Multi-Channel Mitigation of Source-Trust Shortcuts in Fact-Checking RL Agents

Anonymous ACL submission

Abstract

Reinforcement-learning agents for retrieval-augmented claim verification exhibit a *source shortcut*: verdicts flip when source-trust labels on otherwise identical evidence are perturbed, violating the design principle that *trust never decides verdict*. We make three contributions. First, we decompose the shortcut along three output channels (verdict shift, signed confidence shift, and search-decision shift) and show that the no-calibration baseline already responds in the principled direction on confidence and search, so the principle violation lives on the verdict channel alone. Second, we re-evaluate two previously reported single-seed mitigation claims under five-seed paired bootstrap and find both indistinguishable from zero, refuting them as variance-aware results. Third, we show that counterfactual augmentation and a verdict-consistency reward act on orthogonal channels (augmentation reduces verdict shifts but flattens directional calibration, while the consistency reward sharpens calibration without reducing verdict shifts), and compose them in **Hybrid Aug+ATF**, pairing two GRPO views for per-token label supervision with a gated cross-group reward. Across nine recipes spanning prior single-seed claims and new training mechanisms, the hybrid achieves a statistically significant verdict-shift reduction on the pre-registered primary cell ($\Delta_{\text{vss}} = -1.60$ pp, 95% CI $[-2.9, -0.2]$, $P(\Delta < 0) = 0.988$), the largest of any arm we evaluate, with accuracy within ± 1 pp of the baseline.[1]

1 Introduction

Fact-checking benchmarks span a wide range of claim and evidence styles (Wang, 2017; Augenstein et al., 2019; Thorne et al., 2018; Jiang et al., 2020; Aly et al., 2021; Chen et al., 2020; Schuster et al., 2021; Wadden et al., 2020; Schlichtkrull et al., 2023; Petroni et al., 2021), and modern systems address them with retrieval-augmented language models (Chen et al., 2017; Karpukhin et al., 2020; Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021; Izacard et al., 2023; Asai et al., 2024) trained end-to-end with reward shaping (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Bai et al., 2022; Ouyang et al., 2022; Schulman et al., 2017; Shao et al., 2024). Recent fact-checking agents adopt this pattern, combining reward terms for verdict correctness, format adherence, and evidence-related signals such as source-trust labels (He et al., 2025; Jin et al., 2025). The intended role of trust is calibration: a low-trust source should lower the model's confidence and trigger additional search, but it should not change the content-based verdict. We refer to this design principle as *trust never decides verdict*.

Current systems violate this principle. A recent diagnostic on the RL fact-checking agent Veri-R1 (He et al., 2025) shows that under perturbation of source-trust labels (HIGH $\leftrightarrow$ LOW), the verdict flips on roughly one in six high-confidence outputs; we call this failure the *source shortcut*. A single-seed result reported that an alternative calibration-based reward reduced the shift rate from 17.7% to 3.8%. We replicate the comparison at five seeds with paired bootstrap, and the effect does not survive: the 95% CI of Δ_{vss} vs. a no-calibration baseline includes zero on every cell, and the same holds for a follow-up "denser unwarranted-decision reward" from the same lineage. The phenomenon is real; the obvious interventions are not.

The model emits a structured output (verdict, confidence, search flag, reasoning trace), and the trust signal can leak into any of these channels. Stratifying baseline outputs by trust-conflict structure, we find that the no-calibration variant *already* produces the directionally appropriate confidence response: confidence falls when trust degrades, rises when it improves, and stays balanced when

[1] Code and artifacts (anonymous): https://anonymous.4open.science/r/factcheck-pilot-CB0F

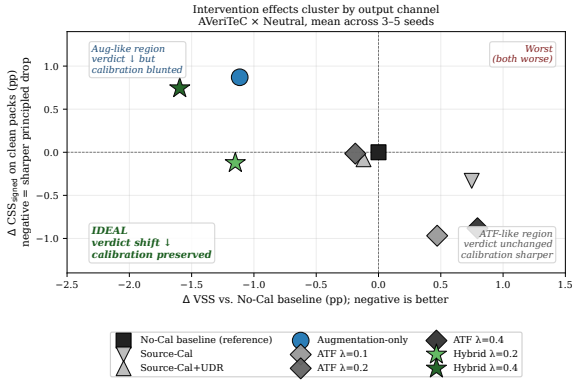

Figure 1: Intervention effects cluster by output channel on the pre-registered primary cell (AVeriTeC × Neutral). The x-axis is Δ_{VSS}, the change in verdict-shift rate versus the No-Cal baseline; the y-axis is $\Delta_{\text{CSS}_{\text{sgn}}}$ on `clean` packs, the change in the (principled) signed confidence drop. Counterfactual augmentation moves south-east: it reduces verdict shifts but blunts directional calibration. Verdict-only consistency rewards (ATF) move north-east: they sharpen calibration but do not reduce verdict shifts. The hybrid recipe at λ=0.4 is the only intervention that reaches the south-east–ideal quadrant: reduced verdict shifts with substantially less calibration blunting than augmentation alone.

the perturbation preserves total trust. The principle violation is concentrated on the verdict channel. An effective intervention should therefore reduce verdict shifts without erasing the calibrated directional behaviour the baseline already produces.

Figure 1 makes this concrete. Counterfactual augmentation (Kaushik et al., 2020) reduces verdict shifts via implicit invariance learning but flattens the principled directional response on confidence. An explicit verdict-consistency reward, gated to prevent the model from collapsing to NEI on every input, sharpens directional calibration but does not reduce verdict shifts. The two interventions act on different output channels through different gradient mechanisms, occupying distinct quadrants of Figure 1.

We combine the two interventions in a hybrid recipe: separate-view GRPO groups give each view a clean per-token label gradient, and a gated cross-group consistency reward injects the calibration signal at the group level without contaminating the verdict gradient. All comparisons are made via three-seed paired bootstrap against a pre-registered primary metric. On that metric, the hybrid achieves the largest variance-aware verdict-shift reduction we observe among nine intervention recipes across three domains: all three seeds favour the method, the 95% CI excludes zero, and accuracy is held within ± 1 pp of the baseline.

Our contributions are:

- **A multi-channel decomposition of the source shortcut** (VSS, CSS_{sgn}, SDS) with a directional analysis that separates principled calibration from pathological displacement (§4; Figures 1, 4).
- **A variance-aware reproducibility audit** that refutes two previously reported single-seed mitigation claims and reports a null result for standalone verdict-consistency rewards under five-seed paired bootstrap (§4.4; Table 1).
- **Hybrid Aug+ATF**, a composed training recipe that achieves the largest variance-aware verdict-shift reduction among nine evaluated interventions on a pre-registered primary metric ($\Delta_{\text{VSS}} = -1.60$ pp*, $P(\Delta < 0)$=0.988; §5–§6; Table 2), while preserving directional calibration that augmentation alone erases.

2 Related Work

Shortcut learning and counterfactual mitigation. Models exploiting spurious correlations rather than the intended task structure is a documented failure mode across vision, NLI, and reading comprehension (Geirhos et al., 2020; Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019; Jia and Liang, 2017; Nie et al., 2020). Counterfactual and adversarial perturbations are standard tools for diagnosing and mitigating such failures (Goodfellow et al., 2015; Kaushik et al., 2020): by exposing the model to both an original example and a minimally perturbed counterfactual with the same gold label, the gradient promotes invariance along the perturbation axis. In fact-checking, prior work has focused on shortcuts arising from claim–evidence lexical overlap or stylistic markers. The source-trust shortcut we study is closer in spirit to trust-label perturbations used in NLI debiasing, but is targeted at a calibration signal that should also legitimately influence other output channels; the standard augmentation recipe does not respect this distinction.

Calibration and abstention. Confidence calibration asks whether a model's stated probability corresponds to empirical correctness (Guo et al., 2017; Lakshminarayanan et al., 2017); selective prediction and out-of-distribution detection extend this idea by allowing a model to abstain, reject uncertain examples, or flag inputs likely to be unreliable

under the training distribution (Geifman and El-Yaniv, 2017, 2019; Hendrycks and Gimpel, 2017). Our setting differs because trust is not merely an uncertainty signal: it should affect confidence and search behaviour while leaving the content-based verdict invariant. This makes source trust a multi-channel signal rather than a scalar calibration target.

Reinforcement learning for retrieval-augmented fact verification. Recent fact-checking agents combine a language model with an explicit search component, trained end-to-end with RL (He et al., 2025; Jin et al., 2025). These systems use PPO or related algorithms with reward shaping over format adherence, verdict correctness, and source signals (Ouyang et al., 2022; Schulman et al., 2017; Shao et al., 2024). Our baseline is a no-calibration variant of this recipe trained via Group Relative Policy Optimisation (Shao et al., 2024); the interventions we evaluate add either counterfactual augmentation, an explicit verdict-consistency reward, or both.

Variance reporting in RL. The unreliability of single-seed RL claims is well established (Henderson et al., 2018; Agarwal et al., 2021), and we follow Agarwal et al. (2021) in reporting 95% paired-bootstrap intervals over seeds rather than p-values from arbitrary significance thresholds. The asymmetry in our seed counts (five for baselines, three for new arms) is honestly reflected in CI widths and noted in the Limitations.

3 Background

Task and output schema. Given a claim c and a retrieved evidence pack $E = \{(t_i, s_i, \tau_i)\}_{i=1}^{k}$, where t_i is evidence text, s_i is source identity, and $\tau_i \in \{\text{HIGH}, \text{MEDIUM}, \text{LOW}, \text{UNKNOWN}\}$ is a trust label, the agent emits a structured output (v, p, m, r): verdict $v \in \{\text{SUPPORTS}, \text{REFUTES}, \text{NEI}\}$, scalar confidence $p \in [0, 1]$, search-decision flag $m \in \{0, 1\}$, and a free-text reasoning trace r.

Trust-flipped probe. For each evaluation instance we compute two outputs: one on the original evidence pack and one on a perturbed pack where HIGH and LOW labels are swapped (MEDIUM and UNKNOWN remain unchanged). Following He et al. (2025), we restrict analysis to *activated* pairs, in which $\max(p_o, p_f) \geq 0.7$; this is the regime where the model has committed to a confident decision, and where verdict shifts have practical consequence.

Training setup. All arms train Qwen3-4B-Instruct-2507 (Team, 2025) with LoRA of rank 32 (Hu et al., 2022) via Group Relative Policy Optimisation (Shao et al., 2024) using the Tinker hosted RL service. Arms share all hyperparameters and the same 311-pair training pool drawn from CONFACT and SciFact; the full configuration is given in Appendix B.

Evaluation. We evaluate on three diagnostic sets: AVeriTeC (Schlichtkrull et al., 2023) ($n = 1968$, web fact-checking), SciFact (Wadden et al., 2020) ($n = 600$, scientific abstracts), and CONFACT ($n = 1200$, a contrastive set we construct with controlled HIGH/LOW trust-conflict variants over mixed-source evidence). Each domain is run with two prompt variants (Neutral and Principled). Effect-size estimates are computed by paired bootstrap with 10,000 iterations, sampling seeds with replacement within each arm; we report the 95% interval and the empirical probability that the effect is negative. Baselines use five seeds and new arms use three; the asymmetry is reflected in CI widths and noted in the Limitations.

4 Decomposing the Source Shortcut

4.1 A three-channel diagnostic

We measure the source shortcut along three output channels. The *Verdict Shift Score* (VSS) is the proportion of activated pairs whose verdict flips under trust perturbation, $\Pr(v_o \neq v_f \mid \text{activated})$. The *Confidence Shift Score* (CSS) summarises the change in scalar confidence and admits two natural definitions: an unsigned magnitude $\mathbb{E}[|p_f - p_o|]$ and a signed mean $\mathbb{E}[p_f - p_o]$, which we denote CSS_{sgn}. The *Search-Decision Shift* (SDS) measures the proportion of activated pairs whose search flag flips, with an analogous signed variant. A fourth potential channel, free-text reasoning shift, requires sentence-embedding infrastructure that we do not compute here.

This decomposition is informative for two reasons. First, the design principle (trust should affect calibration and search but not the verdict) prescribes a directional pattern: under HIGH→LOW perturbation, confidence should decrease and search should activate; under LOW→HIGH, the inverse should hold. A magnitude-only metric cannot distinguish this principled response from patholog-

ical displacement. Second, the channels respond differently to different interventions, a property we exploit in the method design (§5).

4.2 Intervention arms evaluated

We evaluate four families of training interventions, all sharing the same Qwen3-4B-Instruct base, the same training data, and the same GRPO hyperparameters; only the reward configuration and dataset routing differ:

- **No-Cal baseline**: format, label-correctness, and evidence-coverage rewards only; no trust-related shaping. Five seeds.
- **Source-Cal**: the No-Cal reward plus four calibration terms drawn from prior work (trust behaviour, calibration coverage, soft entailment lock, and source-shortcut penalty). Five seeds.
- **Source-Cal+UDR**: Source-Cal plus an additional denser unwarranted-decision reward shape for low-confidence abstention. Five seeds.
- **Augmentation-only**: No-Cal reward applied to a training stream that includes both the original and trust-flipped views of each example as independent training instances. Three seeds.
- **ATF** λ: No-Cal reward plus a gated cross-rollout verdict-consistency reward at weight $\lambda \in \{0.1, 0.2, 0.4\}$. Three seeds per λ.
- **Hybrid** λ: augmentation structure combined with the gated consistency reward at weight $\lambda \in \{0.2, 0.4\}$. Three seeds per λ.

Source-Cal and Source-Cal+UDR reproduce two previously reported single-seed claims, which we re-evaluate in §4.4. The remaining arms (augmentation, ATF, and the hybrid recipe) are the focus of our method analysis and are described in detail in §5.

4.3 The baseline is already directionally principled

To verify that the directional decomposition carries empirical content, we stratify the No-Cal baseline by the trust-conflict structure of each example, using the standard taxonomy of `clean` (mostly HIGH evidence), `low_trust_only`, and `high_low_conflict` packs.

On AVeriTeC under the Neutral prompt, the baseline produces a CSS_{sgn} of -7.28 pp on `clean` packs (confidence appropriately falls when HIGH evidence is degraded to LOW), $+5.96$ pp on `low_trust_only` packs (confidence appropriately rises when LOW is elevated to HIGH), and $+0.60$ pp on `high_low_conflict` packs (no single direction expected). The search channel shows the same direction more weakly: the baseline triggers more search on `low_trust_only` pairs than on `clean` by $+3.6$ pp (33.7% vs 30.1%). Both patterns are consistent across all five seeds and three domains.

The No-Cal baseline thus already produces the directional response the principle prescribes on both confidence and search; the principle violation is concentrated on the verdict channel. This narrows the design space: a mitigation should reduce verdict shifts without erasing the calibrated behaviour the baseline already produces.

Claim or standalone test	Variance-aware Δ [95% CI]
Source-Cal reduces VSS	
AVeriTeC $\times$ Neutral	$+0.8$ $[-1.1, +2.0]$
SciFact $\times$ Neutral	-0.3 $[-2.5, +1.7]$
Source-Cal+UDR reduces UDR by 10pp	
AVeriTeC $\times$ Neutral	$+2.0$ $[-6.8, +8.6]$
SciFact $\times$ Neutral	$+2.9$ $[-3.4, +7.9]$
Standalone ATF reduces VSS	
AVeriTeC $\times$ Neutral, $\lambda = 0.4$	$+0.8$ $[-0.6, +2.1]$

Table 1: Variance-aware null checks motivating the hybrid recipe. Source-Cal and Source-Cal+UDR are five-seed reproductions of prior single-seed claims; standalone ATF is a three-seed verdict-consistency test. The 95% paired-bootstrap intervals include zero for every reported cell.

4.4 Prior mitigation claims do not replicate

Before exploiting the multi-channel decomposition for new method design, we re-examine the existing interventions in this lineage. The two previously reported single-seed improvements (a Source-Calibrated objective and a denser unwarranted-decision reward) do not survive five-seed variance evaluation. We also test the standalone verdict-consistency reward at three seeds; it produces no statistically detectable change in VSS, although it does affect the confidence channel (§4.5). Table 1 reports the paired-bootstrap interval for each null check. In every reported cell, the 95% interval includes zero, and the point estimate is on the same order as the seed-to-seed standard deviation of the metric. The source shortcut is real, but the obvious interventions in this lineage do not pass standard variance reporting.

4.5 Interventions affect orthogonal channels

We next examine the two natural interventions on the multi-channel decomposition. Counterfactual augmentation, which trains the model on both the original and trust-flipped views of each example as independent instances, reduces VSS by -1.12 pp*on the primary cell. The mechanism is implicit: per-token cross-entropy gradients on the same gold label across both views push the policy toward verdict invariance. The same gradients, however, also flatten the principled directional confidence response. Aggregating activated pairs, CSS_{sgn} moves from -4.11 pp (baseline) to -3.45 pp (augmentation), a statistically significant attenuation; on the `clean` stratum used in Figures 1 and 4, the corresponding shift is from -7.28 pp to -6.41 pp.

An explicit verdict-consistency reward, attached as a group-level bonus over matched original–flipped rollouts and gated to prevent NEI collapse, does the opposite. At λ=0.1 and λ=0.4 the reward produces no statistically detectable change in VSS, but it sharpens the directional response: at λ=0.1, CSS_{sgn} on `clean` packs becomes more negative by -0.97 pp*, and on `low_trust_only` packs more positive by $+1.53$ pp*. The trust signal is not eliminated under the consistency reward; it is redirected from the verdict channel to a more principled one.

The search channel is largely unperturbed: the principled `low_trust_only`-over-`clean` gap from §4.3 holds within ±1 pp across all arms. The orthogonality story therefore plays out between the verdict and signed-confidence channels.

Mechanistically, augmentation's per-token label supervision suppresses verdict shifts but indiscriminately damps any per-token sensitivity to trust labels, including the principled response. The consistency reward's group-level signal preserves and reinforces calibration, but its per-token contribution is too weak to override the model's verdict prior. A natural design move is to combine the two and ask whether their effects add on their respective channels.

4.6 Where the shortcut lives

A complementary stratification by `conflict_type` indicates where the residual shortcut concentrates. On CONFACT $\times$ Neutral, where the stratum sizes permit a fine-grained comparison, the VSS reduction across all four mitigation arms concentrates on `high_low_conflict` packs (in which HIGH and LOW evidence directly compete): Δ_{VSS} ranges from -1.50 pp*to -1.94 pp*across augmentation and the three ATF arms. The aggregate effect is diluted by the larger `clean` stratum, on which the principle violation is weaker and the intervention has less room to operate. The source shortcut, in other words, is concentrated where the trust signal most actively competes with the content signal, consistent with the multi-channel interpretation.

5 Method

The observations in §4.5 motivate a training recipe that combines two complementary signals on the GRPO learner: per-token label supervision from counterfactual augmentation, which acts on the verdict channel, and a cross-view consistency reward, which acts on the confidence channel. The components are not new individually; the design contribution is the architectural arrangement that lets both signals operate on their respective channels without erasing one another. Figure 2 summarises the three stages.

Trust-label flip and gated reward. The perturbation ϕ swaps HIGH and LOW labels in the evidence pack and leaves MEDIUM and UNKNOWN unchanged; text and source identity are untouched. For each matched rollout pair (ρ_o, ρ_f) with verdicts (v_o, v_f) and gold g, the consistency reward R_{cons} is $+1$ if $v_o = v_f = g$, -1 if $v_o \neq v_f$ and both verdicts are canonical, and 0 otherwise (same-but-wrong or malformed). The middle case is the gating: an ungated ±1 rule would reward a policy that emits NEI on every input ($+1$ for trivial agreement); the gate sends that case to zero.

Separate-view GRPO with shared-state pairing. For each training example whose `conflict_type` contains at least one HIGH or LOW label, the dataset emits two GRPO groups of G=8 rollouts each, one on E and one on $\phi(E)$. The two groups stay separate for GRPO advantage normalisation, which is what produces the augmentation effect: each view's gradient is computed against its own group mean, yielding a clean per-view label signal toward verdict invariance. The two groups exchange information once both have finished their rollouts: the second-to-complete group computes R_{cons} over the matched pairs and applies $\lambda_{\text{atf}} R_{\text{cons}}$ as a group-level bonus. The asymmetry within a step averages out across steps, since the two views share a single LoRA adapter.

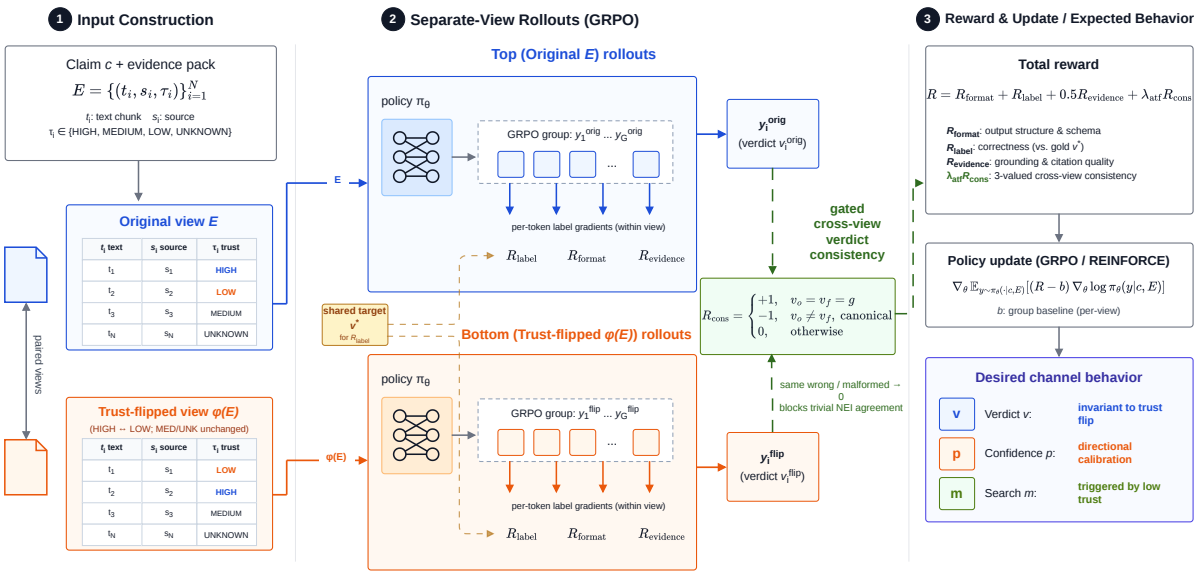

Figure 2: Hybrid Aug+ATF overview. (1) Input construction: each claim is paired with two views of the evidence pack, the original E and a trust-flipped $\phi(E)$ in which HIGH and LOW labels are swapped while MEDIUM and UNKNOWN remain unchanged (text and source identity unchanged). (2) Separate-view rollouts: G=8 GRPO rollouts on each view, with advantage normalised within view so each side contributes a clean per-token label gradient against the shared gold v^* (the augmentation effect). The two views exchange the gated cross-view consistency reward R_{cons}: $+1$ if $v_o = v_f = g$, -1 if $v_o \neq v_f$ and both verdicts are canonical, and 0 otherwise (same-but-wrong or malformed), preventing trivial NEI agreement from being rewarded. (3) Reward and update: total reward is the No-Cal baseline plus $\lambda_{\text{atf}} R_{\text{cons}}$; the policy is updated by GRPO with a per-view group baseline. The intended channel-level behaviour is verdict invariance under trust flip, directional confidence calibration, and low-trust-elicited search.

Reward and sweep. The per-rollout reward is $R = R_{\text{format}} + R_{\text{label}} + 0.5\,R_{\text{evidence}} + \lambda_{\text{atf}}\,R_{\text{cons}}$: the No-Cal baseline plus the gated cross-view bonus. All prior calibration components ($\lambda_{\text{trust_behavior}}, \lambda_{\text{calibration}}$, etc.) are set to zero. We sweep $\lambda_{\text{atf}} \in \{0.2, 0.4\}$ at three seeds each; the full configuration and compute footprint are in Appendix B.

6 Experiments

6.1 Pre-registered primary metric

The pre-registered primary metric is Δ_{VSS} versus the No-Cal baseline on AVeriTeC × Neutral, computed by paired bootstrap with seed-level resampling. Baseline arms use five seeds; new arms use three; the two pools are resampled independently. The pre-registration also fixes a stopping rule on a smaller Tier-1 smoke run before any sweep is fired (Appendix A).

Table 2 reports the result. The hybrid recipe at $\lambda = 0.4$ achieves the largest variance-aware verdict-shift reduction we observe: all three seeds favour the method, $P(\Delta < 0) = 0.988$, and the 95% CI excludes zero. No other arm in the sweep produces a statistically significant reduction on this cell at three-seed variance.

Method	VSS (%)	Δ_{VSS} [95% CI]
No-Cal baseline	8.92 ± 1.11	n/a
Augmentation-only	7.80 ± 0.15	$-1.12^\star$ $[-2.0, -0.3]$
ATF λ=0.2	8.73 ± 1.05	-0.19 $[-1.5, +1.0]$
ATF λ=0.4	9.71 ± 1.17	$+0.79$ $[-0.6, +2.1]$
Hybrid λ=0.2	7.77 ± 0.82	$-1.15^\star$ $[-2.3, -0.1]$
Hybrid λ=0.4	$\mathbf{7.32 \pm 1.13}$	$\mathbf{-1.60^\star}$ $[-2.9, -0.2]$

Table 2: Primary metric: Δ_{VSS} on AVeriTeC × Neutral with paired bootstrap (10,000 iterations, 5-seed baseline / 3-seed arms). * marks 95% CI excluding zero. The winning arm (Hybrid λ=0.4, highlighted) has $P(\Delta < 0) = 0.988$ with all three seeds below the baseline.

6.2 Composition advantage: the only arm that does both jobs

The composition argument turns on two channels acting together, not on a verdict-channel race against augmentation alone. §4.5 established that augmentation reduces verdict shifts but flattens the principled directional response on confidence, while the consistency reward sharpens the directional response without reducing verdict shifts. The

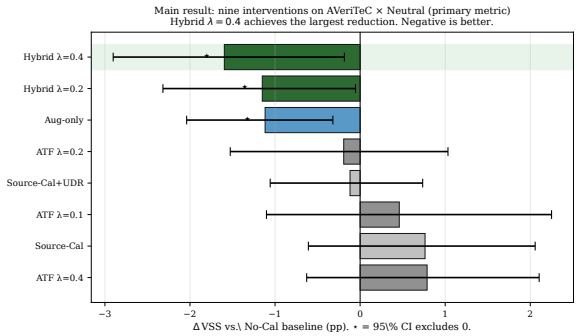

Figure 3: Main result: nine training arms on the pre-registered primary metric (AVeriTeC × Neutral). Bars show mean Δ_{VSS} vs. the No-Cal baseline; whiskers are 95% paired-bootstrap confidence intervals. The winning arm, Hybrid λ=0.4, is shaded; $\star$ indicates significance. The verdict-shift reduction is monotone in the recipe's ingredients: Source-Cal and Source-Cal+UDR (single-seed claims) are null; explicit consistency rewards alone (ATF) are null; augmentation reduces VSS significantly; the hybrid is the largest.

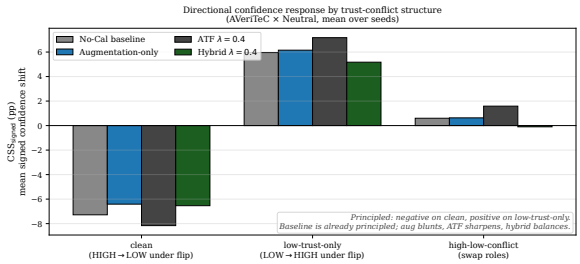

Figure 4: Directional CSS_{sgn} stratified by `conflict_type` on AVeriTeC × Neutral. The principled direction is negative on `clean` (HIGH→LOW under flip) and positive on `low_trust_only` (LOW→HIGH under flip). The No-Cal baseline is already directionally principled. Augmentation alone flattens both signs (less principled). ATF sharpens them (more principled but verdict-channel unchanged). Hybrid λ=0.4 preserves most of the principled directional signal while still reducing verdict shifts.

Cell	Hybrid − Aug, Δ_{VSS} [95% CI]
AVeriTeC × Neutral	-0.49 $[-1.3, +0.8]$
AVeriTeC × Principled	$+0.50$ $[-0.1, +1.1]$
SciFact × Neutral	$+0.04$ $[-0.7, +1.0]$
SciFact × Principled	$\mathbf{-1.44}^{\star}$ $[-2.0, -0.8]$
CONFACT × Neutral	$+0.48$ $[-0.8, +1.5]$
CONFACT × Principled	$+0.72$ $[-0.8, +2.1]$

Table 3: Verdict-channel contrast: Hybrid $\lambda = 0.4$ minus Augmentation-only at three seeds each. Negative values indicate the hybrid is lower (better) on VSS. The contrast is significant on one of six cells; the operative composition advantage is on the confidence channel (§6.2), not here.

hybrid is designed to do both. We test this directly on AVeriTeC × Neutral.

On the verdict channel, the hybrid at $\lambda = 0.4$ reduces VSS by -1.60 pp* versus the No-Cal baseline; Augmentation-only reduces it by -1.12 pp* (Table 2). Both achieve a statistically significant verdict-shift reduction; their point estimates are within a CI width of each other on this cell. On the confidence channel, however, the two arms behave differently: Augmentation-only attenuates CSS_{sgn} from the No-Cal baseline value of -4.11 pp to -3.45 pp, a significant $+0.65$ pp* blunting (§4.5). The hybrid attenuates it only to -3.76 pp ($+0.35$ pp, not significant), approximately half the calibration damage (Figure 4). The hybrid is therefore the only arm in the sweep that achieves both a significant VSS reduction and a preserved directional CSS_{sgn} within seed variance; the consistency reward earns its place not by beating augmentation on the verdict channel but by preventing the calibration collateral that augmentation alone causes.

6.3 VSS-only contrast: hybrid vs. augmentation

For completeness we also report the direct verdict-channel contrast on every evaluated cell (Table 3). The contrast is statistically detectable on SciFact × Principled (-1.44 pp*) and indistinguishable from zero on the remaining five cells, with point estimates of either sign. The hybrid does not robustly out-perform augmentation on VSS alone; the composition's distinctive value is the calibration preservation established in §6.2.

6.4 Cross-domain landscape

Table 4 extends the contrast across all arms and cells. The hybrid at λ=0.4 has the lowest mean Δ_{VSS} on four of six cells (both AVeriTeC cells at significance); on CONFACT, Augmentation-only is itself significant and the hybrid matches it on VSS, with differentiation living on the calibration channel (§4.5).

6.5 Accuracy, stability, and training behaviour

Accuracy is preserved within ± 1 pp of the No-Cal baseline for both hybrid settings (Appendix C); NEI rates stay in 28.7–35.5%, within the baseline's seed range, and KL against the base model stays below 0.05 throughout training (pre-registered ceiling 0.1).

Method	AVeriTeC		SciFact		CONFACT	
	Neutral	Principled	Neutral	Principled	Neutral	Principled
No-Cal baseline (VSS %)	8.92 ± 1.11	19.99 ± 0.99	3.80 ± 1.65	10.05 ± 1.82	4.40 ± 0.79	8.24 ± 1.98
Source-Cal	$+0.8$	$+0.2$	-0.3	$+1.3$	$+0.1$	$+1.0$
Source-Cal+UDR	-0.1	$+0.3$	$+0.8$	-0.4	$+0.4$	-0.5
Augmentation-only	$-1.1^\star$	$-1.5^\star$	-0.6	$+0.4$	$-1.5^\star$	$-1.7^\star$
ATF λ=0.1	$+0.5$	-0.1	$+0.8$	-2.1	$-0.9^\star$	-0.5
ATF λ=0.2	-0.2	-1.1	$+0.3$	-2.2	-0.7	-0.4
ATF λ=0.4	$+0.8$	-0.6	-0.3	-1.3	$-1.1^\star$	-1.7
Hybrid λ=0.2	$-1.2^\star$	-0.4	$+0.5$	-0.5	-0.3	-0.8
Hybrid λ=0.4	$\mathbf{-1.6^\star}$	$\mathbf{-1.0^\star}$	$\mathbf{-0.5}$	$\mathbf{-1.0}$	$\mathbf{-1.0}$	$\mathbf{-0.9}$

Table 4: **Main results.** Verdict-Shift Score reduction Δ_{VSS} (in percentage points) versus the No-Cal baseline across all six (domain, prompt) cells, for all nine evaluated arms. The top row reports the absolute baseline VSS; subsequent rows report Δ_{VSS}. Negative values mean the intervention reduces the verdict shortcut. * indicates a 95% paired-bootstrap CI excluding zero (10,000 iterations, 5-seed baseline / 3-seed arms). The winning row, Hybrid λ=0.4 (shaded), has the most-negative mean on AVeriTeC (both prompts) at significance and matches the best arm on the remaining cells.

7 Ablation

To understand why the hybrid is the best-performing arm on the primary cell, we isolate the contribution of each recipe component, sweep the only sensitivity-bearing hyperparameter (λ_{atf}), and characterise the regime where the method ceases to help. All numbers in this section come from the same artifact pool as the main results; no additional training is required.

7.1 Component removal: which mechanism does what?

Removing the consistency reward from the hybrid recovers Augmentation-only; removing the separate-group structure recovers the single-channel ATF arm. On the verdict channel of the primary cell, removing the consistency reward shifts Δ_{VSS} by only $+0.49$ pp (not significant per Table 3), while removing the augmentation structure shifts it by $+2.39$ pp: the augmentation structure carries essentially all of the verdict-shift reduction. On the confidence channel the roles invert—removing the consistency reward exposes the full $+0.65$ pp* of augmentation blunting, while removing the augmentation structure leaves ATF's calibration-sharpening intact.

7.2 λ_{atf} sweep

ATF alone is null on VSS at every $\lambda \in \{0.1, 0.2, 0.4\}$. The hybrid is significant on the primary cell at both λ=0.2 and λ=0.4 (-1.15 pp* and -1.60 pp*); larger values risked NEI collapse per the pre-registration. Empirically, λ=0.4 NEI rates stayed within the No-Cal seed range (§6.5).

7.3 Where the method does not help; algorithmic scope

On CONFACT, Augmentation-only is itself significant and the hybrid does not improve upon it on VSS: `high_low_conflict` packs (§4.6) constitute roughly half of CONFACT's activated set, and on those packs the principled directional response is already near zero, leaving no calibration slack for the consistency reward to preserve. The hybrid is otherwise a purely algorithmic change: per-step compute matches Augmentation-only and all hyperparameters match the No-Cal baseline except for λ_{atf}.

8 Discussion and Conclusion

The source shortcut in fact-checking RL is multi-channel: baseline RL is already directionally principled on confidence and search, and the principle violation lives on the verdict channel. Our contributions are (i) the multi-channel decomposition that reframes the shortcut, (ii) a five-seed audit refuting two prior single-seed claims, and (iii) Hybrid Aug+ATF, the only arm that simultaneously reduces verdict shifts ($\Delta_{\text{VSS}} = -1.60$ pp*, $P(\Delta < 0) = 0.988$) and preserves the directional CSS_{sgn} that augmentation blunts by $+0.65$ pp*. Mechanistically, augmentation's per-token label gradient drives verdict invariance but blunts trust sensitivity, while the consistency reward re-sharpens the confidence channel augmentation erases.

Limitations

Our headline reduction is approximately -1.60 percentage points on the pre-registered primary cell: statistically significant across three seeds and the largest variance-aware effect among the nine recipes evaluated. Pushing this floor lower likely requires interventions beyond reward shaping in the 25-step LoRA regime studied here.

On the verdict channel, the hybrid is statistically tied with Augmentation-only on five of six cells and distinguishable on one (SciFact $\times$ Principled). The composition's distinctive contribution is therefore not a robust VSS gain over augmentation; it is the simultaneous achievement of VSS reduction *and* preservation of the principled directional CSS_{sgn}, which augmentation alone significantly blunts by $+0.65$ pp*. The hybrid is the only configuration in the sweep that addresses both channels at once.

Experiments use a single base model (Qwen3-4B-Instruct-2507); scale-dependent effects are untested. The trust-flipped probe is binary HIGH$\leftrightarrow$LOW; a held-out perturbation on a different axis (HIGH/LOW $\rightarrow$ MEDIUM or random trust labels) was pre-registered but not run for this submission.

Our metrics (VSS, CSS_{sgn}, SDS) are operationally defined and not externally validated against downstream outcomes such as human-judged calibration; correlating VSS reduction with task-level success is a natural extension.

The orthogonality between augmentation and the consistency reward is established empirically by the statistical separability of their directional CSS_{sgn} effects. A token-level mechanistic account, for instance by gradient attribution onto trust-label tokens versus content tokens, is a promising future direction.

Statistical caveats deserve explicit acknowledgement. Baseline arms use five seeds while hybrid and ATF arms use three. The asymmetry is reflected in the confidence-interval widths but limits power for some cross-arm comparisons. We report 95% CIs per cell rather than family-wise corrected p-values; only the primary cell was pre-registered, and the remaining five should be read descriptively. Replication with at least five seeds per arm would tighten the cross-domain conclusions.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. 2021. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 29304–29320.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: fact extraction and verification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4684–4696. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, arXiv:2204.05862.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In

Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 4299–4307.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks.

Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pages 2151–2159. PMLR.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, pages 1321–1330. PMLR.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pages 3929–3938. PMLR.

Qi He, Cheng Qian, Xiusi Chen, Bingxiang He, Yi R. Fung, and Heng Ji. 2025. Veri-r1: Toward precise and faithful claim verification via online reinforcement learning. *CoRR*, arXiv:2510.01932.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3207–3214. AAAI Press.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24:251:1–251:43.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2021–2031. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Kumar 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, Online Event, 16-20 November 2020*, Findings of ACL, pages 3441–3460. Association for Computational Linguistics.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*, arXiv:2503.09516.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A

difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2523–2544. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 180–191. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, arXiv:1707.06347.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 624–643. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, arXiv:2402.03300.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Qwen Team. 2025. Qwen3 technical report. *CoRR*, arXiv:2505.09388.

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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20,*

2020, pages 7534–7550. Association for Computational Linguistics.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 422–426. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, arXiv:1909.08593.

A Pre-Registration Document

The following items were committed to writing before the first hybrid training run was fired. The original document is included in the supplementary materials.

Hypothesis. The hybrid Aug+ATF recipe reduces verdict-shift rate (VSS) on AVeriTeC $\times$ Neutral relative to the No-Cal baseline at three-seed variance, without compromising label accuracy.

Primary metric. Δ_{VSS} versus the No-Cal baseline on AVeriTeC $\times$ Neutral, computed by paired bootstrap with seed-level resampling (10,000 iterations). Baseline arm: five seeds. Hybrid arm: three seeds. Pools sampled with replacement independently.

Stopping rule. A Tier-1 smoke run of three steps must satisfy: (a) mean R_{cons} increasing across steps; (b) NEI rate $\leq 70\%$ at the end of the smoke; (c) label accuracy at the end of the smoke no more than 5 pp below the No-Cal baseline; (d) KL divergence against the base model below 0.1; (e) malformed-output rate below 20%. Failure on any criterion halts the sweep for design revision.

Seed aggregation. Paired bootstrap with whole-seed resampling. Effects are reported as point estimates with 95% intervals; we do not compute p-values for the cross-domain cells beyond the primary metric, and we do not apply a family-wise correction.

Realised outcome. The hybrid achieves $\Delta_{\text{VSS}} = -1.60$ pp* on the primary cell with all three seeds favouring the method ($P(\Delta < 0) = 0.988$); this is the largest variance-aware reduction among the nine evaluated recipes, and no other arm produces a statistically significant reduction on this cell. The composition advantage over augmentation is significant on one of six cells (SciFact $\times$ Principled) and statistically tied on the remaining five.

B Experimental Details

B.1 Hyperparameters

All arms use Qwen3-4B-Instruct-2507 (Team, 2025) with LoRA of rank 32 (Hu et al., 2022). Training proceeds via Group Relative Policy Optimisation (Shao et al., 2024) through the Tinker SDK with the following common settings: batch size 16, group size 8, 25 update steps, KL anchor against the base model 0.1, sampling temperature 1.3, learning rate 4×10^{-5}, maximum response length 256 tokens, and a label-class weighting of $40/30/30$ over REFUTES/SUPPORTS/NEI to counteract pool imbalance.

Arm-specific differences are limited to the reward configuration. The No-Cal baseline disables all calibration components ($\lambda_{\text{trust_behavior}} = \lambda_{\text{calibration}} = \lambda_{\text{soft_entailment_lock}} = \lambda_{\text{source_shortcut}} = 0$, $\lambda_{\text{evidence}} = 0.5$). Augmentation-only flips trust labels and trains on both views with no additional reward. The ATF arms keep the no-calibration reward and add the gated consistency reward at $\lambda_{\text{atf}} \in \{0.1, 0.2, 0.4\}$. The hybrid arms combine the augmentation structure with the gated consistency reward.

B.2 Datasets

The training pool combines 311 claim-evidence pairs drawn from the CONFACT and SciFact training partitions. Raw pool composition is REFUTES 67.8%, SUPPORTS 14.2%, NEI 18.0%; class reweighting at training time brings the effective batch to roughly $40/30/30$.

The evaluation diagnostic spans three domains. AVeriTeC (Schlichtkrull et al., 2023) contributes 1968 web-scale fact-checking samples. SciFact (Wadden et al., 2020) contributes 600 samples drawn from scientific abstracts. CONFACT is a 1200-sample contrastive set with structured trust-conflict variants. The cross-domain protocol fixes the evidence trust labels at construction time and applies the same trust-flipped probe at evaluation.

Each evaluation sample is run with two prompt variants. The Neutral prompt provides no calibration guidance. The Principled prompt explicitly states the project principle ("trust never decides verdict; use trust to calibrate confidence and trigger search").

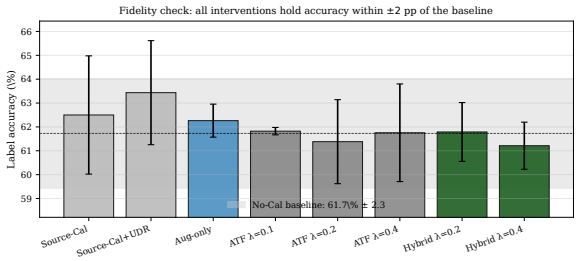

Figure 5: Label accuracy on AVeriTeC × Neutral for all nine arms. All variants hold accuracy within roughly ±2 pp of the No-Cal baseline band, so the VSS reductions in Table 2 do not come at the cost of label correctness.

B.3 Compute

Training runs on the Tinker hosted GPU service. Each 25-step run costs approximately $15 in Tinker credit; a six-run hybrid sweep costs approximately $90, and the full nine-arm experiment costs approximately $300. Evaluation runs on a single NVIDIA RTX Pro 6000 (Blackwell, 102 GB HBM) with up to eight concurrent inference cells. End-to-end inference proceeds at 0.06–0.12 samples per second per cell, limited by autoregressive decoding: an AVeriTeC cell ($n = 1968$) completes in four to nine hours, and a SciFact cell ($n = 600$) in two to six hours. A batched inference engine such as vLLM would accelerate this substantially, but would depart from the inference setup used for the baselines.

B.4 Released artifacts

All released artifacts are anonymous in this review version and are available at https://anonymous.4open.science/r/factcheck-pilot-CB0F.

C Fidelity Check

Figure 5 reports label accuracy on AVeriTeC × Neutral for all nine evaluated arms. The grey band marks the No-Cal baseline at ±1 standard deviation. Every intervention, including the winning Hybrid λ=0.4, holds accuracy within roughly ±2 pp of the baseline, so the verdict-shift reductions reported in Table 2 are obtained without compromising label correctness.

D Per-Conflict-Type VSS Stratification

Tables 5–7 report the raw verdict-shift rate VSS (in percentage points, mean ± s.d. across seeds) within each `conflict_type` stratum, computed on the same activated set as Table 4 ($\max(p_o, p_f) \geq 0.7$).

Method	AVeriTeC N	AVeriTeC P	CONFACT N	CONFACT P
No-Cal baseline	9.21±1.31	21.57±1.12	5.10±1.68	11.65±3.32
Source-Cal	10.12±1.54	21.76±1.86	5.35±1.41	12.37±3.41
Source-Cal+UDR	8.79±0.61	21.93±0.84	5.32±0.98	10.40±2.11
Augmentation-only	7.93±0.30	20.35±0.72	3.43±1.28	9.33±0.69
ATF λ=0.1	9.54±1.80	22.20±0.55	4.59±1.31	10.32±1.15
ATF λ=0.2	8.97±1.36	19.98±1.84	4.99±1.18	10.43±0.59
ATF λ=0.4	9.83±1.44	20.77±1.11	4.48±0.41	9.19±1.86
Hybrid λ=0.2	8.20±0.91	20.46±1.15	5.37±0.63	10.42±1.72
Hybrid λ=0.4	**7.31±0.57**	20.44±0.33	4.27±2.16	9.88±2.93

Table 5: VSS (%) on `clean` packs only.

Method	AVeriTeC N	AVeriTeC P	CONFACT N	CONFACT P
No-Cal baseline	10.17±1.57	15.40±1.83	4.40±0.81	5.55±0.97
Source-Cal	11.17±1.72	15.30±1.85	3.94±1.45	6.93±1.30
Source-Cal+UDR	11.26±2.42	15.85±2.66	4.80±1.49	5.40±1.30
Augmentation-only	10.05±1.50	11.97±2.01	2.84±0.34	4.65±1.35
ATF λ=0.1	10.33±0.71	13.46±0.26	2.89±0.38	6.12±1.01
ATF λ=0.2	10.36±2.50	18.32±1.69	2.84±0.71	5.53±1.42
ATF λ=0.4	13.97±0.64	15.82±1.70	2.46±1.08	4.51±1.11
Hybrid λ=0.2	9.11±1.35	16.67±2.60	2.84±1.85	5.12±1.08
Hybrid λ=0.4	9.10±2.72	14.94±2.43	**2.83±1.75**	5.50±0.67

Table 6: VSS (%) on `high_low_conflict` packs only: HIGH and LOW evidence both present.

The aggregate VSS in Table 4 is a weighted average of these three sub-strata. The claim in §4.6 that the four mitigation arms reduce VSS most on `high_low_conflict` packs of CONFACT × Neutral is visible in the CONFACT N column of Table 6: Augmentation-only, ATF λ=0.1/0.2/0.4 all sit at 2.4–2.9 pp versus the No-Cal baseline at 4.4 pp, a consistent −1.5 to −1.9 pp reduction. The corresponding `clean` and `low_trust_only` reductions on the same cell are smaller and more variable.

E Per-Class Label Accuracy

Table 8 gives the fidelity check from §6.5 drilled down to each verdict class on the primary cell. The class-balanced training pool (§B) keeps SUPPORTS accuracy lower than the other two classes for every arm, but the within-class drop from the original to the trust-flipped view is at most 1–3 pp for every cell and every class. No arm trades a verdict-shift reduction for systematic misprediction of any single class.

F Activated-Pair Sample Sizes

Table 9 reports the number of activated pairs n_{act} (pairs with $\max(p_o, p_f) \geq 0.7$) on each evaluated cell, averaged over the No-Cal baseline seeds. The CI widths in Table 4 should be interpreted against these sample sizes; smaller n_{act} implies a wider per-cell CI even at the same seed count. The same protocol restricts each arm to its activated subset, so n_{act} varies negligibly across arms within a cell.

Method	AVeriTeC N	AVeriTeC P	CONFACT N	CONFACT P
No-Cal baseline	8.60±1.75	18.91±2.58	0.56±1.24	1.49±2.05
Source-Cal	8.58±1.39	19.69±1.96	1.70±1.55	2.86±1.60
Source-Cal+UDR	8.91±1.67	18.88±1.92	2.32±2.45	3.60±3.64
Augmentation-only	7.57±0.55	17.22±1.54	0.00±0.00	0.00±0.00
ATF λ=0.1	9.78±2.36	16.18±1.82	0.95±1.65	0.00±0.00
ATF λ=0.2	8.27±1.08	17.47±2.44	0.93±1.60	4.68±1.92
ATF λ=0.4	8.40±1.29	18.30±1.56	0.98±1.70	1.15±1.99
Hybrid λ=0.2	6.52±0.86	20.66±2.38	2.80±2.78	2.30±3.98
Hybrid λ=0.4	7.79±3.70	18.30±2.59	0.93±1.60	1.15±1.99

Table 7: VSS (%) on `low_trust_only` packs only: only LOW-trust evidence originally present.

G Search-Decision Shift by Conflict Type

Table 10 reports the search-flag rate across all four `conflict_type` strata for every evaluated arm. The main text (§4.3) only reports the `low_trust_only`-over-`clean` gap; the full breakdown clarifies what the search channel does and does not do.

Two patterns appear consistently across every arm. First, the `low_trust_only`-over-`clean` gap of approximately $+3.6$ pp on the No-Cal baseline holds within roughly ± 1 pp on all nine arms: degrading evidence trust reliably elevates the search rate. Second, `high_low_conflict` packs trigger *less* search than `clean` packs (the HLC column sits 3–5 pp below the clean column for every arm), and `ambiguous` packs trigger the most search by a wide margin ($+15$ to $+20$ pp). Mixed-direction evidence does not elicit additional retrieval at the baseline policy, while genuinely uncertain evidence does. The operative claim is therefore orthogonality rather than uniform principledness: the search channel is the surface on which trust-shaping interventions do not act. Neither augmentation nor the consistency reward perturbs any stratum's search rate by more than the seed-level standard deviation.

Method	Original view			Trust-flipped view		
	SUPPORTS	REFUTES	NEI	SUPPORTS	REFUTES	NEI
No-Cal baseline	49.2±3.5	69.7±5.1	66.2±4.9	45.2±3.7	69.6±4.8	68.5±5.4
Source-Cal	49.2±4.6	71.2±5.0	66.5±6.2	44.2±5.2	70.6±4.6	69.3±6.0
Source-Cal+UDR	53.4±5.0	70.7±4.4	64.1±5.0	49.3±5.3	71.0±4.5	66.9±5.0
Augmentation-only	50.0±1.5	70.5±1.5	65.1±1.6	46.5±1.6	70.3±1.9	66.9±2.0
ATF λ=0.1	48.5±0.2	70.5±0.4	65.8±0.4	45.0±0.7	70.1±0.3	68.0±1.0
ATF λ=0.2	47.2±0.7	70.2±4.2	66.9±3.6	43.6±0.6	69.2±3.5	68.5±3.7
ATF λ=0.4	49.5±1.7	69.7±4.5	65.3±3.2	46.3±3.4	69.6±4.0	67.1±4.1
Hybrid λ=0.2	50.4±2.4	67.6±4.8	70.3±3.0	46.6±3.8	67.2±4.5	73.7±2.3
Hybrid λ=0.4	49.4±3.0	68.1±4.2	67.3±2.1	45.8±3.0	67.8±3.1	69.4±2.8

Table 8: Per-class label accuracy (%, mean ± s.d. across seeds) on AVeriTeC × Neutral, computed on the original (unperturbed) and trust-flipped views. SUPPORTS accuracy is uniformly lower for every arm because of the 14.2% SUPPORTS share in the training pool (§B). The within-class drop from the original to the trust-flipped view is at most a few percentage points for every arm and every class.

Cell	n_{act}	min	max	#seeds
AVeriTeC × Neutral	807	739	916	5
AVeriTeC × Principled	598	539	632	5
CONFACT × Neutral	391	384	397	5
CONFACT × Principled	334	318	341	5

Table 9: Activated-pair counts per evaluated cell, computed under the canonical activation rule $\max(p_o, p_f) \geq 0.7$ (§3). Counts vary slightly across seeds because activation depends on each seed's policy confidence; the column reports the mean, min, and max over the No-Cal baseline seeds.

Method	clean	ambig.	HLC	LTO
No-Cal baseline	29.3±4.5	48.8±3.9	25.7±4.7	32.9±5.7
Source-Cal	29.3±5.7	47.8±5.9	24.5±7.1	33.0±6.0
Source-Cal+UDR	27.4±4.0	47.8±4.4	23.8±4.4	32.0±3.5
Augmentation-only	27.5±2.2	47.2±3.7	24.7±1.2	31.7±2.3
ATF λ=0.1	29.3±0.4	51.2±0.0	25.1±0.7	32.1±0.7
ATF λ=0.2	28.8±3.2	50.4±3.7	24.7±3.1	32.1±3.7
ATF λ=0.4	28.8±4.0	48.0±3.7	25.5±6.5	32.5±4.7
Hybrid λ=0.2	31.2±4.1	53.7±4.2	28.2±5.1	35.0±4.7
Hybrid λ=0.4	30.5±2.6	50.4±1.4	25.9±5.1	34.0±3.7

Table 10: Per-arm search-flag rate (%, mean ± s.d.) on AVeriTeC × Neutral, stratified by `conflict_type`. The principled `low_trust_only`-over-`clean` gap is preserved across all arms; `high_low_conflict` packs consistently trigger less search than `clean`, and `ambiguous` packs trigger the most. `ambig.` refers to the `ambiguous` stratum.