

EGR-FV: Evidence-Grounded Remixing for Debaised Fact Verification via Counterfactual Necessity-Guided Training

Jun Yang¹, Dandan Song^{1*}, Zhijing Wu¹, Yuhang Tian¹, Huipeng Ma¹,
Chenhao Li¹, Yizhou Jin¹, Guangyuan Feng¹, Xudong Li¹

¹School of Computer Science and Technology, Beijing Institute of Technology, China
{yangjun123, sdd}@bit.edu.cn

Abstract

Automatic fact verification (FV) predicts the veracity of a claim against retrieved evidence, yet modern verifiers often behave as *shortcut-driven verifiers*, predicting from surface regularities in the claim rather than the factual content of the evidence. Existing debiasing methods over-correct this by uniformly enforcing evidence reliance on every sample, thereby suppressing the valid information already carried by the claim itself. We propose **Evidence-Grounded Remixing for Fact Verification (EGR-FV)**, which estimates a per-sample evidence necessity score by contrasting claim-with-evidence and claim-only predictions, and uses it to remix training batches so that debiasing is adapted to each sample’s actual reliance on evidence. On FEVER, PolitiHop, and HOVER together with their symmetric counterparts, EGR-FV achieves state-of-the-art accuracy and robustness, validating the effectiveness of EGR-FV.

1 Introduction

Automatic fact verification (FV) (Thorne et al., 2018; Jiang et al., 2020; Liu et al., 2020; Zhou et al., 2019; Chen et al., 2022), which assesses the truthfulness of a claim based on evidence, plays a critical role in curbing the spread of online misinformation (Vosoughi et al., 2018; Allcott and Gentzkow, 2017; Kogan et al., 2019).

Following FEVER (Thorne et al., 2018), GEAR (Zhou et al., 2019), KGAT (Liu et al., 2020), DREAM (Zhong et al., 2020), EvidenceNet (Chen et al., 2022), BEVERS (DeHaven and Scott, 2023), and Claim-Dissector (Fajcik et al., 2023) improve FV via a three-stage pipeline of document retrieval, evidence selection, and claim verification, instantiating graph- or semantic-reasoning verifiers and end-to-end pipeline systems. More recently, with the rapid progress of large language models, DagFC (Wu et al., 2026) and SLED (Yang

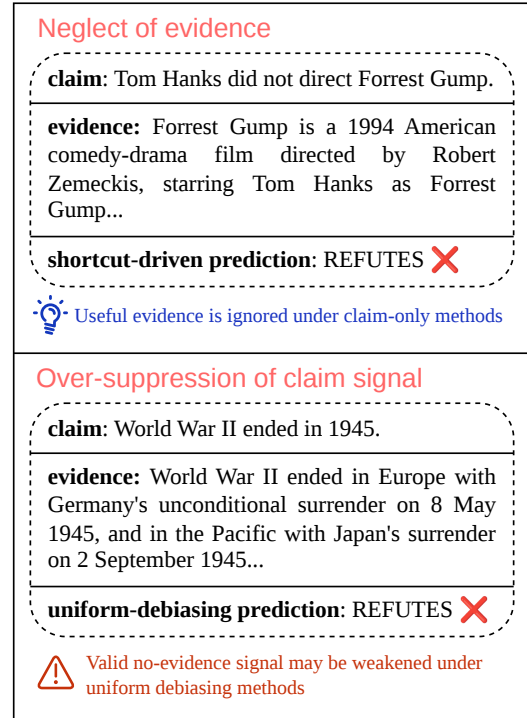


Figure 1: Two complementary failure modes that motivate EGR-FV. (a) A shortcut-driven verifier mispredicts a negated true claim. (b) Uniform debiasing suppresses the valid claim signal on a self-evident claim, leading to an incorrect prediction.

et al., 2026a) further boost verification performance with LLMs through compositional, retrieval-augmented, or debate-based reasoning.

Despite these advances, existing verifiers often behave as **shortcut-driven verifiers** that predict the verdict from surface regularities in the claim (e.g., keywords, syntactic templates, or label-correlated phrasings) rather than from the factual content of the evidence. Figure 1(a) provides a concrete instance: having internalized a spurious “negation $\rightarrow$ REFUTES” correlation, the verifier mispredicts the true claim “*Tom Hanks did not direct Forrest Gump.*” as REFUTES, despite the gold label being SUPPORTS.

*Corresponding author

Several works attempt to remove this shortcut via counterfactual subtraction (Xu et al., 2023), front-door adjustment (Zhang et al., 2024), and dual-branch counterfactual inference (Yang et al., 2026b). However, by uniformly enforcing evidence reliance on every sample, these methods overlook claims that are legitimately decidable from their own content and end up suppressing such valid, non-spurious claim-side signals. As shown in Figure 1(b), a uniform-debiasing verifier on the well-known fact “World War II ended in 1945.” treats the high claim-only confidence as a shortcut and subtracts it, mispredicting the claim as REFUTES despite the gold label being SUPPORTS. These methods overemphasize the role of evidence on every sample, thereby suppressing the valid information already carried by the claim itself.

To overcome the limitations of uniform debiasing, we propose **Evidence-Grounded Remixing for Fact Verification (EGR-FV)**. EGR-FV intervenes on the evidence to model an explicit counterfactual world that conditions only on the claim, and contrasts it against the factual world that conditions on both claim and evidence. The disagreement between the two worlds proxies the individual treatment effect of evidence on the verdict and is summarized as a per-sample *evidence necessity score*. Using this score, EGR-FV routes each training example into one of three subsets—grounded-needed, hard, or bias-easy—and then composes each training batch by remixing the three subsets at fixed proportions. This sample-level routing together with batch-level remixing makes full use of the information in every example while ensuring that all examples in a batch push the optimizer in a single, evidence-sensitive direction.

Systematic experiments on FEVER, PolitiHop, and the multi-hop HOVER benchmark together with their symmetric counterparts show that EGR-FV achieves state-of-the-art performance, demonstrating the effectiveness of our method. Our contributions are:

- We introduce a *sample-level evidence necessity* perspective, observing that different samples depend on evidence to substantially different degrees.
- We propose **Evidence-Grounded Remixing for Fact Verification (EGR-FV)**, which estimates per-sample evidence necessity and remixes training batches to counterfactually rebalance the training distribution.

- Systematic experiments on FEVER, PolitiHop, and HOVER show that EGR-FV improves both accuracy and robustness over strong debiasing baselines.

2 Related Work

2.1 Fact Verification

Early fact verification (FV) methods predict veracity from the claim alone, relying on lexical or stance cues (Wang, 2017; Pérez-Rosas et al., 2018). On benchmarks such as FEVER (Thorne et al., 2018), HoVer (Jiang et al., 2020), FaVIQ (Park et al., 2022), and CHEF (Hu et al., 2022), GEAR (Zhou et al., 2019), KGAT (Liu et al., 2020), DREAM (Zhong et al., 2020), EvidenceNet (Chen et al., 2022), and LisT5 (Jiang et al., 2021) adopt a three-stage claim–evidence pipeline of retrieval, sentence selection, and verification with graph- or semantic-reasoning verifiers, while BEVERS (DeHaven and Scott, 2023) and Claim-Dissector (Fajcik et al., 2023) push the pipeline-level design further. More recently, DagFC (Wu et al., 2026) and SLED (Yang et al., 2026a) extend FV with LLM-based compositional or debate-based reasoning. However, despite strong benchmark accuracy, these methods often behave as *shortcut-driven verifiers* that predict from surface regularities in the claim rather than from the factual content of the evidence.

2.2 Debiasing Fact Verification

A growing body of work mitigates this bias from different angles: re-weighted training with a symmetric test set (Schuster et al., 2019), contrastive evidence augmentation (VitaminC (Schuster et al., 2021), CrossAug (Lee et al., 2021)), counterfactual subtraction (CLEVER (Xu et al., 2023)), front-door adjustment (Causal Walk (Zhang et al., 2024)), dual-branch counterfactual inference with a Jensen–Shannon dynamic constraint (Dual-Debias (Yang et al., 2026b)), and related ideas in adversarial OOD debiasing (Liu et al., 2024) and VQA constraint-based debiasing (Li et al., 2023). However, these methods apply a uniform correction to every sample, overemphasizing the role of evidence and thereby suppressing the valid information already carried by the claim itself.

3 Method

We propose **Evidence-Grounded Remixing for Fact Verification (EGR-FV)**, a debiasing framework built on the premise that samples differ in

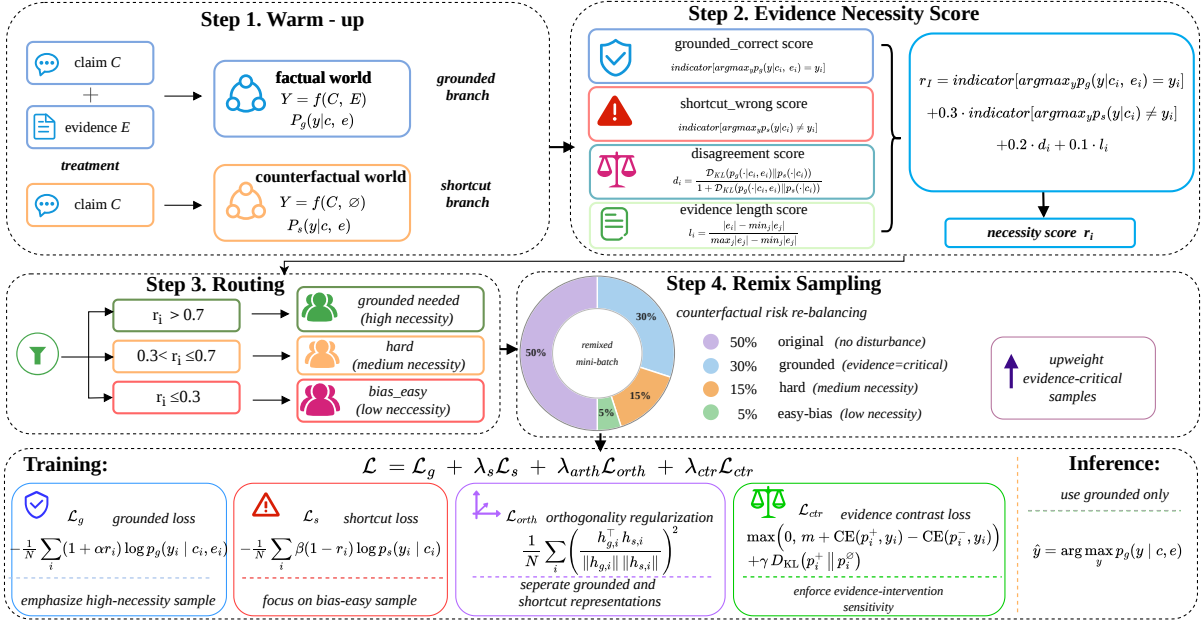


Figure 2: Overview of the EGR-FV framework. After a warm-up stage, EGR-FV computes the evidence necessity score r_i , uses it for sample-level routing, performs proportional remix sampling, and yields the debiased verifier through multi-objective training.

how much they need evidence, so debiasing should be adapted per sample. As illustrated in Figure 2, EGR-FV jointly trains a grounded branch and a shortcut branch, derives a per-sample evidence necessity score from their disagreement, and uses it to route examples into subsets and remix mini-batches, steering optimization toward evidence-grounded decisions.

3.1 Warm-up

Following Xu et al. (2023); Zhang et al. (2024), given a claim c_i and evidence e_i , we build a grounded branch f_g to model the factual world conditioned on (c_i, e_i) ,

$$p_g(y_i | c_i, e_i) = f_g(c_i, e_i), \quad (1)$$

where $y_i \in \mathcal{Y} = \{\text{SUPPORTS}, \text{REFUTES}\}$.

We then intervene on evidence with $\text{do}(E = \emptyset)$ and build a shortcut branch f_s to model the counterfactual world in which evidence is removed,

$$p_s(y_i | c_i) = f_s(c_i) \approx p(y_i | c_i, E = \emptyset). \quad (2)$$

The gap between the two worlds reflects how much the prediction depends on evidence rather than claim-only shortcuts.

3.2 Evidence Necessity Estimation

To quantify whether an instance requires evidence-grounded reasoning, we define an *evidence neces-*

sity score $r_i \in [0, 1]$ for each training example:

$$\begin{aligned} r_i = & 0.4 \cdot \mathbb{1} \left[\arg\max_y p_g(y | c_i, e_i) = y_i \right] \\ & + 0.3 \cdot \mathbb{1} \left[\arg\max_y p_s(y | c_i) \neq y_i \right] \\ & + 0.2 \cdot d_i + 0.1 \cdot \ell_i, \end{aligned} \quad (3)$$

where $\mathbb{1}[\cdot]$ is the indicator function: the first term equals 1 when the grounded branch predicts correctly, and the second term equals 1 when the shortcut branch predicts incorrectly. d_i measures the prediction disagreement between the grounded and shortcut branches, and ℓ_i denotes a normalized evidence-length score. The four weights are ordered by signal reliability, and a sensitivity analysis confirming this design is reported in Appendix F. The disagreement term is defined as the KL divergence between the two branches squashed into a bounded range,

$$d_i = \frac{\mathcal{D}_{\text{KL}}(p_g(\cdot | c_i, e_i) \| p_s(\cdot | c_i))}{1 + \mathcal{D}_{\text{KL}}(p_g(\cdot | c_i, e_i) \| p_s(\cdot | c_i))}, \quad (4)$$

which bounds $d_i \in [0, 1)$ and prevents a few outliers from dominating the routing.

The length score ℓ_i is used as a weak proxy for evidence complexity. Let $|e_i|$ denote the number of tokens in the evidence. We compute

$$\ell_i = \frac{|e_i| - \min_j |e_j|}{\max_j |e_j| - \min_j |e_j|}. \quad (5)$$

A high r_i indicates that the example is likely to require evidence for correct verification. In contrast, a low r_i suggests that the label may be predictable from claim-only shortcuts.

3.3 Routing Policy

Based on r_i , EGR-FV partitions $\mathcal{D} = \mathcal{D}_{\text{grounded}} \cup \mathcal{D}_{\text{hard}} \cup \mathcal{D}_{\text{bias}}$, where each instance is assigned by

$$z_i = \begin{cases} \text{grounded}, & r_i > 0.7, \\ \text{hard}, & 0.3 < r_i \leq 0.7, \\ \text{bias-easy}, & r_i \leq 0.3. \end{cases} \quad (6)$$

Examples in $\mathcal{D}_{\text{grounded}}$ are those for which the grounded branch succeeds while the shortcut branch tends to fail. These examples are considered evidence-critical. Examples in $\mathcal{D}_{\text{hard}}$ are difficult for both branches and usually require more robust semantic reasoning. Examples in $\mathcal{D}_{\text{bias}}$ can often be solved by the shortcut branch alone, suggesting that they may contain strong claim-only artifacts. The empirical distribution of r_i and the resulting group sizes on each dataset are reported in Appendix D.

3.4 Remix Sampling

The original training distribution may contain a large proportion of examples with weak treatment effects, i.e., examples for which

$$p(y_i | c_i, e_i) \approx p(y_i | c_i, E = \emptyset). \quad (7)$$

Such examples encourage the model to rely on the shortcut path $c_i \rightarrow y_i$ rather than the grounded path $(c_i, e_i) \rightarrow y_i$.

To reduce shortcut dominance, we construct a counterfactual remixing strategy that rebalances the training batches according to the routed subsets. At each training step, a mini-batch is sampled from a mixture of the original training set and the routed subsets, $\mathcal{B} = \mathcal{B}_{\text{orig}} \cup \mathcal{B}_{\text{grounded}} \cup \mathcal{B}_{\text{hard}} \cup \mathcal{B}_{\text{bias}}$, with sampling proportions $\pi = [0.50, 0.30, 0.15, 0.05]$. That is, each batch contains 50% original examples, 30% grounded examples, 15% hard examples, and 5% bias-easy examples.

This remixing policy increases the frequency of evidence-critical examples while retaining a small portion of bias-easy examples to preserve coverage of the original data distribution. In this way, the training distribution is counterfactually rebalanced toward examples where

$$p(y_i | c_i, e_i) \neq p(y_i | c_i, E = \emptyset), \quad (8)$$

thereby encouraging the model to learn evidence-sensitive decision boundaries.

Formally, the remixing objective can be interpreted as minimizing the expected risk under a reweighted training distribution q_{EGR} :

$$R_{\text{EGR}} = \mathbb{E}_{(c,e,y) \sim q_{\text{EGR}}} [\ell(f_g(c, e), y)], \quad (9)$$

where q_{EGR} assigns larger sampling probability to evidence-critical and hard examples than the empirical training distribution.

3.5 Training Objective

EGR-FV jointly optimizes the grounded branch and the shortcut branch while regularizing their representations and enforcing evidence contrast. The overall loss is

$$\mathcal{L} = \mathcal{L}_g + \lambda_s \mathcal{L}_s + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}} + \lambda_{\text{ctr}} \mathcal{L}_{\text{ctr}}, \quad (10)$$

where $\mathcal{L}_g$ is the grounded loss, $\mathcal{L}_s$ is the shortcut loss, $\mathcal{L}_{\text{orth}}$ is an orthogonality regularizer, and $\mathcal{L}_{\text{ctr}}$ is the evidence contrastive loss.

Grounded and shortcut losses. The grounded branch upweights evidence-critical examples while the shortcut branch focuses on bias-easy ones via complementary weights $(1 + \alpha r_i)$ and $\beta(1 - r_i)$:

$$\begin{aligned} \mathcal{L}_g &= -\frac{1}{N} \sum_i (1 + \alpha r_i) \log p_g(y_i | c_i, e_i), \\ \mathcal{L}_s &= -\frac{1}{N} \sum_i \beta(1 - r_i) \log p_s(y_i | c_i). \end{aligned} \quad (11)$$

Orthogonality regularization. To prevent the two branches from collapsing onto identical representations, we penalize the squared cosine similarity between their [CLS] hidden states $h_{g,i}$, $h_{s,i}$:

$$\mathcal{L}_{\text{orth}} = \frac{1}{N} \sum_i \left(\frac{h_{g,i}^\top h_{s,i}}{\|h_{g,i}\| \|h_{s,i}\|} \right)^2. \quad (12)$$

Evidence contrastive loss. To make the model sensitive to valid evidence, we instantiate three evidence interventions per claim— $E_i^+ = e_i$, $E_i^- = \text{do}(E = e_{\pi(i)})$, and $E_i^\emptyset = \text{do}(E = \emptyset)$ —corresponding to factual, wrong-evidence, and null interventions, where $\pi(i) \neq i$ cyclically rotates evidence within the mini-batch so that E_i^- keeps the same domain and length as e_i but is no longer the gold evidence (full construction details are given in Appendix E). Let p_i^+ , p_i^- , $p_i^\emptyset$ denote $p_g(\cdot | c_i, \cdot)$ under the three conditions; the contrastive loss is

$$\begin{aligned} \mathcal{L}_{\text{ctr}} &= \max(0, m + \text{CE}(p_i^+, y_i) - \text{CE}(p_i^-, y_i)) \\ &\quad + \gamma \mathcal{D}_{\text{KL}}(p_i^+ \| p_i^\emptyset). \end{aligned} \quad (13)$$

	FEVER			PolitiHop			HOVER		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
SUPPORTS	80,035	3,333	3,333	79	21	21	9,920	1,103	2,000
REFUTES	29,775	3,333	3,333	386	98	150	6,433	715	2,000
Total	109,810	6,666	6,666	465	119	171	16,353	1,818	4,000
Sym. test	717 (379/338)			342 (171/171)			—		

Table 1: Per-label split statistics for the three FV benchmarks under the 2-class setup; *Sym. test* denotes the symmetric test split (SUPPORTS/REFUTES).

The margin term prefers the gold evidence over corrupted evidence, and the KL term forces the grounded branch to encode information that cannot be recovered from the claim alone.

3.6 Inference

At inference time, we use the grounded branch as the final verifier:

$$\hat{y} = \arg \max_y p_g(y | c, e). \quad (14)$$

The shortcut branch is used only during training to estimate and regulate shortcut behavior. This design allows EGR-FV to benefit from shortcut diagnosis during training while relying on evidence-grounded reasoning during prediction.

4 Experiments Settings

4.1 Datasets and Metrics

We evaluate EGR-FV on three FV benchmarks of increasing reasoning complexity—**FEVER** (Thorne et al., 2018) (single-hop, Wikipedia), **PolitiHop** (Ostrowski et al., 2021) (multi-hop political fact-checking), and **HOVER** (Jiang et al., 2020) (multi-hop Wikipedia, stratified by 2/3/4-hop)—and additionally evaluate on their symmetric counterparts (Schuster et al., 2019) that decorrelate claim wording from labels to directly stress claim-only shortcuts. We report accuracy on FEVER and PolitiHop and macro-F1 on HOVER, all under gold evidence so that the gains reflect claim–evidence reasoning rather than retrieval. For symmetric testing, models are trained and selected only on the original training/development splits, and the symmetric labels are used strictly for final evaluation; details of the paired construction and evaluation protocol are given in Appendix B. Per-label split statistics are reported in Table 1; per-dataset descriptions are in Appendix A.

4.2 Baselines

We compare against four families of methods, all using RoBERTa-base as the backbone for fair comparison unless otherwise noted. **Standard classifiers:** *Claim-only* fine-tunes RoBERTa on the claim alone, and *Claim-evidence* is a standard claim+evidence cross-encoder. **Evidence-aware verifiers:** *GEAR* (Zhou et al., 2019) aggregates evidence with a graph attention network; *LisT5* (Jiang et al., 2021) performs listwise evidence reasoning with T5; *DeBERTaV3-NLI* (He et al., 2023) is a strong off-the-shelf NLI verifier. **Debiasing methods:** *CICR / CICR-graph* (Tian et al., 2022) apply claim–evidence causal intervention with and without a graph-structured evidence encoder; *CLEVER / CLEVER-graph* (Xu et al., 2023) apply counterfactual subtraction to remove the claim-only direct effect, with the -graph variant adding the same graph encoder; *Causal Walk* (Zhang et al., 2024) adopts a front-door adjustment tailored to multi-hop FV. All debiasing baselines are evaluated in the same 2-class, gold-evidence setting as EGR-FV, with no tuning on symmetric test labels; Appendix C gives the implementation and reporting details. **Large-model verifier pipelines** (HOVER only): *ProgramFC* (Pan et al., 2023) ($n = 1, 5$), *PACAR* (Zhao et al., 2024), *LoCal* (Ma et al., 2025), *BiDeV* (Liu et al., 2025), and *DagFC* (Wu et al., 2026) are recent LLM-based compositional verifiers that substantially exceed RoBERTa in parameter count, providing a stress test against EGR-FV.

4.3 Implementation Details

Both branches use RoBERTa-base (~ 125 M parameters); the grounded branch tokenises $[\text{CLS}; c; \text{SEP}; e; \text{SEP}]$ and the shortcut branch tokenises $[\text{CLS}; c; \text{SEP}]$. We train with AdamW (Loshchilov and Hutter, 2019) at learning rate 2×10^{-5} , weight decay 0.01, dropout 0.1, batch size 16, gradient accumulation 2, for 3–5 epochs (dataset-dependent), with checkpoint selection on the validation macro-F1.

For routing we use $K = 5$ folds; necessity-score weights $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.4, 0.3, 0.2, 0.1)$; routing thresholds $\tau_h = 0.7$ and $\tau_l = 0.3$. The remix sampler uses ratios (50%, 30%, 15%, 5%) over (ORIGINAL, grounded_needed, hard, bias_easy). Loss weights are $\lambda_s = 0.3$, $\lambda_o = 0.05$, $\lambda_c = 0.2$, with contrastive margin $m = 0.3$ and null-KL coefficient $\gamma = 0.1$. A sensitivity analysis showing that EGR-FV is robust to these choices is

Method	FEVER		PolitiHop	
	orig.	sym.	orig.	sym.
claim-only	78.13	38.63	80.06	54.39
claim-evidence	90.28	70.75	78.88	57.31
GEAR	92.17	44.91	75.50	51.17
CICR	92.41	74.76	88.30	59.65
CICR-graph	92.63	75.31	85.96	60.53
CLEVER	84.98	86.75	76.00	42.40
CLEVER-graph	85.63	88.42	78.00	52.05
CausalWalk	91.60	72.66	80.00	57.02
EGR-FV (ours)	94.09	83.26	88.30	63.45

Table 2: Accuracy on FEVER and PolitiHop (original / symmetric splits). Symmetric splits stress claim-only shortcuts; bold marks the best score in each column.

Category	Method	2-hop	3-hop	4-hop
Small-model classifiers	claim-only	59.39	67.88	57.23
	claim-evidence	78.82	79.64	77.96
	LisT5	56.15	53.76	51.67
	DeBERTaV3-NLI	77.22	65.98	60.49
	GEAR	74.70	75.04	72.85
Large-model pipelines	ProgramFC ($n=1$)	74.10	66.13	65.69
	ProgramFC ($n=5$)	75.65	68.48	66.75
	PACAR	76.86	70.10	69.95
	LoCal	79.93	73.26	70.14
	BiDeV	77.59	69.91	70.63
Debiasing methods	DagFC	73.15	62.76	58.67
	CICR	77.94	74.38	69.31
	CICR-graph	68.64	74.88	72.89
	CLEVER	80.99	77.71	71.51
	Causal Walk	67.14	75.09	74.19
	EGR-FV (ours)	82.94	84.07	83.05

Table 3: Macro-F1 on HOVER by reasoning depth. EGR-FV (last row) is grouped with the other debiasing methods and outperforms small-model classifiers, large-model verifier pipelines, and prior debiasers at every hop depth.

provided in Appendix F. For the evidence-contrast loss, E^- is sampled at each step uniformly from $\{\text{SHUFFLED}, \text{CORRUPTED}\}$ and $E^\emptyset$ is the empty string. All results are averaged over 5 random seeds and run on a single NVIDIA A6000 (48GB); a full EGR-FV training run (warm-up folds plus main training) completes in roughly 6–10 GPU-hours per dataset per seed.

5 Results and Analysis

5.1 Main Results

Tables 2 and 3 report accuracy on FEVER and PolitiHop, and macro-F1 on HOVER.

Standard / evidence-aware verifiers rely on claim shortcuts. The claim-evidence cross-

encoder, GEAR, CICR, and LisT5 score well on the original splits but collapse on the symmetric ones ($90.28 \rightarrow 70.75$ for claim-evidence on FEVER-sym; GEAR even falls to 44.91/51.17). A verifier that truly conditions on E should be evidence-stable across symmetric pairs; the collapse instead shows the model has absorbed the shortcut path $C \rightarrow Y$ rather than the intended $(C, E) \rightarrow Y$ reasoning.

Explicit debiasing helps. CLEVER, Causal Walk, and CICR outperform un-debiased baselines on symmetric splits (e.g. CLEVER-graph 88.42 vs. claim-evidence 70.75 on FEVER-sym), confirming that without an explicit counter-shortcut signal, dataset biases are silently absorbed.

Necessity-guided debiasing generalises across datasets. EGR-FV is the best method on six of seven evaluation columns, losing only FEVER-sym to CLEVER. Yet CLEVER’s FEVER-sym lead does not transfer: it falls to 42.40 on PolitiHop-sym (vs. our 63.45) and 71.51 on HOVER 4-hop (vs. our 83.05). Per-sample necessity instead produces a debiasing signal that generalises across single-hop, multi-hop, and shortcut-stressing regimes.

EGR-FV beats LLM pipelines on HOVER. EGR-FV outperforms every LLM-based compositional verifier at every hop depth—ProgramFC, PACAR, LoCal, BiDeV, and DagFC. A natural reading is that these pipelines suffer from knowledge conflict between the model’s parametric memory and the retrieved evidence (Wang et al., 2024; Mallen et al., 2023): the LLM brings priors about the claim and may answer from them rather than strictly from E , mirroring the claim-only shortcut EGR-FV explicitly suppresses.

Depth-stable as reasoning grows deeper. From 2- to 4-hop, several baselines drop monotonically (CLEVER $80.99 \rightarrow 71.51$, DeBERTaV3-NLI $77.22 \rightarrow 60.49$, DagFC $73.15 \rightarrow 58.67$), while EGR-FV stays essentially flat (82.94/84.07/83.05). Deeper claims demand more cross-evidence integration, so any residual claim-shortcut reliance is amplified—the flat curve indicates that EGR-FV’s decision is genuinely driven by the evidence chain.

5.2 Ablation Study

Table 4 ablates the four design choices of EGR-FV. Since the symmetric FEVER and PolitiHop splits directly stress shortcut robustness, we repeat the same ablation protocol on these splits in Appendix G (Table 8).

Variant	2-hop	3-hop	4-hop
EGR-FV (full)	82.94	84.07	83.05
<i>w/o evidence contrast</i> – $\mathcal{L}_{\text{ctr}}$	82.58	83.69	82.00
<i>w/o routing</i>			
Random remix-only	81.40	81.75	81.48
Hard routing (fixed weights)	81.14	82.88	80.62
In-sample routing	81.58	81.94	81.29
<i>w/o remix</i>			
Full w/o remix	80.12	81.19	80.58
<i>w/o routing & remix & contrast</i>			
claim-evidence	78.82	79.64	77.96
Two-branch joint	79.40	80.01	77.61

Table 4: Ablation study on HOVER (macro-F1).

Two-branch joint alone is not enough. Training p_g and p_s with equal weights and no routing, remix, or contrast lands at 79.40/80.01/77.61, on par with the vanilla CLAIM-EVIDENCE classifier; merely parameterising the shortcut path does not debias the model.

Remix carries the largest single share. Removing the remix sampler drops 2/3-hop by 2.82/2.88 points—the biggest gap of any single ablation—corroborating that rebalancing the training distribution toward high-necessity samples is the primary debiasing lever.

Routing must be guided and out-of-fold. All three routing variants—RANDOM REMIX-ONLY, HARD ROUTING, and IN-SAMPLE ROUTING—underperform the full pipeline by roughly 2 points, with IN-SAMPLE ROUTING confirming that out-of-fold scoring is necessary to prevent training bias from leaking into the routing decision.

Evidence-contrast loss adds a smaller but consistent gain. Removing $\mathcal{L}_{\text{ctr}}$ costs under 1 point on average, smaller than remix or routing but the only term that directly enforces sensitivity to evidence intervention.

5.3 Parameter Analysis: Sample Selection via r_i

Figure 3 partitions the FEVER test split into five r_i buckets and reports EGR-FV’s clean-correct, shuffled-wrong, and null-wrong rates within each. As r_i grows, the clean-correct rate rises sharply (from $\sim 9\%$ in the lowest bucket to $\sim 98\%$ in the highest), while the perturbation-induced error rates decrease overall. This confirms that r_i effectively isolates the samples on which evidence carries the verdict, validating its role as the routing and remix

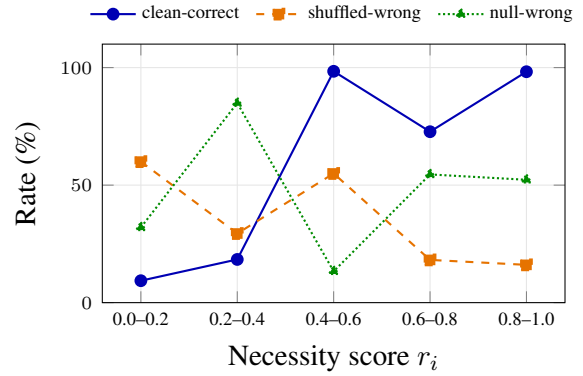


Figure 3: FEVER test samples partitioned by the necessity score r_i . Bucket counts $n = 75/272/5164/11/1144$. EGR-FV’s clean-correct, shuffled-wrong, and null-wrong rates are reported within each bucket.

Data	Method	Clean	Shuf.	Null	Wrong
HOVER	claim-evidence	81.75	40.26	40.99	39.86
	w/o contrast	82.97	44.33	44.14	44.71
	w/o remix	82.75	35.05	57.62	34.76
	EGR-FV	83.05	33.69	61.05	33.69
FEVER	claim-evidence	93.92	48.82	69.58	48.67
	w/o contrast	94.00	44.11	62.72	44.48
	w/o remix	93.78	38.25	72.13	38.11
	EGR-FV	94.09	40.22	76.71	39.59
PolitiHop	claim-evidence	46.73	46.73	46.73	46.73
	w/o contrast	55.19	50.05	64.78	49.43
	w/o remix	51.42	46.39	46.73	45.54
	EGR-FV	51.42	46.06	61.10	47.22

Table 5: Evidence intervention sensitivity (macro-F1). **Clean**: gold evidence. **Shuf./Wrong**: a real but mis-paired passage. **Null**: empty evidence. A faithful verifier should keep *Clean* high while *dropping* on Shuffled/Wrong and remaining largely insensitive to Null only if the claim is genuinely shortcut-decidable.

signal in EGR-FV.

5.4 Evidence Intervention Sensitivity

A genuinely grounded verifier should be sensitive to evidence interventions: its prediction should change in a principled way when the gold evidence is replaced, shuffled, or removed. To measure this directly, we evaluate four checkpoints (the vanilla CLAIM-EVIDENCE cross-encoder, EGR-FV without evidence contrast, EGR-FV without remix, and the full EGR-FV) on each test set under four evidence conditions: **Clean** (gold), **Shuffled** (evidence randomly re-assigned across samples, label-agnostic), **Null** ([NO_EVIDENCE] placeholder), and **Wrong** (evidence taken from a sample with a *different* gold label).

Data	Group	%	C-only	EGR-FV	Δ_{ERG}
HOVER	bias_easy	20.3	100.00	72.69	-27.31
	hard	52.2	79.45	80.80	+1.35
	grounded_needed	27.5	0.00	95.09	+95.09
FEVER	bias_easy	61.7	100.00	97.45	-2.55
	hard	21.0	77.45	80.97	+3.52
	grounded_needed	17.3	0.00	98.01	+98.01
PolitiH.	bias_easy	86.0	100.00	100.00	0.00
	hard	14.0	12.50	16.67	+4.17
	grounded_needed	0.0	-	-	-

Table 6: Per-routing-group accuracy at test time (%). % is the share of test instances; C-only is the claim-only model; Δ_{ERG} is the *Evidence Reliance Gain* (EGR-FV minus C-only).

Two patterns in Table 5 stand out. **EGR-FV is more evidence-sensitive than the baseline.** On HOVER, the CLEAN→SHUFFLED gap widens from claim-evidence (81.75 → 40.26) to EGR-FV (83.05 → 33.69), and the same trend holds for CLEAN→WRONG. Sensitivity to a real-but-wrong passage is a property the model must acquire, and EGR-FV’s evidence-contrast objective explicitly trains for it. **EGR-FV is less destabilised by Null evidence on shortcut-rich corpora.** On FEVER, the EGR-FV Null score of 76.71 is the highest of any method, and the CLEAN→NULL degradation is smaller than for claim-evidence (94.09 → 76.71 vs. 93.92 → 69.58). On the bias-easy subset of FEVER (61.7% of test instances) the claim alone is genuinely informative, so a well-routed verifier should not collapse when evidence is hidden and should fall back gracefully to the claim-only signal. The combination of high Clean, high Null, and low Shuffled is therefore the correct response surface, and only EGR-FV exhibits all three.

5.5 Evidence Reliance Across Routing Groups

To verify that the necessity score r_i is not just a heuristic, we evaluate the claim-only model and the full EGR-FV on each routing group at test time. The *Evidence Reliance Gain* $\Delta_{\text{ERG}} = \text{acc}_{\text{EGR-FV}} - \text{acc}_{\text{C-only}}$ directly estimates how much of EGR-FV’s accuracy on each subset comes from genuinely using the evidence.

Table 6 shows that routing behaves as intended. On bias_easy the claim-only model is already (near-)perfect, so Δ_{ERG} is non-positive; on grounded_needed the claim-only model achieves zero accuracy by construction, so $\Delta_{\text{ERG}} = +95.09$ (HOVER) and $+98.01$ (FEVER) directly quantify the evidence-driven accuracy that EGR-FV

recovers; on hard, both the claim-only and claim-evidence models struggle and EGR-FV produces a small but consistent positive gain (+1.35 on HOVER, +3.52 on FEVER, +4.17 on PolitiHop). PolitiHop has no grounded_needed group because its claim-evidence model rarely beats the claim-only model, consistent with PolitiHop being the most shortcut-rich of the three benchmarks. Three-seed significance tests, the correlation of r_i with per-sample evidence sensitivity, and a bucketed r_i -stratified analysis are reported in Appendix H.

5.6 Case Study

We highlight two HOVER cases that illustrate the two ends of EGR-FV’s behaviour; the full four-category breakdown with structured logs is in Appendix I.

Evidence-rescue ($r_i = 0.94$). The claim “*Groundhog Day* ... stars Bill Paxton” is surface-plausible and the claim-only model predicts SUPPORTS. The gold evidence reveals the lead is Bill Murray, not Bill Paxton; EGR-FV with clean evidence correctly outputs REFUTES (NLL 0.00) but flips back to wrong under null (1.61) or shuffled (1.08) evidence—confirming the verdict is decided by the evidence, not by claim wording.

Bias-easy trade-off ($r_i = 0.10$). A biographical claim about Eugene E. Lindsey is genuinely shortcut-decidable and claim-only gets it right; EGR-FV, being de-biased, no longer reads the claim-only signal as strongly and flips to incorrect on this rare bias-easy sample—a controllable cost of de-biasing, paid in exchange for the much larger grounded_needed gain.

6 Conclusion

We presented **EGR-FV**, a debiasing framework for fact verification that adapts the strength of debiasing to each sample’s actual reliance on evidence. After a brief warm-up of a grounded branch and a shortcut branch, EGR-FV summarises the disagreement between the two into a per-sample evidence necessity score, routes training examples into evidence-critical, hard, and bias-easy subsets, and remixes mini-batches at fixed proportions so that the verifier is steered toward genuinely evidence-grounded decisions while still preserving the valid claim-side signal that some claims legitimately carry on their own. Experiments on FEVER, PolitiHop, and the multi-hop HOVER benchmark, together with their symmetric counterparts, show that

EGR-FV simultaneously improves clean accuracy and shortcut robustness, with the largest gains on the symmetric and many-hop splits where shortcut learning is most damaging; ablations confirm that the necessity-guided routing and remix sampler are the primary contributors. We view per-sample necessity-guided debiasing as a general principle for evidence-grounded NLP, and plan to extend it to retrieval-augmented question answering and multi-modal claim verification in future work.

Limitations

(1) Compute overhead. EGR-FV requires $K = 5$ warm-up folds before the main training run, multiplying warm-up compute by K (each fold is shorter than a full run). **(2) Heuristic counterfactual evidence.** E^- is constructed by shuffling or sentence replacement, which approximates but does not exhaust meaningful evidence interventions. **(3) Gold-evidence evaluation.** All experiments use gold evidence; end-to-end evaluation jointly with retrieval is left to future work. **(4) English-only benchmarks.** Extending EGR-FV to non-English fact-verification corpora (e.g. CHEF (Hu et al., 2022)) is a natural next step.

Ethical Considerations

Intended use. EGR-FV is a defensive technology for automatic fact verification, aimed at curbing online misinformation (Vosoughi et al., 2018) by grounding verdicts in evidence rather than in claim-only shortcuts.

Potential risks of misuse. Erroneous SUPPORTS / REFUTES verdicts on sensitive claims could be cited as authoritative truth, so EGR-FV should be used as a decision-support tool for human fact-checkers rather than as a standalone arbiter.

Bias statement. The work contains no content that targets, stereotypes, or discriminates against any individual or group, and the experiments do not introduce any social, demographic, or political bias.

Data, licensing, and intended use. All experiments use publicly released benchmarks (FEVER, PolitiHop, HOVER, and their symmetric counterparts) under their original research licenses, and EGR-FV code and checkpoints will be released under the same research-only conditions.

Personally identifying and offensive content. The corpora are derived from English Wikipedia and public political fact-checking articles and may

reference public figures, but contain no private identifiers; we did not introduce additional personal data.

Use of AI Assistants

We used general-purpose AI assistants for language polishing and limited productivity support under author supervision. They were not used to generate research ideas, design experiments, make scientific decisions, or produce the technical claims, results, or conclusions of this paper. All such content is the authors’ own, and the authors take full responsibility for the correctness of the paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2025YFB3003700) and the National Natural Science Foundation of China (Grant No. 62476025).

References

- Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.
- Zhendong Chen, Siu Cheung Hui, Fuzhen Zhuang, Lejian Liao, Fei Li, Meihuizi Jia, and Jiaqi Li. 2022. Evidencenet: Evidence fusion network for fact verification. In *Proceedings of the ACM Web Conference 2022*, pages 2636–2645.
- Mitchell DeHaven and Stephen Scott. 2023. Bevers: A general, simple, and performant framework for automatic fact verification. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 58–65.
- Martin Fajcik, Petr Motlicek, and Pavel Smrz. 2023. Claim-dissector: An interpretable fact-checking system with joint re-ranking and veracity prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10184–10205.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip S. Yu. 2022. Chef: A pilot chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376.

- Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 402–410.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460.
- Shimon Kogan, Tobias J. Moskowicz, and Marina Niessner. 2019. [Fake news: Evidence from financial markets](#). *SSRN Electronic Journal*.
- Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheoneum Park, and Kyomin Jung. 2021. Crossaug: A contrastive data augmentation method for debiasing fact verification models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3181–3185.
- Yu Li, Bojie Hu, Fengshuo Zhang, Yahan Yu, Jian Liu, Yufeng Chen, and Jinan Xu. 2023. A multi-modal debiasing model with dynamical constraint for robust visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5032–5045.
- Qiang Liu, Junfei Wu, Shu Wu, and Liang Wang. 2024. [Out-of-distribution evidence-aware fake news detection via dual adversarial debiasing](#). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6801–6813.
- Yuxuan Liu, Hongda Sun, Wenya Guo, Xinyan Xiao, Cunli Mao, Zhengtao Yu, and Rui Yan. 2025. BiDeV: Bilateral defusing verification for complex claim fact-checking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 541–549.
- Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Jiatong Ma, Linmei Hu, Rang Li, and Wenbo Fu. 2025. LoCal: Logical and causal fact-checking with LLM-based multi-agents. In *Proceedings of the ACM Web Conference 2025*, pages 1614–1625.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9802–9822.
- Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-hop fact checking of political claims. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3892–3898.
- Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6981–7004.
- Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Faviq: Fact verification from information-seeking questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5166.
- Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3391–3401.
- Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643.
- Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.
- Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing NLU models via causal intervention and counterfactual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11376–11384.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.
- William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*.

Zhouhui Wu, Zhuohua Yang, Jiaojiao Jiang, Shuiqiao Yang, and Nan Sun. 2026. *DagFC: Dependency-aware fact-checking via claim-constructed knowledge graphs and large language models*. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining, WSDM '26*, pages 745–755. Association for Computing Machinery.

Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual debiasing for fact verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6777–6789.

Jun Yang, Yuhang Bai, Dandan Song, Zhijing Wu, and Yuhang Tian. 2026a. *A fact-checking framework with denoising evidence retrieval and LLM-based debate verification*. In *Proceedings of the ACM Web Conference 2026, WWW '26*, pages 2094–2104. Association for Computing Machinery.

Jun Yang, Dandan Song, Zhijing Wu, and Yuhang Tian. 2026b. *Dual-debias: A counterfactual inference framework for causally robust fact-checking*. *Neurocomputing*, 694:133901.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. 2024. Causal walk: Debiasing multi-hop fact verification with front-door adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19533–19541.

Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. 2024. PACAR: Automated fact-checking with planning and customized action reasoning using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 12564–12573.

Wanjuan Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. Gear: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901.

A Dataset Details

FEVER (Thorne et al., 2018) is a single-hop FV dataset built from Wikipedia, with original labels {SUPPORTS, REFUTES, NEI}; we follow the

2-class debiasing convention and drop NEI. We follow the standard splits and report accuracy on both the original test set and the symmetric test set of Schuster et al. (2019), which constructs paired samples that decorrelate claim wording from labels, breaking the claim-only shortcut.

PolitiHop (Ostrowski et al., 2021) is a multi-hop political fact-checking corpus drawn from PolitiFact. Its symmetric variant follows the protocol of Causal Walk (Zhang et al., 2024): the same claim is paired with different gold evidence so that the verdict flips with the evidence, isolating the contribution of evidence reasoning.

HOVER (Jiang et al., 2020) is a multi-hop Wikipedia FV benchmark explicitly stratified by reasoning depth (2-hop, 3-hop, 4-hop). Verdicts on deeper hops require integrating more claim–evidence interactions, making HOVER the most demanding split in our evaluation.

B Symmetric Test Settings

We use the symmetric splits only as held-out stress tests. For every method, training uses the original training split, early stopping and checkpoint selection use the original development split, and the symmetric test labels are never used for hyperparameter tuning, routing-threshold selection, or model selection. Evaluation follows the same input format as the original test sets: the model receives the claim and the gold evidence passage(s), and we report accuracy over the binary SUPPORTS/REFUTES labels.

FEVER-sym. We evaluate on the symmetric FEVER test set of Schuster et al. (2019). Because our experiments follow the 2-class FV setting, all NEI instances are excluded consistently with the original FEVER preprocessing. The resulting split contains 717 examples (379 SUPPORTS, 338 REFUTES). Each symmetric item is paired to reduce correlations between claim wording and label, so a method that succeeds by claim-only cues should transfer poorly even if it performs well on the original FEVER test set.

PolitiHop-sym. We use the symmetric PolitiHop protocol adopted by Causal Walk (Zhang et al., 2024). Each claim is evaluated with alternative gold evidence such that the supported/refuted verdict changes with the evidence, producing a balanced 342-example split (171/171). This construc-

tion is especially strict for evidence grounding: the surface claim can remain identical across paired examples, so the prediction must be determined by the evidence rather than by the claim prior.

C Debiasing Baseline Details

All RoBERTa-based debiasing baselines are compared under the same binary label space, gold-evidence input, and development-set model-selection protocol used for EGR-FV. When a method has both a sequence-only and graph-enhanced version, the sequence-only variant uses the same claim–evidence text input as our cross-encoder baselines, while the graph variant additionally encodes sentence/evidence nodes with the graph module specified by the original method. Symmetric test labels are held out for all baselines, matching the protocol in Appendix B.

CICR / CICR-graph. CICR (Tian et al., 2022) treats claim and evidence representations causally and applies intervention-style training to reduce spurious claim–label associations. We report both the RoBERTa cross-encoder version and the graph version, where the latter augments the evidence representation with graph aggregation before the intervention objective is applied.

CLEVER / CLEVER-graph. CLEVER (Xu et al., 2023) estimates the claim-only direct effect through counterfactual evidence masking/substitution and subtracts this shortcut effect from the final decision. The graph variant keeps the same debiasing objective but replaces the flat evidence encoder with a graph-based evidence encoder, making it the closest graph-enhanced counterpart to the sequence-only CLEVER baseline.

Causal Walk. Causal Walk (Zhang et al., 2024) performs a front-door-style adjustment for multi-hop FV by modelling evidence-mediated reasoning paths between the claim and verdict. We include it as the strongest prior causal debiasing baseline on PolitiHop-style symmetric evaluation, where the same claim can require different labels under different evidence.

D Necessity Score Distribution and Routing Statistics

For reproducibility we report the empirical distribution of the evidence necessity score r_i and the resulting routing group sizes on each dataset. With

thresholds $\tau_h = 0.7$ and $\tau_l = 0.3$, the three groups (grounded_needed, hard, bias_easy) cover the full training set without overlap; group shares on the test split are reported in Table 6. Per-dataset histograms of r_i are released with the code and routing caches.

Correlation table on PolitiHop. The PolitiHop correlations omitted from Table 10 are: warm-up indicator Δ_i^{ind} undefined (the grounded warm-up produces identical predictions with and without evidence on essentially all samples, so $\Delta_i^{\text{ind}} \equiv 0$); warm-up NLL Pearson +0.961, Spearman +0.045; EGR-FV indicator Pearson +0.284, Spearman +0.071; EGR-FV NLL Pearson +0.813, Spearman +0.093. The high Pearson but near-zero Spearman pattern reflects that a small number of very-confident samples dominate the NLL scale, while most samples cluster around the same rank, consistent with the observation that PolitiHop carries little evidence-decided signal in the first place.

E Counterfactual Evidence Construction

For the evidence-contrast loss we generate the wrong-evidence intervention E^- on the fly: at each training step E^- is sampled uniformly from two perturbation operators—**SHUFFLED**, which replaces e_i with a real-but-mis-paired evidence passage drawn from another sample in the same mini-batch, and **CORRUPTED**, which applies token-level corruption to e_i . Both yield a passage with the same domain and length distribution as e_i but break its semantic alignment with c_i , so the contrast is forced to come from semantics rather than from surface domain shift. The null intervention $E^\emptyset$ is the empty string.

For the analysis tables in Section 5.4, the **Shuffled** condition implements a more aggressive perturbation that re-assigns evidences across the *entire test split* (not just inside a batch); **Wrong** additionally constrains the re-assigned evidence to come from a sample with a *different* gold label, which gives a strictly harder evidence-sensitivity test.

F Hyper-parameter Sensitivity

Loss weights and routing thresholds. We performed a small grid over $(\lambda_s, \lambda_o, \lambda_c) \in \{0.1, 0.2, 0.3, 0.5\}^3$ on the HOVER development set; final macro-F1 varied within ± 0.5 over the central region of the grid, indicating that EGR-FV is robust to these weights.

Weights ($\alpha_1, \alpha_2, \alpha_3, \alpha_4$)	dev (ep)	test
(0.40, 0.30, 0.20, 0.10) (<i>default</i>)	0.8468 (4)	0.8305
(0.30, 0.35, 0.25, 0.10) weaken α_1	0.8466 (5)	0.8300
(0.45, 0.20, 0.25, 0.10) weaken α_2	0.8403 (5)	0.8276
(0.25, 0.25, 0.25, 0.25) uniform	0.8433 (5)	0.8250

Table 7: Sensitivity of the necessity-score weights on HOVER. Macro-F1 on the dev (best epoch) and test splits.

The routing thresholds (τ_h, τ_l) were swept on $\{(0.65, 0.25), (0.70, 0.30), (0.75, 0.35)\}$ with negligible effect; we report the middle setting throughout.

Necessity-score weights ($\alpha_1, \alpha_2, \alpha_3, \alpha_4$). We additionally swept the four weights in the necessity-score definition (Eq. 3) on the HOVER development set, perturbing each component by ± 0.1 while keeping the weights summed to one. Table 7 reports HOVER dev (best epoch) and test macro-F1 for four representative configurations.

The default configuration is best on both dev and test, and a ± 0.1 perturbation of the dominant indicator α_1 is essentially harmless (-0.0002 dev / -0.0005 test), indicating that routing is already saturated and the framework is locally insensitive to small weight changes. Reducing the shortcut-failure weight α_2 from 0.30 to 0.20 shifts roughly 120 samples from hard into bias_easy and causes the largest dev drop (-0.0065); the test drop (-0.0029) is smaller but visible, confirming that the “shortcut predicts incorrectly” signal is the second most informative ingredient. Setting all four weights to 0.25 shrinks the bias_easy group by about 16% and produces the largest test drop (-0.0055); flattening the weights weakens the prior that some signals are more reliable than others, exactly the prior that the ordered (0.4, 0.3, 0.2, 0.1) encodes. Overall, the framework is robust within a neighborhood of the default weights, and the ordered weighting in Eq. 3 is a meaningful, not arbitrary, design choice.

G Ablation on Symmetric Splits

Table 8 complements the HOVER ablation in Table 4 by evaluating the same model variants on the two symmetric test splits. These splits are designed to break claim-only label shortcuts, so they provide a direct stress test of whether each component helps the verifier rely on evidence rather than surface claim cues.

Variant	FEVER-sym.	PolitiHop-sym.
EGR-FV (full)	83.26	63.45
<i>w/o evidence contrast</i> – $\mathcal{L}_{ctr}$	82.81	59.65
<i>w/o routing</i> Random remix-only	79.50	63.16
Hard routing (fixed weights)	82.67	59.36
In-sample routing	81.87	56.73
<i>w/o remix</i> Full w/o remix	79.64	55.46
<i>w/o routing & remix & contrast</i> claim-evidence	79.22	50.00
Two-branch joint	79.50	55.17

Table 8: Ablation study on the symmetric FEVER and symmetric PolitiHop splits (accuracy).

The full EGR-FV model achieves the best performance on both symmetric splits, indicating that the components are complementary under shortcut-stressing evaluation. On FEVER-sym, removing remix or replacing necessity-guided routing with random remixing causes a clear drop (83.26 to 79.64/79.50), showing that sample selection remains important even when evaluation pairs are explicitly balanced. On PolitiHop-sym, removing remix is especially harmful (63.45 to 55.46), and in-sample routing also degrades sharply (56.73), which supports the need for out-of-fold necessity estimation. The evidence-contrast loss contributes a smaller but consistent gain on both datasets, while the claim-evidence and two-branch joint baselines confirm that simply adding evidence input or a second branch is insufficient without the routing, remixing, and contrastive objectives.

H Additional Analyses

This appendix collects supporting analyses that complement Section 5.5: a three-seed significance test, the correlation of the necessity score r_i with per-sample evidence-sensitivity, and an r_i -bucketed intervention-sensitivity analysis on HOVER.

H.1 Statistical Significance Across Seeds

To rule out seed variance as an explanation for the gains, we repeat the four most informative methods (CLAIM-EVIDENCE, TWO-BRANCH JOINT, FULL W/O REMIX, and FULL EGR-FV) across three random seeds {13, 42, 2024} on the splits that probe shortcut robustness (FEVER-sym, PolitiHop-sym, HOVER macro-F1). For each method we report

Method	FEVER-s.	PolitiH.-s.	HOVER
claim-evidence	79.9 $\pm$ 1.6	50.0 $\pm$ 0.0	82.96 $\pm$ 0.6
Two-branch joint	78.8 $\pm$ 1.0	55.2 $\pm$ 3.3	84.06 $\pm$ 0.8
Full w/o remix	80.8 $\pm$ 2.0	55.5 $\pm$ 6.3	83.73 $\pm$ 0.4
Full EGR-FV	83.3$\pm$1.7	59.9$\pm$5.7	83.89$\pm$0.6
Δ vs. baseline	+3.42	+9.94	+0.93
p -bootstrap	$< 10^{-4}$	$< 10^{-4}$	2×10^{-4}
p -randomization	$< 10^{-4}$	$< 10^{-4}$	2×10^{-4}

Table 9: Three-seed mean $\pm$ std (FEVER-s. / PolitiH.-s.: accuracy; HOVER: macro-F1) and paired bootstrap / randomization p-values of Full EGR-FV vs. the claim-evidence baseline. p-values are Fisher-combined across seeds.

mean $\pm$ standard deviation and two paired significance tests against CLAIM-EVIDENCE on the same seed: a paired bootstrap over sample ids (10 000 resamples, percentile p-value centred at the observed effect) and an approximate randomization test (10 000 method/baseline label swaps). Per-seed p-values are combined with Fisher’s method.

Table 9 confirms that all three gains are statistically significant at $p < 10^{-3}$ even after combining across seeds. On HOVER the symmetric-style stress is implicit in the multi-hop split, so the absolute Δ is smaller (+0.93) than on FEVER-sym, but the bootstrap p-value remains $\leq 2 \times 10^{-4}$.

H.2 Correlation of r_i with Evidence Sensitivity

For each sample we measure how much adding evidence helps the model, in two forms:

$$\Delta_i^{\text{ind}} = \mathbb{1}[f(c_i, e_i) = y_i] - \mathbb{1}[f(c_i, \emptyset) = y_i],$$

$$\Delta_i^{\text{NLL}} = \ell(f(c_i, \emptyset), y_i) - \ell(f(c_i, e_i), y_i),$$

and report Pearson and Spearman correlations between r_i and Δ_i .

Across HOVER and FEVER, r_i correlates consistently with both forms of Δ_i , and the correlation is preserved (or strengthened, in the FEVER case) by the full EGR-FV checkpoint: the model EGR-FV trains is not merely accurate on average but *accurate on the samples its routing score flags as evidence-needed*, the operational definition of an evidence-grounded predictor. On PolitiHop the indicator Δ_i^{ind} is degenerate for the warm-up because $f(c, e)$ and $f(c, \emptyset)$ produce identical predictions on virtually all examples; the NLL form recovers a positive correlation but is dominated by a few high-confidence samples (PolitiHop numbers reported in Appendix D).

Data	Model / Δ	Pearson r	Spearman ρ
HOVER	warm-up, indicator	+0.517	+0.581
	warm-up, NLL	+0.490	+0.567
	EGR-FV, indicator	+0.420	+0.453
	EGR-FV, NLL	+0.479	+0.623
FEVER	warm-up, indicator	+0.188	+0.260
	warm-up, NLL	+0.218	+0.314
	EGR-FV, indicator	+0.424	+0.393
	EGR-FV, NLL	+0.418	+0.538

Table 10: Correlation between the necessity score r_i and the per-sample evidence-sensitivity Δ_i for the grounded warm-up and the full EGR-FV. All listed correlations are significant at $p < 10^{-50}$ except Δ_i^{ind} on FEVER warm-up, which is significant at $p < 10^{-54}$.

I Case Study (Full Table)

Table 11 expands the case study of Section 5.6 to four representative HOVER cases, one per behavioural signature of EGR-FV. Each row reports the gold label, the claim-only prediction, the claim-evidence prediction, and EGR-FV under three evidence conditions: clean (gold), null ([NO_EVIDENCE]), and shuffled (a real but mis-paired passage); NLL values are on the gold label.

(A) Evidence-rescue. The *Groundhog Day* claim is surface-plausible and the claim-only model predicts SUPPORTS; EGR-FV under clean evidence correctly outputs REFUTES (NLL 0.00) and reverts to wrong under null (NLL 1.61) or shuffled (NLL 1.08) evidence, confirming that the verdict is decided by the evidence rather than by claim wording.

(B) Evidence-sensitive. Both claim-only and claim-evidence err on the Shughart claim due to its surface plausibility; only EGR-FV produces REFUTES under clean evidence (NLL 0.03) and degrades sharply under null (NLL 3.27). The +3.24 NLL gap directly visualises evidence reliance.

(C) Bias-easy trade-off. A biographical claim about Eugene E. Lindsey is genuinely shortcut-decidable: claim-only happens to predict the correct REFUTES from claim wording alone, while EGR-FV—deliberately de-biased—no longer trusts that signal and predicts SUPPORTS. This is a controllable cost of de-biasing, traded for the much larger grounded_needed gain documented in Table 6.

(D) Hard-bucket win. The Iqaluit Airport claim is genuinely multi-hop. Both claim-only and claim-evidence predict SUPPORTS, but EGR-FV under clean evidence flips to the correct REFUTES; null/shuffled revert the prediction, ruling out mem-

Category	Claim	Gold evidence (truncated)	r_i	Gold	c-only	c+e	EGR-FV (C/N/S)	NLL (C/N/S)
A. Evidence-rescue	The movie <i>Groundhog Day</i> was directed by Harold Ramis and stars Bill Paxton.	<i>Groundhog Day</i> is a 1993 American fantasy-comedy film directed by Harold Ramis, starring Bill Murray, Andie MacDowell, and Chris Elliott...	0.94	refutes	sup. ✗	ref. ✓	✓ / ✗ / ✗	0.00 / 1.61 / 1.08
B. Evidence-sensitive	Randy Shughart's father was Sergeant First Class when he died.	Randall David "Randy" Shughart was a United States Army soldier of 1st SFOD-D / Delta Force, posthumously awarded the Medal of Honor for Battle of Mogadishu...	0.45	refutes	sup. ✗	sup. ✗	✓ / ✗ / ✗	0.03 / 3.27 / 1.13
C. Bias-easy trade-off	Eugene E. Lindsey, born in Sprague, WA, 2 July 1905, graduated from the US Naval Academy in 1927; served with an observation squadron...	USS <i>Maryland</i> (BB-46), also known as "Old Mary", was a <i>Colorado</i> -class battleship, commissioned in 1921...	0.10	refutes	ref. ✓	sup. ✗	✗ / ✗ / ✗	3.34 / 0.78 / 1.04
D. Hard-bucket win	Iqaluit Airport and Canadian North are based out of Montreal.	Iqaluit Airport (CYFB) serves Iqaluit, Nunavut, Canada; operated by the government of Nunavut. Hosts scheduled service from Ottawa, Montreal, Rankin Inlet, Kuujuaq...	0.40	refutes	sup. ✗	sup. ✗	✓ / ✗ / ✗	0.28 / 1.89 / 1.13

Table 11: Qualitative case study on HOVER. C/N/S = clean / null / shuffled evidence; ✓/✗ = correct/incorrect; NLL = negative log-likelihood on the gold label.

orisation.