeping the claim and prediction system fixed; the modular predictor formulation is given in Appendix A.2.

Row-level interleaving. For batch-adaptive methods, we interleave matched and swapped evidence over approximately half of each evaluation batch, so that the intervention itself does not become an unrestricted batch-level shift cue. The explicit intervention mask and construction are given in Appendix A.3.

Diagnostic controls. When a gain remains after the same-label swap, we use targeted interventions to test for event overlap, fusion effects, source style, evidence availability, and batch-level adaptation. These controls identify four recurring sources of masked retrieval gain: label transfer, fusion shift, style, and availability, and the intervention for each control is provided in Section 4.

3.3 DREF: Repairing Evidence

The GainMasque audit determines how much of a retrieval gain depends on claim-specific evidence. Then, DREF controls how the evidence signal is allowed to modify a frozen prediction. Let $s_i^{(0)}$ denote the log-odds produced by the frozen base predictor for x_i, and let e_i denote the evidence score produced by a trained evidence scorer. We write the evidence-aligned residual as

$$\Delta_i = \beta \tilde{e}_i + b, \tag{3}$$

where $\tilde{e}_i$ is the aligned evidence score and (β, b) are fitted on labeled calibration data.

Let $\mathcal{S}^+$ represent the set of evidence sources admitted on held-out calibration data. The effective contribution of evidence on instance x_i in test batch B is

$$\omega_{i,B} = \mathbb{I}\big[S_i \in \mathcal{S}^+\big]\, \mathbb{I}[E_i \neq \varnothing]\, \tau_B, \qquad \omega_{i,B} \in [0,1], \tag{4}$$

where τ_B is a label-free trust coefficient defined below. The repaired prediction is then

$$s_i^\star = s_i^{(0)} + \omega_{i,B} \Delta_i \tag{5}$$

and therefore evidence acts only as a controlled residual update to the frozen base prediction.

Base-referenced alignment. In our DREF framework, the evidence score is used as a residual rather than as an independent prediction; its test-time behavior is aligned relative to the frozen base. Let

$$\hat{y}_i^{(0)} = \mathbb{I}\Big[s_i^{(0)} > 0\Big] \tag{6}$$

denote the base prediction, and let $\mathcal{B}_E = \{i \in B : E_i \neq \varnothing\}$ denote the evidence-bearing instances in test batch B. We estimate the test-time offset

$$\delta_B = \arg\min_{\delta} \sum_{i \in \mathcal{B}_E} \mathrm{BCE}\left(\sigma(e_i + \delta), \hat{y}_i^{(0)}\right), \tag{7}$$

and align the evidence score by

$$\tilde{e}_i = e_i + (\delta_B - \delta_{\mathcal{C}}), \tag{8}$$

where $\delta_{\mathcal{C}}$ is the calibration term estimated on the labeled calibration split $\mathcal{C}$. Thus, the calibration split provides the anchor, while the frozen base provides a label-free test-time reference. At test time, only the alignment offset δ_B is re-estimated; β and b remain fixed.

Dual trust control. Alignment determines the evidence residual proposed by evidence; dual trust determines how strongly that residual may modify the frozen prediction. We use two complementary signals: agreement with the frozen base and claim–evidence relevance. Let $\mathcal{G}(B)$ measure agreement between the aligned evidence scores and base predictions on test-time batch B, and let $\mathcal{R}(B)$ denote the mean claim–evidence relevance score. We define

$$\tau_B^{(a)} = \Pi_{[0,1]}\left(\frac{\mathcal{G}(B)}{\mathcal{G}(\mathcal{C})}\right), \tag{9}$$

and

$$\tau_B^{(r)} = \Pi_{[0,1]}\left(\frac{\mathcal{R}(B) - \mathcal{R}^-}{\mathcal{R}^+ - \mathcal{R}^-}\right)\kappa(B), \tag{10}$$

where $\mathcal{C}$ is the calibration split, $\mathcal{R}^+$ and $\mathcal{R}^-$ are matched- and mismatched-evidence relevance references, respectively, and $\kappa(B) \in [0,1]$ measures consistency with the calibration distribution. The final trust coefficient is the conservative conjunction

$$\tau_B = \min\left\{\tau_B^{(a)}, \tau_B^{(r)}\right\}. \tag{11}$$

The two factors address complementary failure modes: relevance suppresses evidence that is poorly matched to the claim, whereas agreement suppresses evidence whose proposed update conflicts with the frozen base prediction. Detailed definitions of $\mathcal{G}(B)$, $\mathcal{R}(B)$, $\mathcal{R}^+$, $\mathcal{R}^-$, and $\kappa(B)$ are provided in Appendix C.

4 EXPERIMENTS

In this section, we conduct experiments on multiple multimodal benchmark datasets to evaluate our GAINMASQUE audit and DREF repair framework. We first introduce the experimental setup, including the datasets, evaluated models, evidence sources, and experimental settings. We then report the audit results and analyze the sources of masked retrieval gain. Furthermore, we evaluate DREF under different backbones, corrupted evidence, search agents, and external evidence readers to study its robustness and generalization.

4.1 EXPERIMENTAL SETUP

Scope. The GAINMASQUE audit covers six datasets—Twitter (Boididou et al., 2015), Weibo (Jin et al., 2017), NewsCLIPpings (Luo et al., 2021), VERITE (Papadopoulos et al., 2024), MOCHEG (Yao et al., 2023), and AVeriTeC (Schlichtkrull et al., 2023)—with DREF evaluated on MOCHEG and AVeriTeC and its admission mechanism additionally tested on NewsCLIPpings, VERITE, and Twitter. The main backbone is Qwen3-VL-8B; additional experiments cover multiple integration designs and retrieval/search agents, twelve VLMs (2B–72B), and open and closed readers across training-bank, gold, retrieved-passage, and live-web evidence.

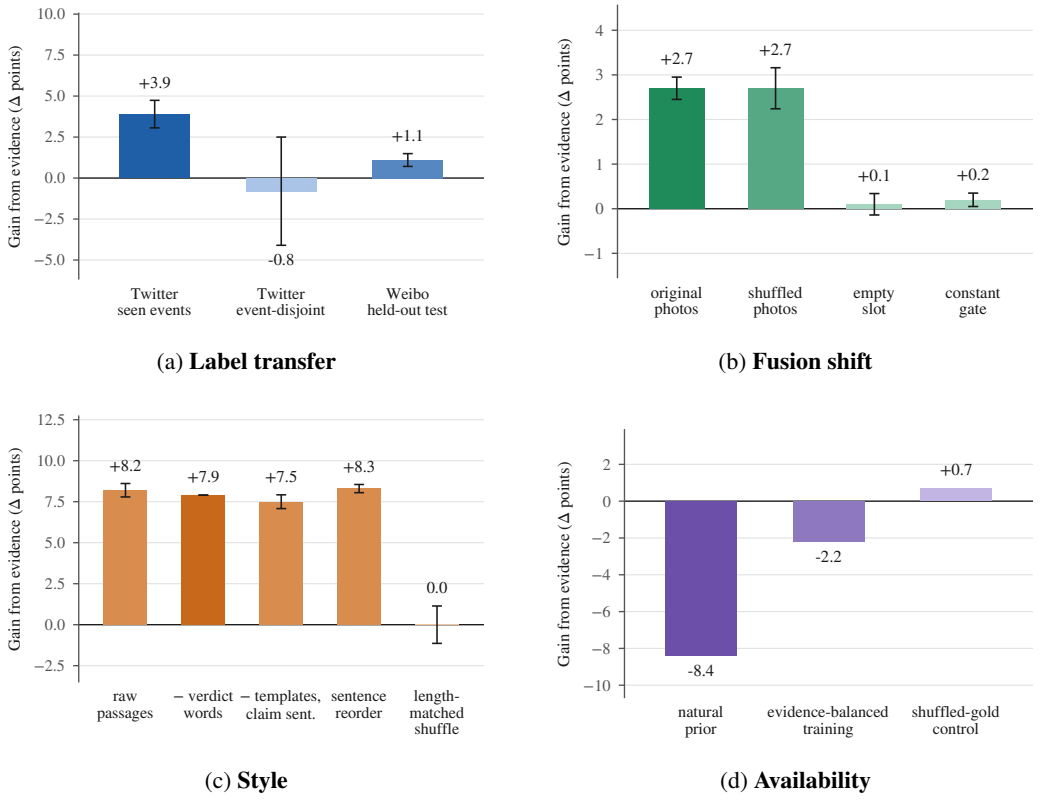
Protocol. Unless otherwise specified, results are mean $\pm$ standard deviation over three seeds; the MOCHEG backbone study reports the median across twelve models. Each EG uses one same-label swap draw, and robustness spans eight corruption families at multiple rates. Full details are in Appendix I.

4.2 RETRIEVAL GAIN DECOMPOSITION

We first examine whether retrieval gain depends on claim-specific evidence. Table 1 shows that most positive training-bank gains and the gated-fusion gain survive same-label swapping; across the six datasets, EG is non-positive in 31 of 36 training-bank measurements. The MOCHEG evidence scorer retains a substantial masked component, whereas the live reverse-image and reader-conditioned settings lose the gain under swapping. Availability is the exception because the swap preserves evidence presence, so its effect is exposed by a separate control. Section 4.3 analyzes the mechanisms behind these masked gains.

Table 1: **Masked retrieval gains.** MG measures the gain that remains after claim–evidence alignment is broken: MG = RG − EG.

Design	Dataset	RG ↑	EG ↑	MG	Diagnosis
Full gain masked Swapping removes little of the reported gain.					
Training bank, seen	Twitter	+3.9	−0.5	**+4.4**	LABEL TRANSFER
Training bank	Weibo	+1.1	−2.8	**+3.9**	LABEL TRANSFER
Gated fusion	NewsCLIPpings	+2.0	+0.2	**+1.8**	FUSION SHIFT
No gain to decompose Retrieval does not improve over the no-retrieval baseline.					
Training bank, unseen	Twitter	−0.8	−0.1	−0.7	NO GAIN
Most gain masked Only a small evidence-specific component remains.					
Evidence scorer	MOCHEG	+7.3	+1.7	**+5.6**	SOURCE STYLE
Availability Evidence availability changes performance.					
Evidence fusion	MOCHEG	−8.4	—	—	AVAILABILITY
Gain is evidence-dependent Swapping removes the gain or makes the evidence noisy.					
Live reverse-image	NewsCLIPpings	+3.4	+3.4	**0.0**	EVIDENCE-DEPENDENT
Reader-conditioned	AVeriTeC	+11.2	+14.4	**−3.2**	EVIDENCE-DEPENDENT



(a) **Label transfer** (b) **Fusion shift**

(c) **Style** (d) **Availability**

Figure 2: Four sources of masked retrieval gain. Bars are gains from evidence in accuracy points under each diagnostic; error bars show sample standard deviations where available.

Table 2: **DREF repair component trade-offs.** RG is the three-seed mean clean gain over the no-evidence base using gold evidence. $\Delta_{\min}$ is the minimum, over corruption families, of the three-seed mean per-seed worst accuracy change across $r \in \{0.25, 0.5, 0.75, 1.0\}$, using corrupted variants of the same evidence (higher is better).

Configuration	MOCHEG		AVeriTeC	
	RG ↑	$\Delta_{\min}$ ↑	RG ↑	$\Delta_{\min}$ ↑
DREF repair variants				
Residual	-8.6	-56.1	+12.8	-2.7
Base-referenced alignment	+7.9	-66.6	+11.7	-2.1
Alignment + relevance trust	+7.9	-12.5	+10.7	-0.6
Alignment + agreement trust	+7.4	-1.6	+10.5	-2.1
DREF (dual trust)	**+7.4**	**-0.6**	**+10.4**	**-0.6**

4.3 Sources of Masked Retrieval Gain

We next ask why retrieval gain can remain after claim–evidence alignment is broken. Figure 2 summarizes four recurring mechanisms identified by the GainMasque audit; complete diagnostic controls are provided in Appendix A.6.

Masked-gain sources. On Twitter, the training-bank gain appears for seen events but disappears on the event-disjoint split; within the bank, 86% of examples share an image with a nearest neighbor with 100% label agreement, supporting label transfer rather than claim-specific evidence use. On NewsCLIPpings, shuffling retrieved images preserves the gain, while empty-input and constant-gate controls nearly remove it, indicating an input-dependent fusion shift. On MOCHEG, the gain survives targeted content removal and sentence reordering but vanishes under length-matched shuffling, implicating label-correlated source style. Separately, balancing evidence availability largely removes the availability effect, showing that evidence presence itself can act as a shortcut that survives the swap.

Evidence-dependent gains. By contrast, the live reverse-image source on NewsCLIPpings and the reader-conditioned scorer on AVeriTeC lose the retrieval gain under same-label swapping. Removing the claim from the AVeriTeC reader further lowers AUROC from 0.964–0.970 to 0.78–0.81. Together, these results support joint claim–evidence conditioning as an important ingredient of evidence-dependent gain.

4.4 DREF repair framework

Table 2 shows that base-referenced alignment restores a positive gain from the failing calibration-fitted residual, but remains highly vulnerable to corrupted evidence; neither batch calibration nor prior adaptation repairs this shift (Table 9). Agreement and relevance trust are complementary across MOCHEG and AVeriTeC, and their combination limits the worst-case loss to -0.6 points on both datasets while retaining most of the aligned gain. Across twelve MOCHEG backbones, the retained gain has a median EG of +0.6, compared with +1.7 before trust. On AVeriTeC, DREF preserves the evidence gain while substantially improving robustness relative to the scorer alone.

4.5 Robustness and generalization

Finally, we test whether DREF reduces evidence sensitivity while retaining reference-evidence accuracy. Figure 3 reports corruption sensitivity,

$$S_{\text{corr}} = \frac{1}{2}\left[A(E_i) - A(E_{\pi(i)}) + A(E_i^{\text{bare}}) - A(\mathcal{R}(E_i^{\text{bare}}))\right],$$

the mean accuracy drop under same-label swapping and reversed-direction evidence, where E_i^{bare} is the matched same-direction control and $\mathcal{R}(E_i^{\text{bare}})$ reverses its decision direction. Across the six forced-choice API readers, DREF reduces sensitivity by 3.0–4.9 points (mean 4.15) for a 0.5–1.0-point reduction in reference-evidence accuracy (mean 0.84); only 0.42 points of the mean sensitivity

Table 3: **DREF across 12 frozen backbones on MOCHEG.** Gold-evidence results are three-seed means; retrieval uses relevance-weighted lexical top-3 passages with dual trust. Rows are ordered by size. Δ, $\Delta_{\min}$, and Δ_{ret} are point changes from base accuracy A_0; $\Delta_{\min}$ is the worst loss across eight corruption families and rates 0.25–1.0. AUROC/NLL show base/DREF values, and τ_B is agreement trust on clean gold evidence. Trust and alignment are re-estimated per evaluated batch.

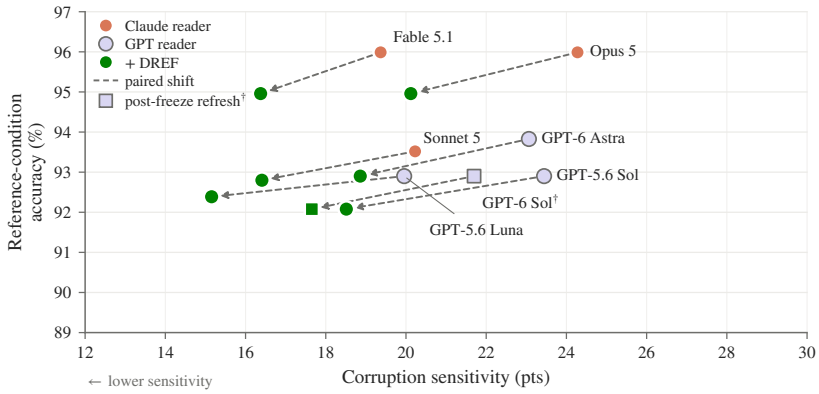
Backbone	Base		Alignment		DREF					Ret.	Audit
	A_0	P_{conf}	Δ	$\Delta_{\min}$	Δ	AUROC	NLL	$\Delta_{\min}$	τ_B	Δ_{ret}	EG
Qwen3-VL-2B	0.615	0.671	+14.1	-61.2	+0.0	0.679 → 0.679	0.678 → 0.678	-0.2	0.00	+5.0	-4.3
Qwen3.5-2B	0.632	0.710	+11.2	-58.0	**+11.1**	0.721 → 0.835	0.655 → 0.521	-0.8	0.79	+5.5	+3.3
Qwen3.5-4B	0.658	0.745	+9.1	-58.5	**+6.8**	0.744 → 0.835	0.622 → 0.525	-0.5	0.60	+4.0	+0.1
Qwen3-VL-4B	0.688	0.779	+8.0	-63.5	**+7.8**	0.763 → 0.867	0.591 → 0.474	-0.2	0.82	+3.7	+0.6
Qwen2.5-VL-7B	0.724	0.813	+6.3	-58.5	**+5.6**	0.789 → 0.874	0.559 → 0.457	-1.2	0.87	+2.2	-0.1
InternVL3.5-8B	0.705	0.792	+6.6	-62.0	**+7.1**	0.782 → 0.858	0.582 → 0.484	-0.1	0.63	+3.5	+3.7
Qwen3-VL-8B	0.708	0.790	+7.9	-66.6	**+7.4**	0.793 → 0.878	0.590 → 0.451	-0.6	0.76	+3.7	+1.2
Qwen3.5-9B	0.700	0.796	+7.0	-56.7	**+5.1**	0.805 → 0.870	0.591 → 0.495	-1.1	0.55	+4.0	+0.6
Gemma-4-12B	0.723	0.821	+7.6	-65.4	**+7.6**	0.821 → 0.892	0.549 → 0.421	-0.4	0.79	+1.9	-0.1
Qwen3-VL-32B	0.738	0.838	+6.7	-63.8	**+6.6**	0.822 → 0.898	0.544 → 0.417	-0.7	0.73	+2.2	+1.9
InternVL3-38B	0.748	0.852	+6.2	-59.2	**+6.0**	0.834 → 0.895	0.528 → 0.418	-0.7	0.80	+3.5	+0.3
Qwen2.5-VL-72B	0.778	0.902	+4.2	-56.1	**+4.1**	0.877 → 0.908	0.468 → 0.404	-0.4	0.73	+1.9	+0.9

Table 4: A closed decision-native reader on AVeriTeC. Accuracy under E_i is evaluated on the 324 test claims, except for the reversed-direction condition $\mathcal{R}(E_i)$, which applies to 136 claims; accuracy under $E_{\pi(i)}$ averages five partner draws per seed. $\Delta_{\min}$: the worst degradation against the claim-based reference on the matching claim set. $\text{NLL}(E_i)$ is negative log-likelihood under original evidence, and $\overline{\text{NLL}}$ is its mean over the three evidence conditions; both measure probabilistic predictive performance. Seed means $\pm$ sample standard deviation.

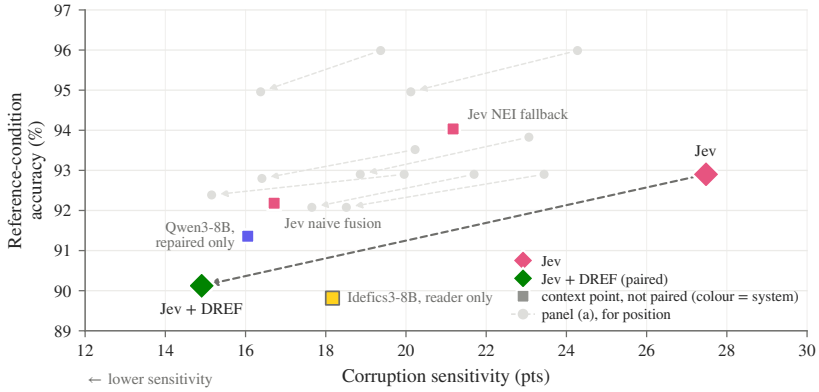
System	Accuracy ↑			Robustness ↑	NLL ↓	
	E_i	$E_{\pi(i)}$	$\mathcal{R}(E_i)$	$\Delta_{\min}$	$\text{NLL}(E_i)$	$\overline{\text{NLL}}$
Reference						
Baseline	0.786 ± 0.012	0.786 ± 0.012	**0.814 ± 0.019**	−−	0.463	**0.461**
External reader and simple integrations						
Jev	0.929 ± 0.000	0.681 ± 0.000	0.640 ± 0.000	-0.174	0.261	1.012
Unweighted sum	0.922 ± 0.002	0.796 ± 0.006	0.743 ± 0.000	-0.071	0.275	0.990
Abstention fallback	**0.940 ± 0.005**	0.789 ± 0.009	0.664 ± 0.024	-0.150	0.241	0.773
Repair (ours)						
Jev + DREF	0.901 ± 0.011	**0.802 ± 0.008**	0.752 ± 0.017	**-0.061**	**0.211**	0.531

reduction is explained mechanically by this accuracy change. Jev shows a larger shift (12.57 points in sensitivity for 2.78 points in reference accuracy), although its abstention behavior makes the displacement not directly comparable with the forced-choice readers.

Backbones and external readers. Across twelve backbones from four architecture families and 2B–72B parameters, DREF retains positive gains on eleven of twelve while keeping the worst corruption loss within 1.2 points of the base (Table 3). Qwen3-VL-2B is the exception: zero agreement trust rejects its large aligned gain and returns to the base, while a similarly sized Qwen3.5-2B retains



(a) Frontier API readers, categorical decisions, and their DREF repairs.



(b) Broader reader landscape, same axes and limits; panel (a) repeated in gray for position.

Figure 3: DREF reduces corruption sensitivity while retaining most reference-evidence accuracy (i.e., accuracy under E_i (%)) on the frozen AVeriTeC test claims. Arrows connect paired readers to their DREF repairs.

its gain, showing that the trust decision is not determined by model size alone. For Jev, DREF gives up 2.8 reference-evidence points but substantially improves both intervention robustness and NLL (Table 4).

Corruption and adversarial retrieval. Across all corruption families and rates, alignment alone can amplify directional failures, whereas relevance and agreement suppress complementary failure modes; the full repair remains within one point of the no-evidence base (Figures 4 and 5). Under system-specific adversarial passage selection, DREF loses 3.4–4.0 points below its no-evidence accuracy, compared with 7.1–15.4 points for the six search agents (Appendix Table 14).

5 Conclusion

Retrieval gain does not necessarily imply claim-specific evidence use. We introduced GainMasque to separate retrieval gain from evidence gain and showed that apparent gains can persist through label transfer, fusion shift, source style, and evidence availability. We further introduced DREF, which keeps the base predictor frozen, aligns the evidence residual, and controls its contribution through relevance- and agreement-based trust. Across datasets, retrieval settings, model families, and closed external readers, DREF preserves useful evidence gains while limiting harmful updates under unreliable retrieval. Together, GainMasque provides a framework for auditing what drives retrieval gains and repairing how retrieved evidence shapes multimodal predictions.

REPRODUCIBILITY STATEMENT

For testing detector, the hyperparameters and the twelve backbones are described in Appendix I, the prompts that generate features in Appendix K, and the fusion in Appendix **??**. The experimental datasets, splits, and three disclosures related to how the numbers should be read are in Appendix G, and the evidence sources in Appendix E. Appendix A shows the swap, including how a partner is selected, what is held fixed and what is not refitted; Appendix C presents the repair, its assumption, admission test, declared constants, and pseudocode; and Appendix D defines the corruption suite. Appendix B gives the statistical methods and supporting results, Appendix H the extended discussion and limitations, and Appendix F the full results for the closed external readers.

The anonymised supplementary materials contain the implementation, the results for every open-model table, per-row predictions, the closed readers' frozen responses, and manifests of the environment, declared constants and input artifacts. Feature extraction is the only step that needs a GPU; everything after it is deterministic and can be re-run on CPU from the frozen feature caches. The materials give the commands, required files, and recorded hashes, and Appendix I documents the one exception to the pinned environment.

Google Cloud Vision (`WEB_DETECTION`) is a live provenance source, and therefore its evidence is reproducible only from our frozen snapshot, not by re-querying. The external reader and the seven frontier readers in Section 4.5 are closed APIs; Appendices J and F record what was frozen, when it was queried, and the settings needed to interpret those results.

AI USE STATEMENT

In this work, we used generative AI tools as assistive tools for implementing comparison methods, processing datasets, and providing feedback on experimental design and the interpretation of results. The authors proposed the research questions and hypotheses, formulated the conceptual framework and methodology, designed and selected the final experiments, produced and analysed the experimental results, and determined the claims and conclusions reported in the paper. Generative AI was not used to formulate mathematical claims, provide proofs, or write proofs. Generating synthetic datasets, translation, and qualitative or thematic data analysis are not applicable to this work.

We additionally used generative AI tools to identify and examine relevant literature, search for information, edit code, modify scientific figures, and edit or rephrase author-written text for clarity and readability. We reviewed all AI-assisted work. LLM-assisted code was read, executed, and tested by one author, and all reported experimental results were obtained by running the resulting code and checked by the authors. Data-processing scripts written with LLM assistance were checked against the raw data, and AI-assisted figures were checked against the experimental results. Prior work surfaced through LLM-assisted search was located, read, and verified against the original publications. AI-assisted edits to the manuscript were checked against the authors' intended meaning and the proposed methods, results, and claims.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Computing Surveys*, 57(3):1–29, 2024.

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14920–14929, 2022. doi: 10.1109/CVPR52688.2022.01452.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J. Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. Explainable Tsetlin machine framework for fake news detection with credibility score assessment. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4894–4903, Marseille, France, 2022. European Language Resources Association.

Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, and Yiannis Kompatsiaris. Verifying multimedia use at MediaEval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, volume 1436 of *CEUR Workshop Proceedings*, 2015.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. ReSearch: Learning to reason with search for LLMs via reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference*, pp. 2897–2905, 2022.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2114–2119, 2020.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 719–729, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 49250–49267, 2023.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. Misleading failures of partial-input baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5533–5538, 2019.

Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Max Glockner, Yufang Hou, and Iryna Gurevych. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5916–5936, 2022.

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys*, 53(4):1–36, 2020.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3929–3938, 2020.

Casper Hansen, Christian Hansen, and Lucas Chaves Lima. Automatic fake news detection: Are models learning to reason? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 80–86, 2021.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 861–877, 2020.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28831–28841, 2025.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 795–816, 2017.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. MVAE: Multimodal variational autoencoder for fake news detection. In *Proceedings of the World Wide Web Conference*, pp. 2915–2921, 2019.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1746–1751, 2014.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4821–4830, 2019.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3122–3130, 2018.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 34892–34916, 2023.

Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6801–6817, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 3818–3824, 2016.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys*, 54(3):1–40, 2021.

Mikel K. Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. Decoding fake news and hate speech: A survey of explainable AI techniques. *ACM Computing Surveys*, 57(7):1–37, 2025.

Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863, 2010.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. VERITE: A robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4, 2024. doi: 10.1007/s13735-023-00312-6.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. Similarity over factuality: Are we making progress on multimodal out-of-context misinformation detection? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5570–5579, 2025.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Devendra Singh Sachan, Manzil Zaheer, and Ruslan Salakhutdinov. Revisiting LSTM networks for semi-supervised text classification via mixed objective function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6940–6948, 2019.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems*, volume 36, pp. 65128–65167, 2023. doi: 10.52202/075280-2842.

Philipp Schmid and Hannah Bauer. Impact of exposure to health misinformation on belief in health misinformation: A meta-analysis of RCTs. *Health Communication*, 41(5):877–887, 2026. doi: 10.1080/10410236.2025.2536772.

Chongyang Shi, Yijun Yin, Qi Zhang, Liang Xiao, Usman Naseem, Shoujin Wang, and Liang Hu. Multiview clickbait detection via jointly modeling subjective and objective preference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11807–11816, 2023a.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227, 2023b.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-Searcher: Incentivizing the search capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025.

Ewout W. Steyerberg, Gerard J. J. M. Borsboom, Hans C. van Houwelingen, Marinus J. C. Eijkemans, and J. Dik F. Habbema. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Statistics in Medicine*, 23(16):2567–2586, 2004.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. Motivations, methods and metrics of misinformation detection: An NLP perspective. *Natural Language Processing Research*, 1(1–2): 1–13, 2020.

Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468): 999–1004, 2019.

Kjerstin Thorson, Emily Vraga, and Brian Ekdale. Credibility in context: How uncivil online commentary affects news credibility. *Mass Communication and Society*, 13(3):289–313, 2010.

Jens Van Nooten and Walter Daelemans. Jump to hyperspace: Comparing Euclidean and hyperbolic loss functions for hierarchical multi-label text classification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4260–4273, 2025.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 849–857, 2018.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

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*, pp. 3367–3378, 2024.

Tong Wu, Feiran Jia, Xiangyu Qi, Jiachen T. Wang, Vikash Sehwag, Saeed Mahloujifar, and Prateek Mittal. Uncovering adversarial risks of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 37456–37495, 2023.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

Chang Yang, Peng Zhang, Wenbo Qiao, Hui Gao, and Jiaming Zhao. Rumor detection on social media with crowd intelligence and ChatGPT-assisted networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5705–5717, 2023.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, 2016.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2733–2743, 2023. doi: 10.1145/3539618.3591879.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024.

Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1942–1951, 2019.

Zhongxing Zhang, Emily K. Vraga, Jisu Huh, and Jaideep Srivastava. BiMind: A dual-head reasoning model with attention-geometry adapter for incorrect information detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15674–15693, 2026.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, 2025.

Yang Zhong and Bhiman Kumar Baghel. Multimodal understanding of memes with fair explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2007–2017, 2024.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *International Conference on Learning Representations*, 2024.

Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40, 2020.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. Similarity-aware multi-modal fake news detection. In *Proceedings of the 24th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 354–367, 2020.

A GainMasque audit

A.1 Expanded retrieval- and evidence-gain definitions

The audit is instantiated on a modular detector whose separate reasoning heads allow GainMasque to isolate the contribution of retrieved evidence and measure how predictions change under no-evidence, original-evidence, and swapped-evidence conditions. The detector is backbone-agnostic and can be instantiated with different pretrained encoders.

Retrieval-augmented prediction is typically evaluated by comparing a model with and without retrieved evidence. Given an information unit x_i with label y_i, let E_i denote its evidence and $f(x_i, E_i)$ represent a model output when provided with evidence E_i. We explore three evidence conditions: the unit's related evidence E_i, no evidence $\varnothing$, and swapped evidence E_j taken from another unit x_j with the same label but $j \neq i$.

Here, $\mathcal{A}_{E_i}$, $\mathcal{A}_{\varnothing}$, and $\mathcal{A}_{E_j}$ denote the model accuracy under these three conditions, respectively:

$$\mathcal{A}_{E_i} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\arg\max_k f_k(x_i, E_i) = y_i\right], \tag{12}$$

$$\mathcal{A}_{\varnothing} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\arg\max_k f_k(x_i, \varnothing) = y_i\right], \tag{13}$$

$$\mathcal{A}_{E_j} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\arg\max_k f_k(x_i, E_{\pi(i)}) = y_i\right]. \tag{14}$$

Here, k is the class index and N is the number of testing samples.

As defined in Section 3.1,

$$\mathrm{RG} = \mathcal{A}_{E_i} - \mathcal{A}_{\varnothing}, \qquad \mathrm{EG} = \mathcal{A}_{E_i} - \mathcal{A}_{E_j}. \tag{15}$$

RG measures the change in prediction accuracy after evidence is introduced. E_i comes from any source, including retrieved evidence or gold evidence. Here, gold evidence refers to the reference evidence provided or annotated by the dataset for a claim, rather than evidence retrieved by the model.

We construct a same-label swapped evidence set E_j, where

$$y_j = y_i, \qquad j \neq i,$$

while matching the evidence source and availability. EG is defined as the accuracy after the same-label swap. Together with RG, $\mathrm{RG} - \mathrm{EG}$ is the residual gain after the swap, which we refer to as a *masked retrieval gain*: an observed retrieval improvement that masks the absence of evidence use.

Our objective is therefore not only to measure whether retrieval improves prediction, but to determine how much of that improvement actually depends on claim-specific evidence, identify why the remaining gain survives, and control how the evidence signal modifies a frozen prediction.

A.2 Modular audit instantiation

Let $\mathcal{D} = \{(x_i, y_i, E_i, S_i)\}_{i=1}^{N}$ denote an evaluation set, where S_i is the evidence source and

$$h_i = \mathbb{I}[E_i \neq \varnothing]$$

indicates evidence availability.

We separate the prediction module into a base predictor and an evidence-conditioned predictor:

$$z_i^Q = f_Q(x_i), \qquad z_i^E(E_i) = f_E(x_i, E_i), \tag{16}$$

where f_Q maps the claim to base logits $z_i^Q \in \mathbb{R}^K$, while f_E maps the claim together with evidence E_i to evidence-conditioned logits $z_i^E(E_i) \in \mathbb{R}^K$.

The final prediction may combine these outputs through an implementation-specific fusion function g,

$$z_i(E_i) = g\Big(z_i^Q, z_i^E(E_i)\Big), \qquad \hat{y}_i(E_i) = \arg\max_k z_{i,k}(E_i). \tag{17}$$

The GAINMASQUE audit intervenes only on E_i while keeping x_i, f_Q, f_E, and g fixed.

A.3 SAME-LABEL SWAP AND ROW-LEVEL INTERLEAVING

For completeness, the compact partner set in Section 3.2 can equivalently be written by defining $j \sim i$ when two units match in label, evidence source, and evidence availability:

$$j \sim i \iff \begin{cases} y_j = y_i, \\ S_j = S_i, \\ h_j = h_i. \end{cases} \tag{18}$$

The swap partners of x_i are then

$$\Omega_i = \{\, j \in [N] \setminus \{i\} \mid j \sim i \,\}. \tag{19}$$

We select $\pi(i) \in \Omega_i$ and define the swapped evidence as $E_{\pi(i)}$. For each x_i, the audit evaluates the same fixed system under E_i, $E_{\pi(i)}$, and $\varnothing$, corresponding to matched evidence, same-label swapped evidence, and no evidence, respectively.

For a batch B, we construct a mixed evaluation batch using a binary intervention mask $m_i \in \{0, 1\}$:

$$\widetilde{E}_i = \begin{cases} E_{\pi(i)}, & m_i = 1, \\ E_i, & m_i = 0, \end{cases} \qquad \sum_{i \in B} m_i = \lfloor |B|/2 \rfloor. \tag{20}$$

This row-level interleaving is used for batch-adaptive methods so that replacing an entire batch does not itself become an unrestricted distribution-level signal. The partner-selection procedure, what remains fixed, and the full-batch control are described below.

A.4 DIAGNOSTIC CONTROLS

When a gain remains after the same-label swap, we use targeted interventions to test for event overlap, fusion effects, source style, evidence availability, and batch-level adaptation. These controls identify four recurring sources of masked retrieval gain: label transfer, fusion shift, style, and availability. The intervention for each control is provided in Section **??**.

A.5 THE SAME-LABEL SWAP

Choosing a partner. For a unit x_i with label y_i, candidate partners are the evaluated units that share y_i and i's evidence availability, excluding x_i itself. A hash of x_i's index selects the partner deterministically; repeated draws offset the hash. The partner's evidence is read from the same array as in the unmodified run, and the source, retriever, k, representation, and pooling remain fixed while the claim–evidence pairing changes. If no candidate exists, the swap is a no-op for that unit; this guard cannot fire on MOCHEG gold, where every test claim has evidence and both labels contain hundreds of examples.

Interleaving. We apply the swap to approximately half of each evaluation batch, selected deterministically by row index: 534 of 1,068 MOCHEG claims and 163 of 324 AVeriTeC claims. Clean and swapped accuracies are both evaluated on these attacked rows, using the same trained model. Interleaving is necessary because the repair is batch-adaptive: its alignment and trust statistics are computed over the deployment batch, so replacing the entire batch could itself trigger the repair. Keeping roughly half of the batch intact reduces this distribution-level signal. Because the deployment statistics are still recomputed on the mixed batch, however, the row-level diagnostic should not be interpreted as a strictly unit-local intervention.

What is not refitted. The evidence scorer, residual, alignment reference and trust references are fitted once on the calibration split with clean evidence. A swapped condition changes only deployment-time evidence; no model component is retrained or recalibrated.

Partner draws. We explore whether the diagnostic depends materially on the particular same-label partner. Across five whole-batch draws, the standard deviation of swapped accuracy is at most 1.16 points; dual trust reduces it to 0.45 on MOCHEG and 0.00 on AVeriTeC. We therefore do not interpret differences of roughly one point between evidence gains.

The bank's swap. Training-bank designs are audited inside the detector. Each attacked query receives the neighbor set, indices and similarities of the next same-label unit in the evaluation batch, cyclically. The bank, k, self-exclusion and twin-exclusion rules, and the label-conditional neighbor distribution remain unchanged. The intervention is applied only at evaluation.

Row-level against full-batch. We use the full-batch swap to distinguish claim–evidence disruption from detection of a distribution-wide intervention. The distinction matters once trust is adaptive: on MOCHEG, adding relevance trust yields only +0.31 row-level EG but +6.24 under a full-batch swap. The latter arises because the trust statistic detects the fully swapped batch and suppresses the residual, rather than because the evidence remains useful for its paired claim. We therefore use the interleaved row-level quantity for the evidence audit and treat the full-batch result only as a diagnostic of batch-level detection.

Where the swap sits among evidence controls. Common evidence controls remove both claim alignment and label information: modalities are permuted or removed (Gat et al., 2021; Hessel & Lee, 2020), while dataset audits remove leaked evidence from the data itself (Glockner et al., 2022). Our control instead performs a deterministic within-class reassignment of the evidence block, related to conditional permutation tests (Ojala & Garriga, 2010) and constrained multimodal comparisons (Liang et al., 2023). Preserving the label while changing the claim–evidence pairing lets the audit distinguish label-level signal from information tied to that pairing. Interleaving further prevents a batch-adaptive method from receiving the intervention itself as an unrestricted distribution-shift cue (Wu et al., 2023).

Why no fraction of one gain in the other is quoted. We report RG and EG separately rather than expressing one as a fraction of the other, since that ratio is informative only when RG is sufficiently positive and $\mathrm{EG} \in [0, \mathrm{RG}]$.

What the swap cannot see. The swap holds evidence availability fixed, and it cannot detect a gain carried solely by evidence presence. Nor does it test train–test overlap, which we audit separately through the split and duplicate checks below. Random shuffling answers a different question because it removes label-level signal together with claim alignment.

A.6 Diagnostics behind the ladder

The controls and what each separates. The remaining controls ask why a gain survives the same-label swap. Split and duplicate checks probe overlap with the evidence pool; empty evidence and a constant gate probe whether receiving evidence itself matters; scrubbing and length-matched shuffling probe source cues and content; availability masking removes the presence cue held fixed by the swap; and the whole-batch swap probes distribution-level detection. Together, these controls help distinguish gains associated with label-level transfer, fusion, source cues, availability and claim–evidence alignment.

Scrub levels. We ask whether the MOCHEG gain can be explained by obvious fact-checking language or claim restatement. Removing verdict and fact-checking vocabulary, deleting sentences flagged as claim-restating, and reordering sentences leaves the 8B alignment-based gain between +7.46 and +8.33 points. In contrast, replacing each claim's passages with a length-matched passage set chosen without using the label reduces the gain to -0.03. The gain therefore survives the tested lexical and structural cues but not the removal of the original claim–evidence pairing, consistent with the relevant signal residing beyond those surface cues.

Split integrity on MOCHEG. We next test whether overlap between train, test and the evidence pool could account for the result. The 1,068 test claims come from distinct fact-check articles absent from training, and only 0.4% of test passages have a near-duplicate in training; no test claim crosses the specified paraphrase threshold with a train or test claim. Grouping the few linked examples before bootstrapping leaves the repair interval essentially unchanged ([+5.43,+11.24] by row versus [+5.53,+11.23] by group). These checks make simple article or near-duplicate overlap an unlikely explanation for the observed gain.

Duplicate images on Twitter. We ask why training-bank retrieval can carry strong label information on Twitter. Across seeds, about 86% of bank-member queries retrieve the same image as their top neighbor, and removing same-image neighbors lowers top-five label consistency from 0.955 to 0.79. The corresponding effect is essentially absent on Weibo, MOCHEG and NewsCLIPpings. This diagnostic is measured on bank members rather than the held-out rows used for the +3.9-point retrieval gain, so it identifies a plausible route for label transfer rather than attributing that gain directly to duplicate images.

Availability. Finally, we test whether the availability of evidence itself provides a class cue. Under the natural availability imbalance, the fusion recovers 0.97 of refuted claims at only 0.53 precision, consistent with substantial reliance on availability. Equalising evidence availability during training removes this advantage: every feature-level fusion of gold text falls $2.15-4.43$ points below its separately trained no-retrieval counterpart, while the gated design's +0.66 is nearly identical to its shuffled control's +0.72. Thus the apparent gain in this setting is not supported as a content-driven evidence effect once the availability cue is removed.

B STATISTICS, ACCURACIES, AND FURTHER RESULTS

Statistical methods. Different comparisons use resampling schemes matched to their experimental unit. The fusion grid of Section 4.2 uses a cluster bootstrap over the six datasets, with each sampled dataset contributing its three seed cells. The confirmatory MOCHEG comparisons use paired per-row bootstraps with 10,000 resamples and are reported separately by seed. As a dependence check, we also bootstrap groups of MOCHEG claims linked by a shared article, near-duplicate passage or paraphrase; Appendix A reports the same inferential conclusion as the row-level analysis. The admission test and batch-detection intervals use 2,000 bootstrap resamples of calibration rows and batch-level statistics, respectively, and the external-reader analysis of Section 4.5 uses 10,000 claim-clustered resamples.

Confirmatory intervals on MOCHEG. We first ask whether the pre-declared MOCHEG comparisons persist across seeds. Six of the eight comparisons exclude zero on all three seeds (Table 5). The two exceptions involve retrieved passages: the dual-trust gain over the base is not established under mean pooling, but relevance-weighted pooling excludes zero on every seed, whereas the own-versus-shuffled retrieved comparison remains not established. Thus, the confirmatory results support the main gold-evidence and robustness comparisons while leaving the retrieved-evidence result more qualified.

Absolute accuracies. Table 6 provides the operating points behind the differences reported in the main text. Absolute accuracy spans about 31 points across the six datasets, so the same change in points can occur in substantially different operating regimes. The two VERITE tasks also do not show a consistent retrieval effect: the training bank changes accuracy by +1.76 and 0.00 points, while shuffled evidence changes it by +0.41 and -0.87. We therefore report these results descriptively rather than interpret the absolute differences alone as claim-specific evidence use.

How many neighbors. We next test whether the training-bank result is an artifact of the default choice of five neighbors. Across the 18 dataset-by-seed cells, the concatenation design's evidence-branch gain moves from +0.06 at one neighbor to -0.37 at five and -0.68 at twenty, while label-adversarial harm increases from +2.21 to +3.78 and +5.32. Increasing the neighborhood therefore does not recover the lost accuracy and instead increases exposure to harmful retrieval. The neighbor-token design shows the same qualitative pattern at higher harm.

Table 5: Pre-declared confirmatory comparisons on MOCHEG with Qwen3-VL-8B. Entries are accuracy differences in points for each seed. The final column indicates whether the paired per-row bootstrap 95% confidence interval, based on 10,000 resamples, excludes zero on all three seeds.

Comparison	seed 1	seed 2	seed 3	all excl. 0
Dual trust vs base, gold	+8.33	+5.99	+7.87	yes
Alignment vs base, gold	+7.87	+7.87	+7.87	yes
Dual trust vs base, retrieved (mean)	+2.72	+1.87	+2.90	no
Dual trust vs base, retrieved (weighted)	+4.49	+3.09	+3.46	yes
Own vs shuffled retrieved (dual)	+2.62	+1.69	+3.00	no
Own vs same-label shuffled gold (dual)	+6.65	+5.15	+5.34	yes
Dual vs alignment, label-adversarial at rate 1	+61.99	+58.71	+58.90	yes
Dual vs alignment, coordinated at rate 0.5	+29.96	+30.43	+30.90	yes

Table 6: Qwen3-VL-8B means over three seeds: accuracy without retrieval, with the training bank under concatenation, and with shuffled evidence. Difference columns are accuracy points relative to the no-retrieval base and are computed from the unrounded seed means. Harm is the signed accuracy change under label-adversarial neighbors at rate one. Weibo's absolute accuracies carry the placeholder disclosure of Appendix G.

Dataset	no retrieval	bank	shuffled	bank − none	shuffled − none	harm
Twitter	0.676	0.662	0.678	-1.34	+0.24	+17.90
Weibo	0.900	0.895	0.897	-0.52	-0.27	+2.76
NewsCLIPpings	0.754	0.758	0.758	+0.41	+0.37	+0.60
VERITE-OOC	0.715	0.733	0.720	+1.76	+0.41	-0.20
VERITE-MC	0.591	0.591	0.582	+0.00	-0.87	+1.08
MOCHEG	0.709	0.712	0.705	+0.34	-0.34	+0.53

The full-batch swap. The full-batch swap asks whether a gate can detect that the evidence set is globally foreign, rather than whether predictions depend on evidence belonging to each claim. This distinction is visible on MOCHEG gold: relevance trust has a row-level EG of only +0.7 but a full-batch effect of +5.2, and dual trust similarly gives +0.6 versus +3.8 (Table 7). The contrast is consistent with the relevance statistic suppressing the residual when an entire batch is recognisably foreign, which does not imply stronger claim-specific evidence use. On the AVeriTeC reader-conditioned scorer, the corresponding row-level and full-batch measurements are +14.4 and +13.2 under alignment and +12.7 and +9.9 under dual trust.

The full integration comparison. Then, we test whether combining prior-work safeguards recovers clean gain without reopening the failure modes of the stacked residual. Individual components improve one side of the trade-off but do not solve both: pseudo-label scaling reaches RG +4.3 with a worst case of -3.9, while the assembled baseline reduces the worst case to -0.5 but retains only +0.8 RG (Table 9). Dual trust matches the assembled baseline's worst case at -0.5 while retaining +6.7 RG. The comparison therefore separates DREF from a simple accumulation of existing safeguards.

What the audit does to a benchmark table. Finally, we ask whether the ranking of a retrieval design survives controls on what enters its bank. On the same 1,091 event-disjoint Twitter posts, the published concatenation design is 1.4 points below its no-retrieval variant, and shuffled neighbors are essentially tied with that base (Table 8). Excluding same-image neighbors raises accuracy from 0.662 to 0.701, showing that the benchmark conclusion is sensitive to retrieval overlap. The table therefore motivates auditing the source of a retrieval gain rather than reading the leaderboard difference alone as evidence use.

Table 7: MOCHEG repair ablation, medians over twelve backbones, in accuracy points. RG is measured against the frozen base without evidence, and EG uses the row-level interleaved swap. Full-batch swap is reported separately as a batch-level diagnostic and is not interpreted as EG. Shuffled is the signed accuracy change from the same no-evidence base when all evidence is replaced by unrelated passages.

Configuration	RG	EG	Full-batch swap	Shuffled
Gold fact-check passages				
Residual only, no alignment	-7.3	-0.4	-1.1	-16.4
Alignment, no trust	+7.3	+1.7	+0.7	-8.6
+ relevance trust	7.2	+0.7	+5.2	-0.3
+ agreement trust	+6.7	+1.2	+0.0	-0.1
DREF (dual trust)	+6.7	+0.6	+3.8	-0.0
Passages retrieved from the pool				
Alignment, no trust	+2.0	+0.1	-0.7	-7.0
DREF (dual trust)	+2.6	0.2	+1.5	+0.0

Table 8: Twitter event-disjoint test split: accuracy and macro-F1 on the same 1,091 posts, reported as means over three seeds. Modular-detector rows use frozen Qwen3-VL-8B heads unless stated; the final block re-trains published architectures with their own encoders on the same split.

Method	Acc	Macro-F1
Text-based head, no retrieval	0.710	0.705
Text and image heads, no retrieval	0.683	0.683
Modular detector without retrieval (the frozen base)	0.676	0.674
Modular detector, concatenation (the published design)	0.662	0.662
Modular detector, FiLM	0.645	0.642
Modular detector, neighbor tokens	0.641	0.635
Modular detector, label vote	0.642	0.641
Modular detector, late gate	0.648	0.648
Modular detector, shuffled neighbors (control)	0.678	0.678
Modular detector, concatenation, same-image neighbors excluded	0.701	0.698

C GainMasque repair

C.1 Base-referenced alignment

Base-referenced alignment re-estimates one quantity at deployment: the additive offset of the evidence score relative to the frozen base's verdicts. The following states what this removes under an idealized shift.

Proposition 1. *Suppose the evidence scores in deployment batch B are a translation of their calibration distribution,*

$$e^B = e^{\mathcal{C}} + \eta,$$

where η is constant across units and classes, and the joint law of $(e^{\mathcal{C}}, \hat{y}^{(0)})$ is otherwise unchanged. Let

$$\delta_X = \arg\min_{\delta} \mathbb{E}\left[\mathrm{BCE}\left(\sigma(e^X + \delta), \hat{y}^{(0)}\right)\right], \qquad X \in \{\mathcal{C}, B\}.$$

Then

$$\delta_B = \delta_{\mathcal{C}} - \eta,$$

and the aligned deployment score

$$\tilde{e}^B = e^B + (\delta_B - \delta_{\mathcal{C}})$$

equals $e^{\mathcal{C}}$ exactly.

Proof sketch. For every δ,

$$\mathrm{BCE}\left(\sigma(e^B + \delta), \hat{y}^{(0)}\right) = \mathrm{BCE}\left(\sigma(e^{\mathcal{C}} + (\delta + \eta)), \hat{y}^{(0)}\right).$$

Table 9: Full component comparison behind Table 10, on MOCHEG gold passages, with medians over twelve backbones. Intermediate rows evaluate individual prior-work components from the same stacked residual and are not cumulative unless explicitly named as assembled. Worst case is the most negative accuracy difference from the no-evidence base over the eight corruption families.

Integration	RG	EG	Worst case
Evidence scorer alone, no frozen base	-18.2	+1.1	-69.8
Stacking on the frozen base	-7.3	-0.4	-54.6
Stacking + batch calibration	-2.5	+0.8	-65.2
Stacking + prior adaptation	-24.3	-0.4	-52.9
Stacking + pseudo-label scaling	+4.3	+0.2	-3.9
Stacking + held-out admission	-2.3	+0.9	-66.8
Stacking + relevance gate	-1.5	-2.4	-19.5
Stacking + agreement statistic	+1.1	+0.3	-2.5
Assembled baseline (admission, gate and agreement)	+0.8	+0.2	-0.5
Assembled baseline with pseudo-label scaling	+2.5	+1.3	-1.7
DREF, Rule 1 only (alignment)	+7.3	+1.7	-60.2
DREF, Rules 1 and 2 (dual trust)	+6.7	+0.6	-0.5

Table 10: The repair against representative baselines built from prior-work components, on MOCHEG gold passages, with medians over twelve backbones. Worst case is the most negative accuracy difference from the no-evidence base over the eight corruption families. Appendix B reports six additional configurations.

Integration	RG	EG	Worst case
Evidence scorer alone, no frozen base	-18.2	+1.1	-69.8
Stacking on the frozen base	-7.3	-0.4	-54.6
+ pseudo-label scaling	+4.3	+0.2	-3.9
Assembled baseline (admission, gate and agreement)	+0.8	+0.2	-0.5
DREF, Rule 1 only (alignment)	+7.3	+1.7	-60.2
DREF, Rules 1 and 2 (dual trust)	+6.7	+0.6	-0.5

Thus, the deployment objective is the calibration objective evaluated at $\delta + \eta$, giving $\delta_B = \delta_{\mathcal{C}} - \eta$. Substitution yields $\tilde{e}^B = e^{\mathcal{C}}$. The same frozen-base verdicts $\hat{y}^{(0)}$ enter both objectives and need not equal the ground-truth labels. □

This is a population identity; on a finite deployment batch, estimating δ_B introduces sampling error of order $|\mathcal{B}_E|^{-1/2}$. The result motivates referencing the same frozen prediction rule at calibration and deployment: under a pure translation, the shift is absorbed by $\delta_B - \delta_{\mathcal{C}}$.

Why only the alignment offset. The label-fitted residual scale β remains fixed at deployment. Re-estimating scale from imperfect frozen-base verdicts attenuates it under symmetric verdict noise, while the offset difference does not require such a refit. This matters empirically: on identical scores and splits, adapting scale together with the offset reduces the 8B gain from +7.87 to +7.05, with the same direction on the 2B backbone. We therefore adapt only the alignment offset and keep the label-fitted residual parameters fixed.

When the translation assumption fails. MOCHEG does not satisfy the proposition's idealized setting: calibration and test rows come from different articles, and both the frozen-base verdict rate and label rate shift between them. Table 11 therefore asks whether base-referenced alignment remains useful outside a pure translation. It recovers the gain almost exactly under additive biases, whereas scale changes and noise are only partly absorbed. A class-dependent perturbation is not a translation and can instead introduce class information, so its large post-alignment gain should not be interpreted as shift correction. Most importantly, alignment alone cannot rescue a sign-inverted scorer, while dual trust rejects that update and returns the frozen base. Thus, alignment handles location shift, whereas trust is needed when the evidence signal itself becomes unreliable.

Table 11: Alignment and its failure modes on MOCHEG gold with the 8B backbone. All entries are accuracy-point changes relative to the frozen no-evidence base, averaged over seeds. Left: alternative correction rules evaluated with the same scorer and split. Right: unaligned, base-referenced alignment, and full DREF under synthetic scorer distortions.

Correction	Δacc.	Distortion	unaligned	alignment	full DREF
No alignment	-8.61	bias -2	+7.8	+7.9	+7.4
Mean centring	-3.56	bias +2	-24.0	+7.9	+7.4
Median centring	+2.31	scale × 0.5	-5.6	+8.1	+5.3
Quantile matching	-2.40	scale × 2	-8.8	+4.3	+7.1
Analytic prior correction	-10.36	noise, s.d. 1	-7.9	+6.7	+5.6
Platt scaling on calibration	-14.20	class-dependent bias	-7.5	+19.9	+18.0
Base-referenced alignment	+7.87	sign inverted	-24.8	-26.6	+0.0
+ scale	+7.05				
+ dual trust	+7.40				
Oracle intercept, label-fitted	+10.71				

C.2 Constants, admission and the algorithm

The evidence scorer. On the training units with evidence, the pooled evidence feature is standardised, reduced to 128 principal components without using labels, and passed to an ℓ_2-regularised logistic regression. Its regularisation strength is selected by an inner five-fold log-loss search over $\{0.003,0.01,0.03,0.1,0.3,1\}$. The log-odds define the evidence score e_i after an analytic prior adjustment

$$\operatorname{logit} \pi_{\mathrm{cal}} - \operatorname{logit} \pi_{\mathrm{ev}},$$

where π_{ev} is the class prior among the scorer's training units; it is 0.73 on MOCHEG because evidence availability is label-skewed. The evidence feature is the mean of the available passage features, and units without evidence contribute no residual through the availability indicator in Eq. (11).

Admission. For each source, backbone, and seed, we cross-fit the residual procedure in five folds on the calibration split. For calibration unit i, let $\Delta_i^{(-f)}$ denote the out-of-fold residual, and define its reduction in negative log-likelihood as

$$d_i = \ell\Big(s_i^{(0)}, y_i\Big) - \ell\Big(s_i^{(0)} + \Delta_i^{(-f)}, y_i\Big).$$

A source is included in $\mathcal{S}^+$ when the 2.5% quantile of 2,000 bootstrap resample means of d_i is positive. The MOCHEG gold source passes this test on all three seeds, with lower bounds between +0.063 and +0.077. After admission, (β, b) are fitted on all calibration units; a rejected source contributes no residual, so deployment reduces to the frozen base.

Dual trust in full. The agreement branch compares aligned evidence scores with the frozen base's confident verdicts,

$$\tau_B^{(a)} = \Pi_{[0,1]}\left(\frac{\mathcal{G}(B)}{\mathcal{G}(\mathcal{C})}\right),$$

where $\mathcal{G}(B)$ is the corresponding agreement slope on the confident subset of evidence-bearing units. The relevance branch is

$$\tau_B^{(r)} = \Pi_{[0,1]}\left(\frac{\mathcal{R}(B) - \mathcal{R}^-}{\mathcal{R}^+ - \mathcal{R}^-}\right)\kappa(B),$$

where $\mathcal{R}(B)$ is mean claim–evidence relevance and $\mathcal{R}^+$ and $\mathcal{R}^-$ are its calibration means under matched and shuffled evidence. The shape factor is

$$\kappa(B) = \Pi_{[0,1]}\left(1 - \frac{\mathrm{KS}(B) - 0.15}{0.35}\right),$$

with $\mathrm{KS}(B)$ the two-sample Kolmogorov–Smirnov statistic between $|\tilde{e}_i|$ on the evidence-bearing units of B and $|e_i|$ on the evidence-bearing calibration units. The final trust coefficient is

$$\tau_B = \min\left\{\tau_B^{(a)}, \tau_B^{(r)}\right\}.$$

Algorithm 1 DREF, per source, backbone, and seed

Calibration
fit the evidence scorer on the training portion and compute e_i on calibration split $\mathcal{C}$
cross-fit the residual procedure on $\mathcal{C}$ and decide whether the source belongs to $\mathcal{S}^+$
if $S \in \mathcal{S}^+$ **then** fit (β, b) on all calibration units
compute $\delta_{\mathcal{C}}$ against frozen-base verdicts $\hat{y}_i^{(0)}$
compute $\mathcal{G}(\mathcal{C})$, $\mathcal{R}^+$, and $\mathcal{R}^-$; store calibration $|e_i|$
Deployment on batch B
$\mathcal{B}_E \leftarrow \{i \in B : E_i \neq \varnothing\}$
if $S \notin \mathcal{S}^+$ or $|\mathcal{B}_E| < 20$ **then return** $s_i^{(0)}$
compute $\delta_B = \arg\min_{\delta} \sum_{i \in \mathcal{B}_E} \mathrm{BCE}\big(\sigma(e_i + \delta), \hat{y}_i^{(0)}\big)$
$\tilde{e}_i \leftarrow e_i + (\delta_B - \delta_{\mathcal{C}})$ ▷ base-referenced alignment
compute $\tau_B^{(a)}$ and $\tau_B^{(r)}$
$\tau_B \leftarrow \min\{\tau_B^{(a)}, \tau_B^{(r)}\}$ ▷ dual trust
$\Delta_i \leftarrow \beta \tilde{e}_i + b$
return $s_i^\star = s_i^{(0)} + \mathbb{I}[E_i \neq \varnothing]\tau_B \Delta_i$

The two branches therefore form a conservative conjunction: relevance responds to poor claim–evidence matching, whereas agreement responds to updates that conflict with the frozen base's confident verdicts.

Constants. All constants are declared before evaluation and recorded in the settings manifest; none is tuned on test labels. Base-referenced alignment requires at least 20 evidence-bearing units in batch B. The confident subset contains evidence-bearing units whose $|s_i^{(0)}|$ is at or above the median $|s_i^{(0)}|$ of the full batch; estimating $\mathcal{G}(B)$ requires at least 10 such units and uses a small ridge penalty. If either requirement fails, the corresponding evidence update is disabled. The shape factor uses a Kolmogorov–Smirnov threshold of 0.15 and width 0.35, and the pair verifier uses regularisation strength 0.1. The deployment stream warms up for 32 units before batch statistics are applied. On 8B MOCHEG, $\mathcal{R}^+$ ranges from 0.477 to 0.490 and $\mathcal{R}^-$ from 0.085 to 0.098 across seeds.

The pair verifier. The verifier is trained on the training portion using principal components for the claim and evidence representations and a least-squares map from evidence space to claim space fitted on matched pairs. An ℓ_2-regularised logistic regression then scores the elementwise product, absolute difference, cosine similarity, and centred cosine similarity of the mapped pair. Each matched pair is trained against one lexically matched hard negative and one random negative.

Algorithm. Algorithm 1 summarizes calibration and deployment. Calibration uses labels to decide admission and fit the residual parameters; deployment re-estimates only batch-level alignment and trust quantities and uses no ground-truth labels.

Batch size and warm-up. We next investigate how much batch context the test-time statistics require. On MOCHEG with the 8B backbone, the gain is zero below the minimum batch size, reaches +4.7 points at 32 units, and approaches the full-split gain by 256 units. A causal growing-prefix evaluation after the 32-unit warm-up reaches +7.3, close to the full-split result. This experiment uses the agreement branch alone and therefore characterizes the batch requirement of $\tau_B^{(a)}$, not the full dual-trust rule.

Observed trust coefficients. On clean MOCHEG gold, the relevance branch remains near its calibration reference, with $\tau_B^{(r)}$ between 0.97 and 1.00, while agreement is more sensitive to the frozen base. Across the tested backbones, $\tau_B^{(a)}$ ranges from 0.55 to 0.87 except for Qwen3-VL-2B, where it closes entirely. In this grid, that transition tracks confident-base reliability rather than model

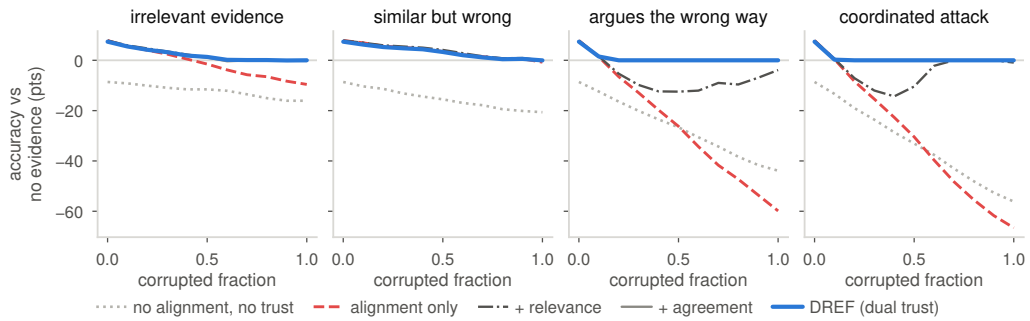


Figure 4: Accuracy against the no-evidence base on MOCHEG gold passages (Qwen3-VL-8B, seed means) as the corrupted fraction grows, for four of the eight corruption families. Curves compare the unprotected configurations, the two single trust rules, and DREF.

size: the gate closes at confident-half precision 0.671 and remains active at 0.710. We therefore treat agreement as a reliability-dependent safeguard rather than as a fixed function of backbone scale.

Scope of the implementation. The full dual-trust rule is used on MOCHEG and AVeriTeC. The NewsCLIPpings live and corpus experiments and the Twitter archive experiment use the agreement-only configuration,

$$\tau_B = \tau_B^{(a)}.$$

Residual clipping is implemented as an optional component but is not used in the reported repair; on retrieved passages it changes accuracy by at most 0.3 points in either direction.

D THE CORRUPTION SUITE AND ROBUSTNESS EVIDENCE

The eight families. Unless otherwise stated, every tabulated worst-case value in the paper minimizes over the following eight corruptions of a unit's evidence set, at rates 0.25, 0.5, 0.75 and 1.0. *Shuffled* replaces the set with another unit's, chosen by a fixed hash, with no control on label or length. *Length-matched shuffled* replaces it with the set of the unit whose total passage length is closest, holding surface length fixed. *Least similar* serves the passages in the pool least similar to the claim. *Deletion* empties the set, after which a method that keeps the base on units without evidence returns the base. *Similarity-matched wrong* serves the three most similar foreign passages, making the replacement similar to the claim without preserving its own evidence. *Label-adversarial stance* takes the twenty nearest foreign passages and serves the three whose scorer log-odds most oppose the true label. *Coordinated* does the same from a pool of a hundred, giving the attacker five times the budget. *Mixed, gradual* assigns each corrupted unit one of six of these families by a hash of its position. The corrupted units are chosen by one deterministic mask shared across families, methods and backbones, and the rate grids are nested, and thus a unit corrupted at rate 0.25 remains corrupted at every higher rate.

We keep three related evaluation sets separate. The eight families above define the registered corruption suite. The same-label swap is instead the audit's instrument and is not included in the registered worst case. Figures 4 and 5 trace the eight corruption families on the finer eleven-rate grid from 0 to 1. Figures 6 and 7 use the ledger's expanded diagnostic sweep, which contains the same-label diagnostic conditions; their horizontal coordinate is therefore the largest observed loss rather than the registered worst case. The leave-one-family-out study below uses a third set, which includes the swap and omits the length-matched shuffle.

The other four families. Do the remaining corruption families change the pattern in Figure 4? Figure 5 shows that they do not: alignment without trust again becomes harmful as corruption grows, while the dual rule remains near the no-evidence base. Deletion provides a useful boundary case, because every rung returns to the base once no evidence remains. Under the length-matched shuffle, alignment alone ends at -0.03 while the trusted variants remain slightly above the base,

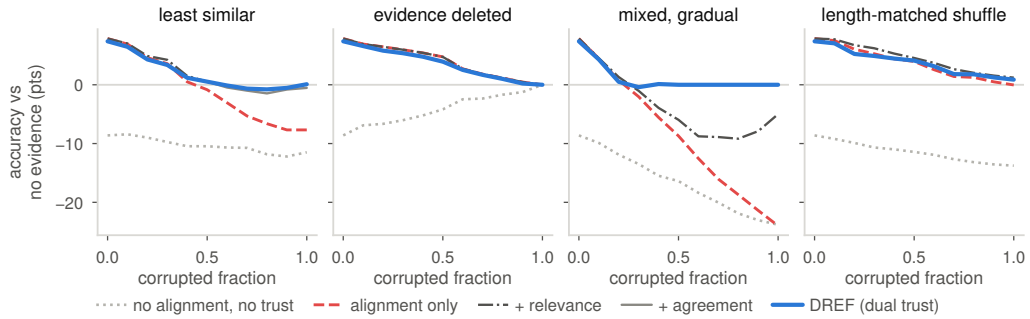


Figure 5: The four additional corruption families on MOCHEG gold passages (Qwen3-VL-8B, seed means), plotted as in Figure 4.

indicating that the alignment-based gain does not survive a length-matched replacement. Across the eight corruption families and the same-label diagnostic on this backbone, the lowest observed dual-trust point is -0.78.

Detecting a corrupted batch. We next ask whether the trust statistic itself signals that the evidence batch has changed. On batches of 128 units, it separates clean from fully corrupted evidence with AUROC 0.95−0.99 for shuffled, label-adversarial, and coordinated evidence, and 0.72−0.95 for similarity-matched wrong evidence. The latter is harder because the replacement passages remain similar to the claim. Qwen3-VL-2B is the exception: its AUROC is at or below chance on three of the five measured conditions because its agreement factor is identically zero. The statistic is therefore informative when the underlying trust signal is active, but it does not provide an independent detector when that signal degenerates.

Leaving a family out. Can selecting the trust rule on observed corruptions improve robustness to an unseen one? We choose among 27 configurations using seven families and evaluate on the held-out eighth. On Qwen3-VL-8B, tuning slightly improves the average held-out value, +1.09 versus +0.85 for the predeclared dual rule, but its worst held-out value is lower, -0.66 versus -0.28. The difference is larger on Qwen3-VL-2B: selection chooses relevance alone and reaches -10.77 on label-adversarial evidence, whereas the predeclared dual rule remains at +0.63. Selecting a rule from the observed families therefore does not improve the worst held-out failure and can expose a failure mode that the fixed dual rule avoids.

Across backbones. Does the protection persist beyond the 8B model? Table 12 shows the same contrast across all twelve backbones. With alignment alone, every backbone reaches a worst accuracy change between -56.1 and -66.6 points; with dual trust, the range contracts to -0.1 to -1.2. Clean retrieval gain also remains positive on eleven of the twelve backbones. The exception is Qwen3-VL-2B, where the agreement factor closes the dual gate and the clean retrieval gain falls to zero.

The repair as a trade. Figure 6 asks how much of the clean evidence benefit is retained when this protection is applied across backbones. From alignment alone to dual trust, the median largest observed loss falls from 60.2 to 0.5 points, whereas median row-level evidence gain moves from +1.7 to +0.6. The robustness change is consistent across all twelve backbones, while the evidence-gain change is not: evidence gain falls on seven and rises on five, with a median paired change of -0.4 points. Thus, the large reduction in observed harm does not correspond to a uniform loss of evidence gain across models.

Where the protection enters. Figure 7 isolates which rung of the repair produces this change. Alignment moves median evidence gain from -1.1 to +1.7, but does not protect against corrupted evidence: the largest observed loss remains 60.2 points. Relevance trust reduces that loss to 13.3, whereas agreement trust reduces it to 0.8 while retaining +1.2 points of evidence gain. Combining

Table 12: MOCHEG, per backbone. Clean retrieval gain (RG) on gold passages and worst accuracy change over the eight corruption families at rates 0.25−1.0, under alignment alone and under dual trust; for dual trust also the family attaining the worst case, the agreement trust τ on the clean gold batch, and RG with the lexical top-3 retrieved passages. Values are accuracy points against each backbone's no-evidence base; seed means. τ compares the scorer's agreement with the base's confident verdicts on the batch to the same statistic on the calibration split and is re-estimated on every batch. least sim.: least-similar passages; adv. stance: label-adversarial stance.

	Alignment only		Dual trust				
Backbone	RG	worst	RG	worst	worst family	τ	RG (retr.)
Qwen3-VL-2B	+14.1	-61.2	0.0	-0.2	least sim.	0.00	+4.3
Qwen3-VL-4B	+8.0	-63.5	+7.8	-0.2	shuffled	0.82	+3.0
Qwen3-VL-8B	+7.9	-66.6	+7.4	-0.6	least sim.	0.76	+2.5
Qwen3-VL-32B	+6.7	-63.8	+6.6	-0.7	least sim.	0.73	+1.2
Qwen2.5-VL-7B	+6.3	-58.5	+5.6	-1.2	adv. stance	0.87	+1.4
Qwen2.5-VL-72B	+4.2	-56.1	+4.1	-0.4	adv. stance	0.73	+1.7
InternVL3-38B	+6.2	-59.2	+6.0	-0.7	adv. stance	0.80	+2.8
InternVL3.5-8B	+6.6	-62.0	+7.1	-0.1	shuffled	0.63	+2.1
Qwen3.5-2B	+11.2	-58.0	+11.1	-0.8	shuffled	0.79	+5.7
Qwen3.5-4B	+9.1	-58.5	+6.8	-0.5	shuffled	0.60	+2.7
Qwen3.5-9B	+7.0	-56.7	+5.1	-1.1	shuffled	0.55	+3.2
Gemma-4-12B	+7.6	-65.4	+7.6	-0.4	least sim.	0.79	+1.7

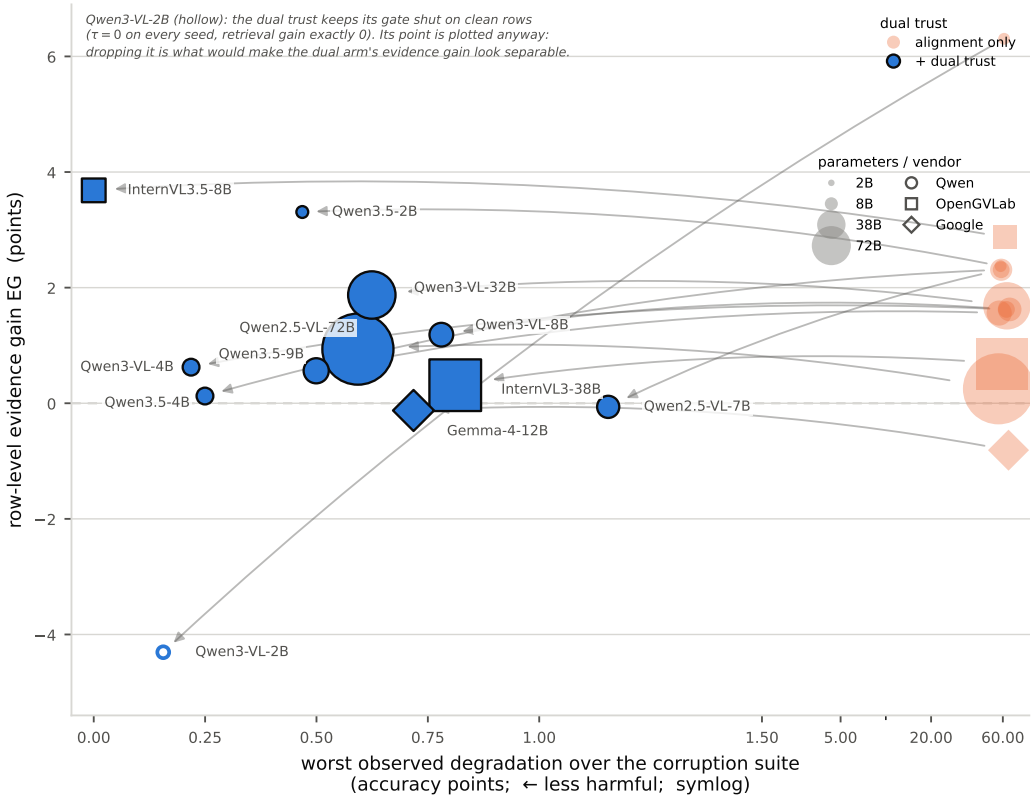


Figure 6: Row-level evidence gain against the largest observed loss over the expanded diagnostic sweep for twelve frozen backbones on MOCHEG gold passages. Each backbone is shown with alignment only and with dual trust, joined by an arrow. Bubble area denotes nominal parameter count and marker shape denotes vendor.

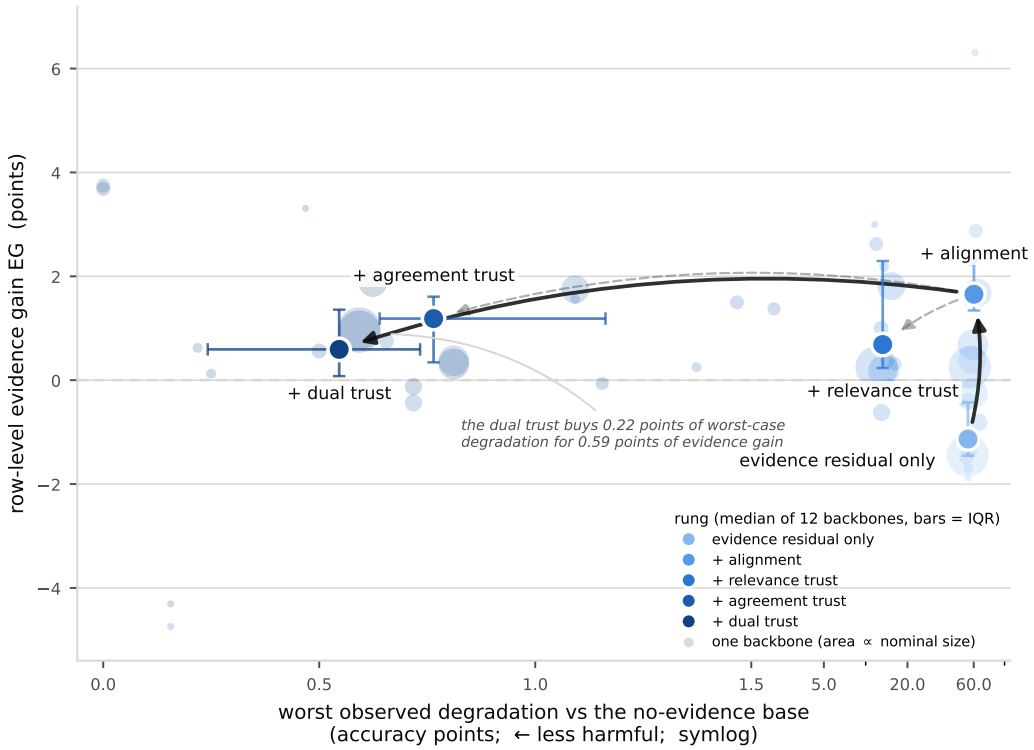


Figure 7: The repair ladder on MOCHEG gold passages, using the axes of Figure 6. Relevance and agreement are sibling trust rules applied to the aligned residual, and DREF combines them by the pointwise minimum of their trust values. Markers show medians over twelve backbones, bars the interquartile range, and faint dots individual backbones.

Table 13: Clean retrieval gain and worst accuracy change (RG / worst, accuracy points) by trust rule. MOCHEG reports the median over twelve backbones; the AVeriTeC columns report the two readers; the final column isolates Qwen3-VL-2B.

Configuration	MOCHEG	AVeriTeC (Qwen)	AVeriTeC (Idefics3)	MOCHEG (Qwen3-VL-2B)
Alignment, no trust	+7.3 / -60.2	+11.7 / -2.1	+10.7 / -1.5	+14.1 / -61.2
+ relevance trust	+7.2 / -12.6	+10.7 / -0.6	+10.4 / -0.5	+14.1 / -13.6
+ agreement trust	+6.7 / -0.7	+10.5 / -2.1	+9.6 / -1.6	+0.0 / -0.2
DREF (dual trust)	+6.7 / -0.5	+10.4 / -0.6	+9.6 / -0.5	+0.0 / -0.2

the two as dual trust reduces the remaining loss to 0.5, with evidence gain of +0.6. On MOCHEG, agreement therefore supplies most of the protection; the next comparison tests whether the same single rule remains sufficient elsewhere.

Which trust each dataset needs. Table 13 shows that the stronger single trust rule changes with the setting. On MOCHEG, agreement alone reaches a worst accuracy change of -0.7, whereas relevance alone leaves -12.6; on AVeriTeC the order reverses, with relevance reaching -0.6 and -0.5 on the two readers while agreement leaves -2.1 and -1.6. Across these settings, dual trust matches or improves the better single-rule worst case. It supports using the two rules together rather than choosing one after observing the dataset. Qwen3-VL-2B marks the boundary: its agreement factor is zero, so the dual rule closes the gate and loses the clean retrieval gain.

The adaptive attacker. Finally, we explore whether an attacker that uses each system's own passage judgments can produce larger failures than the fixed corruption families. For each system, the attacker selects three passages from its nearest foreign candidate pool whose single-passage verdicts

Table 14: AVeriTeC, seed-1 served evidence sets, accuracy in percent. Gain is Gold minus No ev.; the two harm columns report accuracy-point change relative to each system's no-evidence accuracy. "Worst @ 100%" is the minimum across the eight corruption families at full corruption, excluding the same-label swap.

System	No ev.	Gold	Gain	Worst @ 100%	Attacker
DREF, reader-conditioned (Qwen)	78.6	89.0	+10.4	-0.6	-3.4
DREF, reader-conditioned (Idefics3)	77.8	87.3	+9.6	-0.4	-4.0
Search-R1	75.0	88.9	+13.9	-2.5	-7.7
ReSearch	75.0	87.7	+12.7	-2.2	-15.4
R1-Searcher	71.9	84.0	+12.0	0.0	-7.7
DeepResearcher	72.2	89.5	+17.3	0.0	-10.5
CRAG	74.1	84.3	+10.2	-1.9	-10.8
Self-RAG, always retrieve	57.7	60.2	+2.5	-6.8	-7.1

most oppose the true label, using that system's own prompt format. Under this attack, the two DREF readers lose 3.4 and 4.0 points relative to their no-evidence bases, compared with $7.1-15.4$ points for the six search agents in Table 14. The adaptive attack is therefore stronger than the registered family worst for DREF, but the resulting loss remains smaller in these experiments than for the external agents.

E EVIDENCE SOURCES AND FUSION DESIGNS

The audit treats the evidence source as the object of study and compares five sources with a no-retrieval control. The sources differ in claim specificity, whether their content could have been written after fact-checking, and coverage, but all enter the detector through the same evidence slot, separating source differences from architecture. How the evidence is fused is a separate design choice, and the design grid therefore evaluates both accuracy and harm from wrong evidence.

Training bank. The bank contains one multimodal feature for every unit in the seed's training split. Each unit queries the bank with its ℓ_2-normalised multimodal feature; cosine retrieval returns $k = 5$ neighbors, whose features are mean pooled. The query's identifier is masked, and the same-image-excluded arm of Section 4.2 additionally removes entries sharing its caption text or image file.

Eight fusion designs. The audit grid varies how the pooled neighbor information reaches the evidence head. *Concatenation*, used for the main results, concatenates the unit's multimodal feature with the mean-pooled neighbor feature and maps the result through a two-layer perceptron, whereas *FiLM* converts the pooled neighbor feature into a scale and shift applied to the multimodal feature. *Tokens* attends to neighbors individually, representing each by its feature, label embedding, and similarity; *Vote* discards the features and retains only label and similarity statistics. *Late* adds a gated label-vote logit at the output level, while *Single* passes the content, multimodal, evidence, and neighbor features through one perceptron. Two controls complete the grid: *shuffled evidence* assigns each unit another unit's neighbors at both training and test time, and *no retrieval* removes the branch.

Gold evidence. For MOCHEG, gold evidence comprises up to three fact-check passages related to the claim, with HTML removed, verdict words scrubbed, and each passage truncated to 1,200 characters before encoding as in Appendix K. For AVeriTeC, it comprises up to three claim-specific question–answer snippets encoded by an independent text encoder, BGE-M3, rather than by the backbone. The claim-conditioned pair scorer of Section 4.3 reads these same embeddings.

Retrieved evidence. For MOCHEG, the retrieval pool contains 11,414 gold passages from the training and test claims, with the source claim recorded recorded: 8,955 come from training claims and 2,459 from test claims. TF-IDF retrieval uses sublinear term frequency, minimum document frequency 2, unigrams and bigrams, and at most 300,000 features; it is fitted over the passages and claims, queries with claim text, and returns the top three passages. Test passages are deliberately

retained because retrieval is label-free, while a claim's passages are masked in the foreign-evidence conditions; Appendix A reports the corresponding split audit. For AVeriTeC, the pool contains every claim's gold snippets and retrieval is dense in the encoder space.

Live source. For NewsCLIPpings, we queried Google Cloud Vision (`WEB_DETECTION`) for every unique corpus image, using 28,827 images within one window on a single day; 17,888 returned at least one matching page. Each row receives at most three page titles from full or partial image matches in provider rank order, with the corresponding domain. The caption is never used in the query, and it cannot directly provide label information to the source. Images with no matching page receive no evidence, yielding coverage of 0.608 over the 7,264 test rows, or 39.2% without evidence; coverage is balanced across classes, 0.606 versus 0.610. A domain-based leakage filter removes 84 of 195,788 matched pages, and disabling it changes the result by at most 0.01 points. Re-querying 498 images preserves match status for every image and the top-ranked page for 98%.

Our live-source results explores two row sets. The headline gain of +3.4 is computed over 7,264 test rows, whereas the content ladder uses the 7,016 protocol-primary rows after excluding pilot images. On these rows, page titles alone yield +3.47 [+2.80,+4.16], adding the page description yields +3.71 [+3.01,+4.42], and additionally using the first part of the page body yields +4.23 [+3.52,+4.94]. The gain is not recovered from the evidence without the claim: page-text-based and domain-based scorers reach 0.508 and 0.494 and are admitted on no seed, whereas the caption–evidence pair scorer reaches 0.748 and is admitted on all three.

Corpus source. The gated design of Section 4.2 uses an offline source built from genuine image–caption pairs in the NewsCLIPpings training corpus. One construction retrieves corpus captions most similar to the query caption and supplies their associated *photographs*; the mirror construction retrieves photographs by visual similarity and supplies their captions, serving as an offline analogue of reverse-image search. On NewsCLIPpings, neither construction carries detectable label signal: evidence-based probes reach 0.500 and 0.502. It is dataset-specific rather than intrinsic to the construction, since the same corpus evidence reaches 0.617 on VERITE-MC.

F CLOSED EXTERNAL READERS

This appendix reports the full closed-reader results summarized in Section 4.5. Appendix J separately records the decision-native reader's interface, freeze, and query protocol.

Closed-reader results. All analyses use the 324 frozen AVeriTeC development claims, except the reversed-direction condition, which applies to the 136 claims for which a reversible factual unit is available. Same-label swaps use the five frozen partner draws. Worst degradation is the minimum change from the claim-based reference across the evaluated evidence conditions, with each condition compared on its matching claim set. In Table 4, $\overline{\mathrm{NLL}}$ is the mean of the three condition-level NLLs. The closed readers are zero-shot; fitted residual, scorer, and admission components use only the designated training or calibration data.

How the decision-native reader enters the pipeline. The two Jev experiments use the reader output in different ways. In Table 4, Jev's log-odds serve directly as the evidence score in a calibration-fitted residual gated by the preregistered admission test. This configuration does not apply base-referenced alignment or dual trust. In Table 15, Jev outputs instead augment the features of the encoder-based evidence scorer, either through the relevance mass $1 - p(\text{not enough information})$ or through the full verdict distribution; the reader-conditioned scorer is evaluated separately. These experiments therefore test two different uses of the same external reader rather than ablations of a single system.

Admission is not equivalent to an abstention threshold. A reader with an explicit "not enough information" option suggests a simple alternative: fall back to the claim-based score when that probability exceeds a threshold fitted on held-out training data. This heuristic retains more original-evidence accuracy than the admission-gated residual, 0.940 versus 0.901, but is worse by 8.82 points under reversed-direction evidence and by the same amount on worst degradation; both paired intervals exclude zero. The two are not separable under same-label swap. Against an unweighted sum of

Table 15: External-reader features for DREF's evidence scorer. All systems use the preregistered admission path; base-referenced alignment and dual trust are not applied. In the two Jev-augmented variants, Jev outputs are appended to the encoder-based scorer, while the reader-conditioned scorer is evaluated separately. Columns and evaluation conventions follow Table 4. In Table 4, by contrast, Jev's log-odds serve directly as the evidence score.

System	E_i	$E_{\pi(i)}$	$\mathcal{R}(E_i)$	$\Delta_{\min}$	$\mathrm{NLL}(E_i)$	$\overline{\mathrm{NLL}}$
DREF, encoder scorer	0.792 ± 0.010	0.791 ± 0.009	0.814 ± 0.019	+0.000	0.468	0.466
+ Jev relevance	0.792 ± 0.010	0.792 ± 0.010	0.814 ± 0.019	+0.000	0.468	0.466
+ Jev distribution	0.906 ± 0.008	0.803 ± 0.004	0.735 ± 0.007	-0.078	0.254	0.505
DREF, reader-conditioned scorer	0.914 ± 0.005	0.774 ± 0.011	0.738 ± 0.004	-0.076	0.247	0.530

Table 16: **Categorical readers across evidence conditions (AVeriTeC forced-choice arm).** Accuracy is reported in percent on the 324 frozen test claims, except $\mathcal{R}(E_i)$, which applies to 136 claims. The left block reports the categorical readers without DREF; the right block reports the same evidence conditions with DREF. Parentheses report the change relative to the corresponding non-DREF condition, $(\Delta \pm \mathrm{SE})$. Six readers belong to the frozen admitted set; the post-freeze reader was run under the same frozen protocol and gate. The reader of Section 4.5 is not on this axis (Table 4).

Reader	**−DREF**			**+DREF**		
	E_i	$\mathcal{R}(E_i)$	$E_{\pi(i)}$	E_i	$\mathcal{R}(E_i)$	$E_{\pi(i)}$
Admitted categorical readers (frozen set)						
GPT-6 Astra	93.8	65.4	78.6	92.9 (-0.9 ± 0.8)	70.6 $(+5.1 \pm 1.3)$	79.7 $(+1.1 \pm 0.8)$
GPT-5.6 Sol	92.9	65.4	74.7	92.1 (-0.8 ± 0.8)	71.1 $(+5.6 \pm 1.5)$	76.1 $(+1.4 \pm 0.5)$
GPT-5.6 Luna	92.9	72.8	73.6	92.4 (-0.5 ± 1.2)	77.2 $(+4.4 \pm 1.5)$	76.0 $(+2.5 \pm 0.9)$
Claude Fable 5.1	96.0	75.7	77.8	95.0 (-1.0 ± 0.9)	77.9 $(+2.2 \pm 0.7)$	79.6 $(+1.8 \pm 0.8)$
Claude Opus 5	96.0	67.6	76.1	95.0 (-1.0 ± 0.2)	72.3 $(+4.7 \pm 1.1)$	77.8 $(+1.6 \pm 0.7)$
Claude Sonnet 5	93.5	72.1	75.1	92.8 (-0.7 ± 0.6)	76.0 $(+3.9 \pm 0.8)$	77.1 $(+2.0 \pm 1.1)$
Post-freeze check (same protocol and gate)						
GPT-6 Sol[†]	92.9	69.1	74.5	92.1 (-0.8 ± 0.5)	73.0 $(+3.9 \pm 0.4)$	76.6 $(+2.1 \pm 1.2)$

[†]**Post-freeze refresh:** released after the audit froze and run under the same frozen protocol and gate; not part of the ranges quoted in the text. GPT-5.6 Terra, Claude Opus 5.5, and GPT-6 Luna were not admitted by the pre-registered instability screen, with swap-d_0 disagreement of 16%, 17%, and 11%, respectively, above the 10% gate on five repeated calls. They were therefore not run in the forced-choice arm on the test claims and have no DREF-Cat fit; they enter the mechanism arm only, and none was re-run or rescued. Jev (Section 4.5) uses three-option elicitation rather than forced choice and is analyzed separately in Table 4.

the claim-based score and Jev's log-odds, the admission-gated residual has lower original-evidence accuracy but substantially lower mean NLL, while their corrupted-evidence accuracies are not separable. We therefore find a robustness advantage over the abstention heuristic, but do not claim an accuracy advantage over the unweighted residual.

External-reader features introduce a trade-off. Table 15 asks whether Jev provides useful features for DREF's learned evidence scorer under the same admission path. Adding only the relevance mass leaves the encoder-based scorer essentially unchanged. The full verdict distribution raises own-evidence accuracy from 0.792 to 0.906 and reduces own-evidence NLL from 0.468 to 0.254, but reversed-direction accuracy falls from 0.814 to 0.735 and worst degradation moves from 0.000 to -0.078. The verdict distribution therefore supplies useful predictive signal, but the added clean gain does not translate into uniformly more robust evidence use. The encoder-only scorer fails admission on two of three seeds, so its zero worst degradation largely reflects inaction; against the reader-conditioned scorer, the only separable accuracy difference is +2.88 points under same-label swap [+0.72,+5.08].

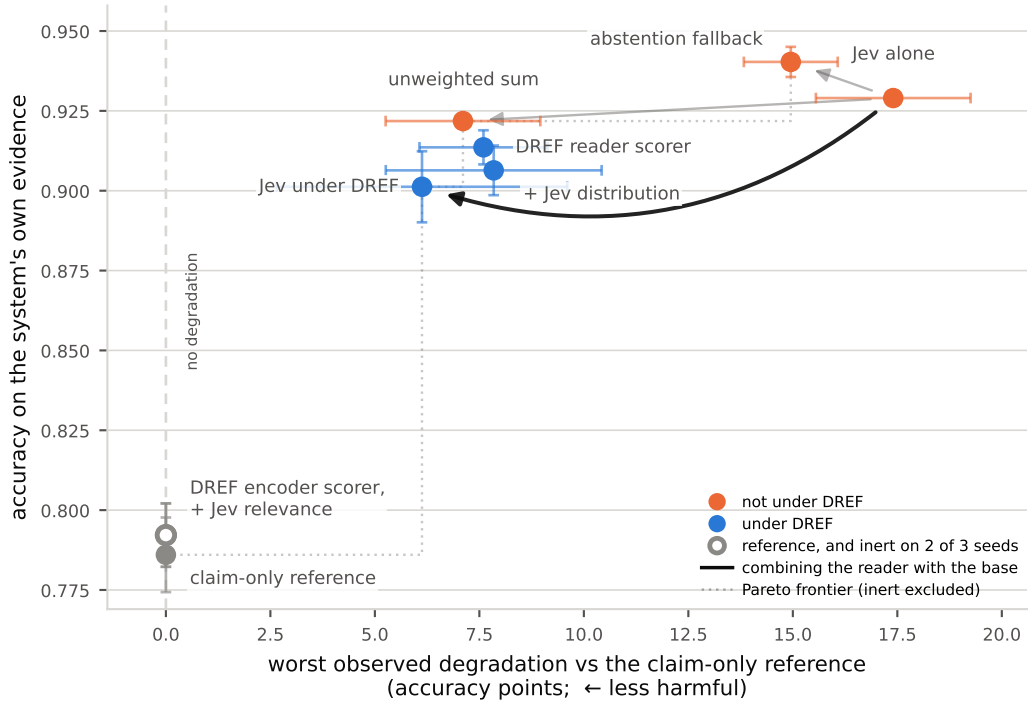


Figure 8: Utility–robustness trade-off for the decision-native reader. The vertical axis is own-evidence accuracy and the horizontal axis is the worst degradation from the claim-only reference across the three evidence conditions. Open marks denote the claim-only reference and configurations that fail admission on two of three seeds; the heavier arrow marks the admission-gated residual.

Forced choice exposes sensitivity to swapped evidence. Table 16 removes the abstention option while keeping the claims and evidence fixed. The six admitted readers achieve 92.9−96.0% accuracy on their evidence, yet same-label evidence from another claim reduces accuracy by 15.2−19.9 points for every reader, with all clustered paired bootstrap intervals excluding zero. Thus, strong original-evidence accuracy does not imply that the decision is stable to evidence replacement. Because the forced-choice and abstention-enabled arms are different elicitations, we use their contrast only qualitatively rather than attributing the loss to abstention itself.

High closed-book accuracy limits what original-evidence accuracy implies. Without supplied evidence, the six readers already answer the forced-choice question correctly on 78.7−92.3% of claims, while supplying each reader's evidence adds only 1.5−14.8 points. Original-evidence accuracy therefore cannot be attributed to the supplied evidence alone. At the same time, every reader performs worse with a same-label partner's evidence than with no evidence, by 3.6−13.7 points on the same claims, showing that corrupted evidence can redirect an otherwise correct closed-book decision. Because the closed-book condition cannot distinguish memorization, world knowledge, and dataset priors, we use it only as a reference decision rather than as a measure of what the reader "knows." All AVeriTeC claims predate the readers' training cutoffs, 72.8% of this cohort is refuted, and no contamination control is available.

The categorical admission-path adaptation. The frontier readers expose only categorical verdicts, so there is no continuous evidence score on which to apply the alignment used elsewhere. We map each forced-choice verdict to $z \in \{-1, +1\}$ and fit its magnitude as a calibration residual $\beta z + b$, gated by the same admission test used for the decision-native reader. The fit uses the same 300 calibration claims per seed. As a zero-cost check, sign-binarising the cached Qwen and Jev scores leaves both admitted on all three seeds. Once admission has passed, the gated and ungated categorical residuals coincide, so Table **??** reports one form. A separate one-dimensional logis-

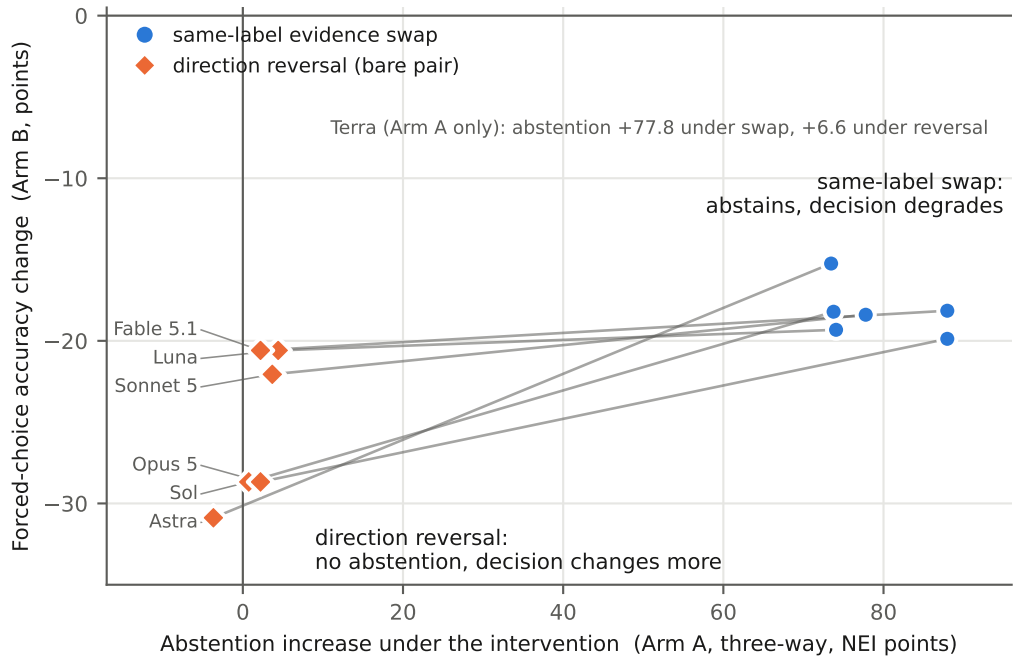


Figure 9: Abstention and decision changes under evidence interventions. Each reader contributes a same-label swap (circle) and a direction reversal (diamond). The horizontal axis is the increase in abstention when that option is available; the vertical axis is the corresponding change in forced-choice accuracy. Terra is reported as a note because it is evaluated only in the abstention-enabled arm.

tic calibration fitted on 1,335 training claims decides by the sign of z and therefore matches the categorical reader's accuracy by construction; its larger fitting set cannot improve that accuracy.

What was queried, and when. The seven frontier readers are closed endpoints, so we identify them by the model identifier sent in each request: `gpt-6-astra`, `gpt-5.6-sol`, `gpt-5.6-terra`, and `gpt-5.6-luna` through OpenAI's Responses API, and `claude-fable-5-1`, `claude-opus-5`, and `claude-sonnet-5` through Anthropic's Messages API. Every response named the requested identifier as the serving model, and no fallback was enabled. All test requests were issued on 20 September 2026 through the standard non-batch interface, using a single user message, a JSON schema restricting the verdict to the arm's option set, the provider's low reasoning-effort setting, an output cap of 4,096 tokens, and no sampling parameter. The mechanism arm issued 1,244 requests per reader; the forced-choice arm added the remaining four swap draws for 2,540 requests per admitted reader, for 23,948 test requests in total. All returned successfully, six after a second attempt; two Claude Fable 5.1 responses were refusals and are scored incorrect, and none was invalid or truncated. The earlier instability screen used training claims only and is excluded from these counts. Because the endpoints are nondeterministic and may change, the reported experiment is the frozen set of raw responses, stored with provider request identifiers and SHA 256 file digests.

The frontier readers' prompt. Both arms use one template fixed before the instability screen and differ only in the option list. The question and option descriptions are the same as for the decision-native reader (Appendix J), and the evidence block is serialized as in Appendix K; in the closed-book condition the block is present but empty.

```
Claim: <claim>

Evidence:
<evidence block>
```

Table 17: Datasets in our experiments. Twitter and Weibo are social-media corpora; NewsCLIPpings and VERITE are image–caption consistency benchmarks; MOCHEG and AVeriTeC are fact-checking corpora with collected evidence. Models trained on NewsCLIPpings are evaluated zero-shot on VERITE.

Dataset	Task	Train	Test	Split
Twitter	post veracity	13,671	1,091	benchmark test split
Weibo	post veracity	7,531	1,996	held-out, events not separated
NewsCLIPpings	image-caption match	24,000	7,264	released split, label-balanced
VERITE-OOC	out-of-context	−−	492	balanced pairwise split
VERITE-MC	miscaptioned	−−	462	balanced pairwise split
MOCHEG	claim verification	5,149	1,068	released split, article-disjoint
AVeriTeC	claim verification	1,935	324	released test set

```
What is the verdict for the claim?

Choose exactly one:
supported - The evidence supports the claim.
refuted - The evidence contradicts the claim.
not_enough_information - The evidence is insufficient to determine the claim.

Answer with the option name only.
```

The forced-choice arm omits the third option line and removes that value from the schema's enumeration.

G Datasets, splits, and disclosures

G.1 Experimental setup

Datasets. We evaluate the GainMasque audit on six multimodal datasets: Twitter (Boididou et al., 2015), Weibo (Jin et al., 2017), NewsCLIPpings (Luo et al., 2021), VERITE (Papadopoulos et al., 2024), MOCHEG (Yao et al., 2023), and AVeriTeC (Schlichtkrull et al., 2023). DREF is evaluated on MOCHEG and AVeriTeC, with its admission mechanism further tested on NewsCLIPpings, VERITE, and Twitter.

Models and evidence. Our main backbone is Qwen3-VL-8B. We additionally evaluate multiple evidence-integration designs, retrieval/search agents, twelve vision-language backbones spanning 2B–72B parameters, and open and closed evidence readers. Evidence comes from four sources: an in-domain training bank, gold fact-checking evidence, retrieved candidate passages, and live web evidence.

Evaluation protocol. Unless otherwise specified, we report mean and standard deviation over three random seeds; the MOCHEG backbone study reports the median across twelve backbones. Each EG uses one same-label swap draw, and robustness is evaluated over eight corruption families at multiple rates. Full experimental details are provided in Appendix I.

G.2 Datasets, splits, and disclosures

Sizes and splits. Table 17 summarizes the dataset sizes and evaluation splits used in our experiments. For each seed, the training pool is partitioned 70 / 15 / 15 as described in Appendix I. For AVeriTeC, the label-stratified split contains 1,335 training, 300 calibration, and 300 gating claims.

The Twitter split. We use the benchmark's original train/test split, constructed to separate events across the two sides, restricted our test to posts that carry an image. We find the weaker property available from these files: the 13,671 training posts draw on 380 distinct image files and the 1,091 test posts on 103, with no image file shared between them. The event separation is a property of the benchmark construction that we rely on rather than re-establish.

Weibo's blank images. When none of the listed Weibo image files can be found, the loader supplies a 336×336 blank white image. Thus, every such sample shares one visual feature. These blank images account for 17.65% of the training data and 11.87% of the test data and are label-correlated: a rule that predicts "real" whenever the image is blank is correct for 73.4% of such training rows and 67.9% of such test rows. Resolving the image paths directly reproduces the same counts. No other dataset has a comparable blank image rate: Twitter is 0.04% on train and 0.00% on test, while NewsCLIPpings, MOCHEG, and both VERITE splits are 0.00%. Twitter also contains a shared visual feature in 7.69% of its training rows, originating from a photograph reused across posts rather than from a blank image.

We therefore report blank-free accuracies alongside the Weibo aggregates and carry the same recomputation through the fusion-grid bootstrap. Removing blank image rows changes any design's difference to concatenation by at most 0.08 points; for concatenation, accuracy changes from 0.8945 to 0.9005 and the evidence-branch gain from +0.08 to +0.32. We do not retrain without blank images, and thus Weibo's accuracies may still be inflated by the training shortcut. We therefore do not use Weibo to support claims of evidence gain.

VERITE's image attrition. We rebuilt the images from the source URLs in the released table and recovered 492 of 663 images; the remainder failed with HTTP errors, malformed URLs, or unreadable payloads. It leaves 723 of the benchmark's 1,001 source items with recoverable images. After constructing the balanced pairwise evaluations, the OOC and MC splits have 492 and 462 rows, respectively, as reported in Table 17. Sample inclusion depends on whether a source URL remained live at fetch time, the resulting subset is therefore not random.

MOCHEG. The binary subset keeps supported and refuted claims and drops the not-enough-information class, because the inherited protocol is two-class; Section 5 discusses the limitation for evidence sufficiency. Of the 5,149 training claims, 949 lack gold passages, and 948 of those are supported, making evidence availability strongly label-correlated in training. All 1,068 test claims have gold evidence, and thus availability cannot operate directly on the test split; the control of Section 4.2 therefore intervenes on training instead. The label prior also shifts from 0.597 refuted on training to 0.427 on test.

NewsCLIPpings. The training corpus is a fixed seeded subsample of 12,000 pristine–falsified pairs (24,000 rows), sampled by pair, preserving class balance and pairing. The subsample seed is fixed independently of the three experiment seeds, and all three runs use the same training corpus and seed variance does not capture the selection of subsample.

Compared systems. The audit focuses on the within-system comparison among the eight fusion designs of Appendix E, the trust variants of Appendix C, and a baseline from the prior-work components of Section 2. We additionally serve six search agents (Asai et al., 2024; Yan et al., 2024; Jin et al., 2025; Chen et al., 2025; Song et al., 2025; Zheng et al., 2025) the same evidence sets on AVeriTeC. We do not import published Twitter numbers because the splits differ, including prior random re-splits that can put the same events on both sides.

H Discussion and limitations

What the audit and repair establish. GainMasque separates retrieval gain (RG) from evidence gain (EG), with the latter as our focus. Based on the tested designs, the swap leaves three gains masked: label transfer from training-bank retrieval, fusion gain reproduced by shuffled evidence, and source-style signal in an evidence scorer; evidence availability requires a separate control. Gains removed by the swap arise from live provenance and a claim-conditioned scorer, although claim conditioning is necessary but not sufficient. DREF adds an admitted evidence residual under reference-based alignment and trust, retaining most clean gain on admitted sources while limiting the worst loss observed in our corruption suite. It also reduces, but does not eliminate, adaptive-attack harm and transfers to closed external readers. Across twelve backbones on MOCHEG, trust retains the gain on eleven and closes on the one base with unreliable confident verdicts. A repair can make a masked gain safer, but cannot make it claim-specific.

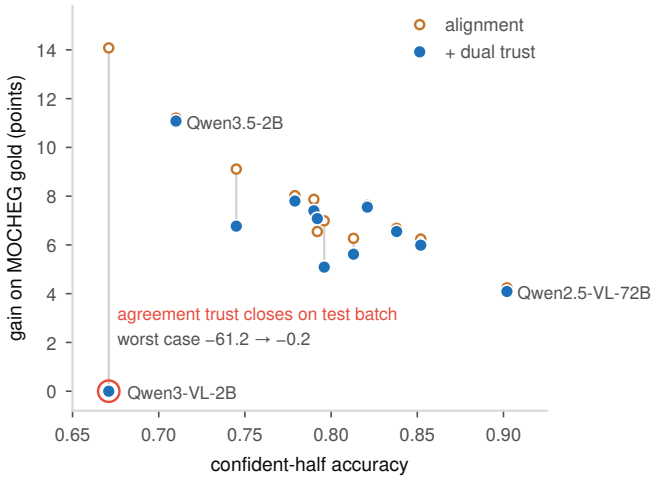


Figure 10: Gain retained by the dual trust on MOCHEG gold passages, for twelve backbones (seed means), ordered by the frozen base's accuracy on its confident half. The gain survives on eleven. On Qwen3-VL-2B, the test-batch agreement rule closes the trust and the clean gain drops from +14.1 to 0.0 points. Its worst loss across the eight corruption families is only -0.2, compared with -61.2 under alignment alone (Table 12).

Table 18: Experimental settings. Values that vary by backbone are given in Table 19.

Training		Architecture and evidence	
Batch size	32	Reasoning heads	3
Learning rate	10^{-4}	Head hidden / prelogit	512 / 128
Weight decay	10^{-4}	Classes	2
Optimizer	Adam	Dropout	0.3
Epochs	5	Fusion gate hidden	64
Checkpoint selection	Best accuracy on gate-fit	Feature token	last hidden state
Auxiliary head weight	0.3	Backbone	frozen
Agreement weight	0.05	Memory bank top-k	5
Train / gate-fit / calibration	70 / 15 / 15	Evidence passages per claim	3
Seeds	1, 2, 3	Evidence characters per passage	1,200

What the audit does not show. Given that the swap holds evidence availability fixed, availability-driven gains are invisible to it; we test this separately on MOCHEG. It also cannot detect evidence–test overlap, and thus we use event-disjoint splits where available and rely on Twitter's benchmark-defined split otherwise (Appendix G). The swap also cannot tell whether the signal is claim-specific or merely comes from information available after the verdict. The masked-gain mechanisms we identify are therefore not complete.

Where the repair does not help. DREF can shift, scale, or suppress an evidence scorer, but cannot change what the scorer reads. On MOCHEG, it makes label-correlated source style safer rather than claim-specific. The agreement rule also requires a base whose confident verdicts are usually reliable; on one of twelve backbones it closes and loses the gain, and we have no label-free rule that both retains that gain and bounds the harm (Section 4.5). The reported worst cases are observed losses within our corruption suite, not formal bounds.

I Implementation and the backbone grid

Here, we report the settings shared by the detector runs and the backbone-specific quantities used in the scale analysis. The backbone is frozen throughout and only the detector heads and fusion module are trained, and a run is determined by the backbone, dataset, fusion design and seed.

Table 19: The twelve MOCHEG backbones. d is the hidden size. Parameter counts are measured by the pipeline in the no-retrieval run and are identical for the concatenation design. Names are abbreviated: the instruction-tuned checkpoint of each is used. Base accuracy is the frozen detector without evidence on the MOCHEG test claims; confident-half accuracy is the accuracy of its verdict on the half of the test split where $|s_Q|$ is at or above the median. The last two columns are the clean gain in accuracy points from the evidence scorer under alignment alone and under dual trust. All are seed means over three seeds.

Backbone	d	dtype	Frozen (B)	Train. (M)	Base acc.	Conf.-half acc.	Align.	Dual trust
Qwen3-VL-2B	2048	4-bit	1.22	32.7	0.615	0.671	+14.08	+0.00
Qwen3-VL-4B	2560	fp16	4.44	50.0	0.688	0.779	+8.02	+7.80
Qwen3-VL-8B	4096	fp16	8.77	124.0	0.709	0.790	+7.87	+7.40
Qwen3-VL-32B	5120	bf16	33.36	191.6	0.739	0.838	+6.68	+6.55
Qwen2.5-VL-7B	3584	fp16	8.29	95.6	0.724	0.813	+6.27	+5.62
Qwen2.5-VL-72B	8192	bf16	73.41	482.6	0.778	0.902	+4.24	+4.09
InternVL3-38B	5120	bf16	38.39	191.6	0.748	0.852	+6.24	+5.99
InternVL3.5-8B	4096	bf16	8.53	124.0	0.705	0.792	+6.55	+7.08
Qwen3.5-2B	2048	bf16	2.21	32.7	0.632	0.710	+11.20	+11.08
Qwen3.5-4B	2560	bf16	4.54	50.0	0.658	0.745	+9.11	+6.77
Qwen3.5-9B	4096	bf16	9.41	124.0	0.700	0.796	+6.99	+5.09
Gemma-4-12B	3840	bf16	11.96	109.3	0.724	0.821	+7.62	+7.55

Detector overview. The detector is an experimental instrument of this work. A frozen vision-language backbone, Qwen3-VL-8B (Bai et al., 2025) in the main experiments and eleven others in the backbone grid, encodes each input. Three heads read content-internal, visual-conditioned, and evidence-augmented representations. Head C gives the prediction without evidence and Head E the prediction with evidence. Thus, the difference between Head C and Head E is the quantity the audit examines. A per-item gate fuses the heads from the predictive entropy and decision margin. We retain this gate because the arrival of evidence can change the fusion weights even when the evidence itself is irrelevant to the claim, making fusion a possible source of masked retrieval gain. Removing the evidence branch defines the base detector left unchanged by the repair, whose log-odds we denote by s_Q. Appendix K gives the prompts.

Training. Each branch representation is the final-layer hidden state at the last position of its prompt (Appendix K). Each reasoning head is a three-layer perceptron. The fusion gate weights the heads = from each head's predictive entropy and top-two margin. The training objective is

$$\mathcal{L} = \mathrm{CE}(z_F, y) + \lambda_{\text{heads}} \frac{1}{|H|} \sum_{k \in H} \mathrm{CE}(z_k, y) + \lambda_{\text{agree}} \mathcal{L}_{\text{agree}}, \qquad \mathcal{L}_{\text{agree}} = \frac{1}{\binom{|H|}{2}} \sum_{j<k} \mathrm{KL}_{\text{sym}}(z_j, z_k), \tag{21}$$

where CE is the unweighted cross-entropy loss, y is the gold label, z_F is the fused logit vector, and z_j and z_k are the logits of two active reasoning heads indexed by j and k. H is the set of active reasoning heads and $|H|$ its size, and $\binom{|H|}{2} = |H|(|H|-1)/2$ is the number of unordered head pairs. $\mathcal{L}_{\text{agree}}$ is the mean pairwise agreement regularizer, weighted by $\lambda_{\text{agree}} = 0.05$; the auxiliary head loss is weighted by $\lambda_{\text{heads}} = 0.3$.

$$p_j = \mathrm{softmax}(z_j), \qquad p_k = \mathrm{softmax}(z_k), \tag{22}$$

$$\mathrm{KL}_{\text{sym}}(z_j, z_k) = \frac{1}{2} \left[\mathrm{KL}(p_j \| p_k) + \mathrm{KL}(p_k \| p_j) \right]. \tag{23}$$

Here, p_j and p_k are the predictive distributions of heads j and k. KL_{sym} is the symmetric KL divergence between these two heads' predictive distributions. When fewer than two heads are active, $\mathcal{L}_{\text{agree}}$ is defined as zero. The agreement term only serves as a training regularizer and does not enter the audit statistics. The remaining shared training settings are given in Table 18.

Splits and seeds. For each seed, the training portion of a dataset is partitioned 70 / 15 / 15 into train, gate-fit, and calibration splits; the gate-fit split also serves as the detector validation split. The partition is redrawn with the run seed. Unless stated otherwise, reported results are means over

seeds 1, 2 and 3. All label-dependent fitting for the repair uses the calibration split, and test labels are never used. At inference, the agreement trust additionally computes its agreement slope from base predictions and evidence scores on the current unlabeled test batch, as defined in Section 3.3.

Backbones. Table 19 lists the twelve MOCHEG backbones and the quantities used in the scale analysis of Section 4.5. Qwen3-VL-8B is the main backbone; the other eleven repeat the audit and repair under the same protocol. All but Qwen3-VL-2B are loaded unquantised in fp16 or bf16 as recorded in the table; Qwen3-VL-2B uses 4-bit NF4 with double quantisation and fp16 compute. Gemma-4-12B-it required an isolated `transformers` 5.17.0 environment; the remaining runs use the default environment below. Artifact names use the `fp16` tag for all unquantised runs, including bf16 runs, because the tag records quantisation status rather than storage dtype. Figure 10 plots the clean gain retained by dual trust against the base detector's confident-half accuracy.

Compute. Feature extraction is the only stage of the detector grid that requires a GPU. It uses 40 GB A100s or a 94 GB H100, with 2 to 4 A100s for Qwen3-VL-32B, InternVL3-38B and Qwen2.5-VL-72B. Once features are frozen, the detector, audit and repair computations are deterministic and run on CPU; one complete MOCHEG seed through all configurations takes about 25 minutes on 8 cores. The logs contain 1,293 CPU tasks with a median runtime of 4.7 minutes and 53 timestamped GPU jobs totaling about 100 job-hours. These figures describe the retained experimental logs rather than the full history of the project.

J CLOSED DECISION-NATIVE READER: JEV

This appendix reports the protocol behind Section 4.5. The external reader, Jev, is a commercial API whose outputs cannot be reproduced from weights, and therefore we describe the frozen version, interface, fitting procedure, queries, and evaluation protocol.

Model and interface. We use TypeSafe System One, requested as `jev-latest` and resolved by the service to `jev-1.13.0`, the version returned on every call reported here. Each request contains the claim and served evidence, the fixed question `"What is the verdict for the claim?"`, and the three options `supported, refuted, and not enough information` in a fixed order and with fixed descriptions. Only the evidence field changes across conditions. Evidence is serialized from the same scrubbed snippets used by our own readers.

Interface variability. The output probabilities are quantized to 0.01 and, in $0.07-0.67\%$ of responses, sum to 0.99; we then renormalize them before treating a response as a distribution. The endpoint is nondeterministic and exposes no seed or temperature. Across 100 claims queried five times each, the mean maximum per-option standard deviation is 0.013, and the predicted decision changes for 2 of 100 claims across byte-identical requests. We therefore treat the frozen responses as the experimental observations and do not average repeated calls when fitting components, which would otherwise make fitted quantities depend on uncontrolled service randomness.

Fitting split and freeze. The residual scale and bias, admission test, alignment and trust constants, and abstention threshold are fitted on the existing AVeriTeC calibration splits rather than on resampled test data: each seed uses stratified draws of 300 calibration and 300 gating claims from the 1,935 training claims. The abstention threshold is fitted only on the gating rows, which no other component reads, by minimizing pooled negative log-likelihood over the two training-side conditions; the resulting thresholds are 0.95, 0.70, and 0.30 across the three seeds. Before any test claim is scored, we freeze all fitted quantities, admission decisions, feature definitions, calibration and gating row identifiers, the deployment rule, and the five partner draws in a manifest. Evaluation reconstructs these objects and asserts equality with the manifest; a mismatch fails rather than silently changing a result.

Queries. The training side requires 7,107 distinct requests after deduplication. Requests are identified by the state, question, and options; a claim shared across seeds is queried once and referenced from each split. Each request is issued once, at a total cost of $\$0.16$.

Statistics. For the reported confidence intervals, we explore a claim-clustered bootstrap with 10,000 replicates. The same resampled claim indices are used for both systems in each contrast, and the five partner draws are averaged within claim before resampling. The worst degradation, $\Delta_{\min}$, is recomputed in every replicate and compares each condition with the claim-based reference evaluated on the same claim set. Metrics under $E_{\pi(i)}$ average the five partner draws before aggregation. $NLL(E_i)$ is computed under the original evidence, while $\overline{NLL}$ is the unweighted mean of the condition-level NLLs over E_i, $E_{\pi(i)}$, and $\mathcal{R}(E_i)$. We employ NLL as a measure of probabilistic predictive performance, not as a calibration metric.

Reversed-direction condition. $\mathcal{R}(E_i)$ is defined on a subset of the test claims. $\Delta_{\min}$ compares each condition with the claim-based reference evaluated on the same claim set. $\mathcal{R}(E_i)$ inverts the stated polarity of one answer while preserving its question; a matched control truncates the same answer without inversion.

K PROMPTS

Every prompt is a single user message and serialized with the loaded backbone's processor chat template; the feature is the final-layer hidden state at the last position of the serialized sequence. Because the grid spans four model families, we do not print a serialized example: the same message content produces a different string under each family's template.

Branch prompts. Three messages are built for each information unit. The content branch is text only,

```
Is the following news real or fake?  News text:  {text}
```

the multimodal branch carries the image before the question,

```
Is the following news real or fake?  Use both the news text
and image.  News text:  {text}
```

and the image branch is

```
Represent this image for visual retrieval.
```

There is no separate evidence prompt: the evidence branch reuses the multimodal message and differs only in how the retrieved neighbors are fused. The image feature enters no reasoning head, and the memory bank is queried with the multimodal feature. It is instead used to construct and rank corpus-photo evidence and to identify blank image placeholders in Weibo.

Evidence prompts. Passages are encoded by the same frozen backbone, one message per passage,

```
Evidence relevant to verifying a news claim:  {text}
```

after HTML is stripped, the text truncated to 1,200 characters, and label or verdict words removed at the scrub level of Appendix A; at most three passages are kept per claim. Image evidence uses

```
Describe the visual content of this image.
```

with images thumbnailed to 448×448 and at most three retained per claim. The prompt string is stored with the resulting features in the cache.

The reader prompt. On AVeriTeC, the reader-conditioned scorer uses the reader model's hidden state from a prompt containing the claim and its served passages, serialised as an evidence block in source order. The text reader truncates at 3,072 tokens with left padding. The claim-withheld control of Section 4.3 removes the claim from this prompt.

L EXTENDED RELATED WORK

Background. Social media platforms carry a large volume of unverified and distorted content (Shu et al., 2017; Zhou & Zafarani, 2020; Abdali et al., 2024; Su et al., 2020; Ngueajio et al., 2025), including multimodal content (Zhong & Baghel, 2024; Huang et al., 2025), which undermines public trust and contributes to suboptimal health beliefs and decisions (Thorson et al., 2010; Schmid & Bauer, 2026); unimodal and multimodal models are proposed to detect and filter such content (Guo et al., 2020; Yang et al., 2023; Shi et al., 2023a; Bhattarai et al., 2022). Most of these methods integrate multiple feature streams, including external knowledge from a knowledge bank or retrieval (Lewis et al., 2020), into a single classifier without separating what is learned from a post's original content from what the knowledge contributes.

L.1 CONTENT-INTERNAL REASONING METHODS

Traditional machine learning methods have been employed to classify the textual content of information units into predefined classes; however, these models typically rely on hand-crafted features and struggle to capture complex semantic patterns, thus limiting detection accuracy (Minaee et al., 2021). As neural networks have emerged, deep learning frameworks have further improved performance by extracting complex content features and capturing nuanced semantic representations, such as convolutional neural networks (Kim, 2014), recurrent neural networks (Ma et al., 2016) and long short-term memory (Sachan et al., 2019). In addition, attention networks improve model performance by incorporating feature embeddings extracted from different dimensions of information units (Yang et al., 2016; Linmei et al., 2019). By introducing pre-trained embeddings, the transformer-based methods have further boosted detection accuracy for incorrect information content (Croce et al., 2020; Van Nooten & Daelemans, 2025). Wu et al. (2024) emphasized the substantive content over stylistic features, using large language models to reframe news articles and focus on content correctness. But, the separation between stylistic and substantive content remains limited.

L.2 KNOWLEDGE-AUGMENTED REASONING METHODS

As data structures and modalities continue to expand, multimodal approaches (**?**) have been proposed to handle more intricate information reasoning tasks by incorporating external knowledge sources and reasoning across text, images, video, audio data, or multiple languages. For instance, Wang et al. (2018) proposed event adversarial neural networks (EANN), which combine an event discriminator into the detection model and learn event features from text and image data; though it has capacity to detect unseen topics or events, EANN relies heavily on the quality of event features. Khattar et al. (2019) presented a multimodal variational autoencoder (MVAE) to learn a shared representation of textual and visual content and capture cross-modality dependency. Zhou et al. (2020) proposed a similarity-aware method (SAFE) that ensures consistency between text and image, but it only identifies one type of inconsistency. To evaluate the ambiguity between textual and visual data, Chen et al. (2022) proposed cross-modal ambiguity learning (CAFE), which adaptively weights text and image representations by measuring their ambiguity. Though achieving better performance, it ignores the contextual knowledge of multimodal information and does not quantify the contributions from different modalities.

Another line of incorrect information detection research focuses on enhancing detection performance through knowledge augmentation. Retrieval-augmented methods (Guu et al., 2020; Lewis et al., 2020) incorporate the retrieved knowledge units into language models to improve factual reasoning, and a dual-head reasoning model (Zhang et al., 2026) separates content reasoning from knowledge reasoning for textual content. Vision-language models (Dai et al., 2023; Liu et al., 2023; Bai et al., 2025) have shown strong potential for multimodal reasoning by jointly encoding textual and visual information; however, these methods still suffer from outdated or incomplete knowledge, hallucinated reasoning, and insufficient uncertainty calibration when handling noisy multimodal evidence. Retrieval has since become adaptive: Asai et al. (2024) train a model to decide when to retrieve and to critique what it reads, Yan et al. (2024) add an evaluator that discards low-quality retrievals, and search agents trained with reinforcement learning interleave queries with reasoning (Jin et al., 2025; Chen et al., 2025; Song et al., 2025; Zheng et al., 2025). On the data side, benchmarks now pair claims with real evidence, fact-check passages in MOCHEG (Yao et al., 2023)

and question-answer evidence in AVeriTeC (Schlichtkrull et al., 2023), while NewsCLIPpings (Luo et al., 2021) and VERITE (Papadopoulos et al., 2024) pair images with captions out of context.

Multimodal incorrect information. Text-based detection models moved from hand-crafted features (Minaee et al., 2021) through neural encoders (Kim, 2014; Ma et al., 2016; Sachan et al., 2019; Yang et al., 2016; Linmei et al., 2019) to pre-trained transformers (Croce et al., 2020; Van Nooten & Daelemans, 2025); Wu et al. Wu et al. (2024) separated substantive content from style on the content side. On the contrary, GAINMASQUE examines a separation on the evidence side. Multimodal detectors add the visual stream through event-invariant features (Wang et al., 2018), shared latent representations (Khattar et al., 2019), text–image consistency (Zhou et al., 2020), and ambiguity-weighted fusion (Chen et al., 2022); vision-language models further (Dai et al., 2023; Liu et al., 2023; Bai et al., 2025) encode text and images jointly. Retrieval-augmented methods instead condition predictions on retrieved knowledge (Guu et al., 2020; Lewis et al., 2020), while a dual-head model separates content from knowledge reasoning (Zhang et al., 2026). Furthermore, retrieval has increasingly adopted adaptive strategies, deciding when to retrieve (Asai et al., 2024), discarding low-quality retrievals (Yan et al., 2024), or interleaving retrieval with reasoning (Jin et al., 2025; Chen et al., 2025; Song et al., 2025; Zheng et al., 2025). Across these approaches, accuracy gains from external knowledge are rarely tested against evidence that retains label-level signal while breaking its alignment with the claim.

Calibration and trust. Prior work provides complementary mechanisms for incorporating an external signal: stacking combines its score with a base prediction (Wolpert, 1992), calibration corrects test-time shifts (Lipton et al., 2018; Zhou et al., 2024; Steyerberg et al., 2004; Platt, 1999), held-out safety tests decide whether an update should be admitted (Thomas et al., 2019), and relevance or prediction agreement can control unreliable evidence (Yan et al., 2024; Baek et al., 2022). However, their references remain largely independent: batch calibration corrects the scorer against its marginal distribution, a held-out safety test depends on a single data split, and relevance and agreement gate the update separately. DREF instead defines each decision relative to the frozen base and the calibration split. Our alignment rule only corrects shifts of the scorer relative to the base prediction, cross-fitted admission explores every calibration row for testing before refitting on the full split, and dual trust caps the update by the weaker of relevance- and agreement-based evidence. This reference-relative formulation preserves substantially more clean gain than these baselines at comparable worst-case loss (Section 4.4), while using the same base model, evidence scorer, verifier, calibration data, and corruptions.

Evidence evaluation. Retrieval robustness is usually tested by replacing evidence with irrelevant or random passages (Shi et al., 2023b; Yoran et al., 2024; Cuconasu et al., 2024). Such controls break both claim–evidence alignment and label-level signal; our GAINMASQUE audit instead preserves the latter while breaking the former, allowing the two sources of gain to be distinguished.