TRUST LABELS SHOULD CALIBRATE, NOT DECIDE: DIAGNOSING AND REDUCING SOURCE-LABEL SHORTCUTS IN RL-TRAINED FACT-CHECKERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented fact-checkers often receive a reliability label, such as HIGH or LOW trust, for each evidence source. These labels should adjust the model's confidence and its decision to search for more evidence, while the verdict should follow the evidence content. We introduce TRUSTSWAP, a counterfactual test that swaps, lowers, or removes source labels while keeping every evidence text fixed, and measures the verdict, the confidence, and the search decision separately. Across untrained and RL-trained models at two scales, three datasets, and two prompts, confidence and search respond to the labels as intended in 49 of 50 comparisons, yet a label change alone alters 4–23% of confident verdicts for Qwen3 models and up to 50% for an existing RL-trained fact-checker. Standard GRPO fine-tuning amplifies this shortcut at 8B in all six settings. To reduce it, we propose trust-swap augmentation (TSA), which trains GRPO on each claim with both its original and its label-swapped evidence under the same gold verdict. At 4B, TSA lowers the verdict flip rate by 7–35% (relative) in four of six settings, keeps accuracy and the intended confidence and search responses, outperforms reward-based alternatives in the main setting, and carries over to an unseen label-removal perturbation. An added consistency reward helps on the trained-on swap but not on unseen perturbations. At 8B, TSA's effect is not detectable, which makes scale the main open question.[1]

1 INTRODUCTION

Automated fact-checking systems decide whether retrieved evidence supports or refutes a claim (Guo et al., 2022; Thorne et al., 2018; Schlichtkrull et al., 2023). Recent systems train language models with reinforcement learning (RL) to search for evidence and answer (Jin et al., 2025), including for claim verification (He et al., 2025). The retrieved evidence comes from sources of very different reliability, so pipelines increasingly attach a reliability label to each source, for example from media-rating services (Media Bias/Fact Check, n.d.; Ge et al., 2025), and ask the model to take source credibility into account (Pan et al., 2024; Deng et al., 2025). This raises a basic question: *what should a model do with these labels?*

We study a simple contract: *the evidence content decides the verdict, and the labels adjust the model's confidence and its decision to look for more evidence.* Two considerations motivate it. First, reliability labels are metadata assigned by the pipeline, not properties of the evidence text. They can be wrong or outdated. Indirect prompt injection demonstrates manipulation of retrieved content (Abdelnabi et al., 2023). Here, we assume that an attacker can change source labels while keeping the evidence text fixed. Second, the model already has two outputs for expressing doubt about sources, namely its confidence and its search decision, so keeping the verdict tied to the content loses no information. Other contracts are defensible; for example, a system could abstain whenever only low-reliability sources are available. We therefore also report verdict reversals separately from abstentions, so that our conclusions hold under this weaker contract as well.

[1] Anonymous code repository: `https://anonymous.4open.science/r/factcheck-pilot-CB0F`.

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

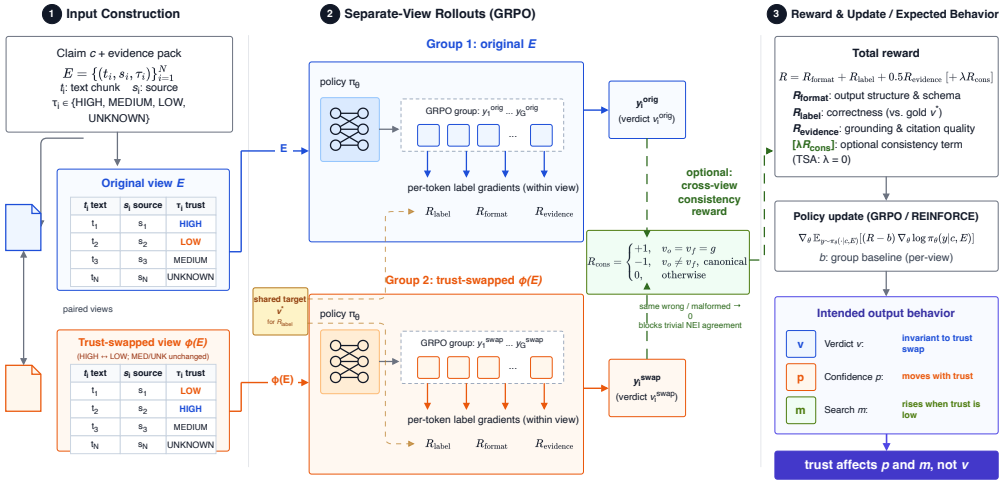

Figure 1: TRUSTSWAP and trust-swap augmentation (TSA). **(1)** Each claim is paired with its evidence pack E and a trust-swapped pack $\phi(E)$ (HIGH↔LOW; texts and sources unchanged); TRUSTSWAP compares the outputs on the two. **(2)** TSA trains GRPO on both packs as separate rollout groups with the same gold verdict $v^\star$; an optional consistency reward R_{cons} links matched rollouts (v_o, v_f: verdicts on E and $\phi(E)$; g: gold). **(3)** TSA uses $\lambda = 0$. Trust should change the confidence p and the search decision m, but not the verdict v.

To test the contract, we introduce TRUSTSWAP (§3). For each claim, it exchanges the HIGH and LOW labels in the evidence pack, keeps every evidence text and source name fixed, and compares the model's outputs before and after the change (Figure 1, left). Two further perturbations, which lower or remove the labels, test whether a model trained on the swap has learned something more general. Because the contract expects the verdict to stay fixed but confidence and search to move, TRUSTSWAP measures the three outputs separately.

TRUSTSWAP shows that the shortcut is specific to the verdict (§4). Across five models, three datasets, and two prompts, **confidence and search already follow the contract** in 49 of 50 comparisons, so the models do read the labels as reliability signals. **The verdict does not follow the contract:** a label change alone alters 4–23% of confident verdicts for Qwen3 models and 16–50% for Veri-R1, an existing RL-trained fact-checker. **RL fine-tuning can amplify the shortcut:** GRPO training of Qwen3-8B raises the flip rate in all six settings, by 44% in relative terms in the main setting, while accuracy in that setting changes by less than 2 percentage points (pp).

Because the failure is confined to the verdict, a fix should target the verdict and leave confidence and search free to respond. We propose **trust-swap augmentation (TSA)**, a counterfactual data augmentation (Kaushik et al., 2020) recipe for GRPO: each training claim appears twice, with its original and its label-swapped evidence, under the same gold verdict, and the two versions form separate rollout groups (Figure 1; §5). At 4B, TSA lowers the flip rate in four of six settings (by 7–35% relative, all significant), keeps or slightly improves accuracy in all six, outperforms trust-aware reward shaping and a verdict-consistency reward in direct comparisons on the main setting, and carries over to an unseen perturbation that removes all labels. We also characterise where TSA stops helping: its gains come from fewer abstention changes rather than fewer reversals, and they do not appear at 8B.

Our contributions are:

- **TRUSTSWAP**, a counterfactual test of source-label use: 1,584 claims from three datasets, three label perturbations, and separate measures for verdict (split into reversals and abstentions), confidence, and search (§3).
- **A diagnosis across models and scales**: the source-label shortcut lives in the verdict, not in confidence or search, and standard GRPO fine-tuning amplifies it at 8B (§4).

- **Trust-swap augmentation**, a simple GRPO recipe that reduces the verdict shortcut at 4B without reducing accuracy, outperforms reward-based alternatives, and generalizes to an unseen perturbation, with its scope mapped at 8B (§5–§6).

2 RELATED WORK

Fact-checking with retrieval and RL. Fact-checking benchmarks cover political claims, Wikipedia, science, and web evidence (Wang, 2017; Thorne et al., 2018; Jiang et al., 2020; Aly et al., 2021; Wadden et al., 2020; Schlichtkrull et al., 2023), and retrieval-augmented models combine retrieved evidence with a language model (Lewis et al., 2020; Asai et al., 2024). RL-trained agents learn to search and answer from reward signals (He et al., 2025; Jin et al., 2025; Shao et al., 2024). We show that such training can increase reliance on source labels even when the reward never mentions them.

Source credibility and knowledge conflicts. Studies of knowledge conflicts ask how models weigh competing evidence and their own knowledge (Xie et al., 2024; Xu et al., 2024), and credibility-aware methods give models reliability scores so that they rely less on unreliable context (Pan et al., 2024; Deng et al., 2025; Ge et al., 2025). These works treat credibility as an input that should shape the answer; we ask *which outputs* it should shape and find that the verdict is the one that should not change.

Calibration and abstention. Confidence calibration asks whether stated confidence matches accuracy (Guo et al., 2017), and selective prediction lets a model abstain when it is uncertain (Geifman & El-Yaniv, 2019). In our setting, trust labels should move the confidence and the search decision but not the verdict, so an abstention caused only by a label change counts as a verdict change, which we report separately from reversals.

Shortcuts and counterfactual training. Shortcut learning, where models rely on features that correlate with the label but do not determine it, is well documented in NLP (Geirhos et al., 2020; Gururangan et al., 2018; McCoy et al., 2019). Counterfactual and behavioural tests expose such shortcuts (Gardner et al., 2020; Ribeiro et al., 2020), and counterfactual augmentation and consistency penalties are standard mitigations (Kaushik et al., 2020; Garg et al., 2019; Veitch et al., 2021). In our setting the perturbed feature must be ignored by one output but used by the others, so we evaluate every method on all outputs at once.

3 TRUSTSWAP: A COUNTERFACTUAL TEST FOR SOURCE-LABEL USE

Task and outputs. The input is a claim c and an evidence pack $E = \{(t_i, s_i, \tau_i)\}_{i=1}^{N}$, where t_i is the evidence text, s_i the source domain, and $\tau_i \in \{\text{HIGH}, \text{MEDIUM}, \text{LOW}, \text{UNKNOWN}\}$ the source's trust label (we use "trust label", "source label", and "reliability label" interchangeably). The model answers in a fixed text format with a verdict $v \in \{\text{SUPPORTS}, \text{REFUTES}, \text{NEI}\}$, a confidence $p \in [0, 1]$, a search flag $m \in \{0, 1\}$ that states whether more evidence is needed, and a one-sentence rationale (prompts in Appendix B). The search flag is an output only; no second retrieval round is run.

Perturbations and pack types. The *label swap* ϕ exchanges HIGH and LOW labels and leaves MEDIUM, UNKNOWN, every text t_i, and every source s_i unchanged. It is the perturbation used for training in §5. Two further perturbations are never used in training: a *downshift* (HIGH→MEDIUM, MEDIUM→LOW) and *label removal* (every label set to UNKNOWN). The contract requires the verdict to stay fixed under all three. For the swap, the direction of the change depends on the pack. In a *high-trust pack* (at least one HIGH source and no LOW source), the swap lowers trust; in a *low-trust pack* (at least one LOW source and no HIGH source), it raises trust; in a *mixed pack* (both), it reverses which sources are trusted, so no single direction is expected; *unlabeled packs* (neither) are unchanged.

Metrics. We compare the original and perturbed predictions (v, p, m) and (v', p', m') on *confident pairs*, where $\max(p, p') \geq 0.7$, because a verdict change matters most when the model commits to

Claim	Evidence (source: label before → after the swap)	Original	Swapped
(a) Typical case: labels lowered on a high-trust pack. Confidence and search respond as intended, but the verdict also changes.			
ICE removed thousands of gang members.	[1] `ice.gov`: HIGH → LOW. "5,396." [2] `ice.gov`: HIGH → LOW. "5,872."	SUPPORTS conf. 0.95 search: no	**NEI** conf. 0.3–0.5 search: yes
(b) Reversal: only one label is raised, and the verdict switches polarity with unchanged high confidence.			
Social distancing [. . .] was not practiced at the funeral of civil rights icon John Lewis.	[1] `perma.cc`: UNKNOWN. "...the church will only accommodate 240 people in total, due to social distancing precautions." [2] `facebook.com`: LOW → HIGH. "Yes. The church's senior pastor ... referred to social distancing ..." [3] `gettyimages.com`: UNKNOWN. "Yes. Photo by ..."	SUPPORTS conf. 0.95 search: no	**REFUTES** conf. 0.95 search: no

Figure 2: The label swap on two AVeriTeC claims. Only the reliability labels change; every evidence text and source name is identical before and after. Outputs are from the 4B GRPO baseline, and both verdict changes occur in all five training seeds (the confidence range in (a) spans seeds). Under our contract, the confidence drop and the search request in (a) are correct, but neither verdict change is. The gold label is SUPPORTS in (a) and REFUTES in (b).

an answer. One measure follows each output. The **flip rate** is the percentage of confident pairs with $v \neq v'$; we split it into the *reversal rate* (SUPPORTS↔REFUTES) and the *abstention-flip rate* (one of v, v' is NEI). The **confidence shift** is the mean of $p' - p$ and the **search shift** is the mean of $m' - m$, both in percentage points (pp) and reported per pack type. The contract predicts a flip rate of zero, a negative confidence shift and a positive search shift on high-trust packs, and the opposite signs on low-trust packs. To summarise each output in one number, the *confidence response* is the pack-size-weighted mean of the confidence drop on high-trust packs and the confidence rise on low-trust packs, and the *search response* is defined in the same way from the search shift; larger values mean a stronger response in the intended direction. Because decoding is greedy, every difference between the two predictions is caused by the label change.

Data. TRUSTSWAP covers three datasets. AVeriTeC (Schlichtkrull et al., 2023) provides 984 real-world claims with web evidence (700 high-trust, 158 low-trust, 85 mixed, and 41 unlabeled packs). CONFACT (Ge et al., 2025) provides 400 claims whose sources often conflict (197 high-trust, 36 low-trust, 167 mixed). SciFact (Wadden et al., 2020) provides 200 claims over scientific abstracts, all in high-trust packs. Trust labels come from Media Bias/Fact Check ratings (Media Bias/Fact Check, n.d.), a manual list of about 150 domains, and top-level-domain rules (Appendix B). We use two prompts. The *neutral* prompt explains the labels without advice. The *principled* prompt states the contract, and it also advises the model to request more search instead of committing to a verdict when only low-trust evidence is available. Our main setting is AVeriTeC with the neutral prompt, chosen because it has the most confident pairs (about 800).

Models. We test untrained Qwen3-4B-Instruct-2507 and Qwen3-8B (thinking disabled) (Qwen Team, 2025); Veri-R1, a Qwen2.5-3B model trained with online RL for claim verification (He et al., 2025); and a *GRPO baseline* for each Qwen3 size, trained as described in §5. Evaluation uses greedy decoding.

4 WHAT MODELS DO WITH TRUST LABELS

Confidence and search follow the contract. Figure 2 shows two example cases, and Figure 3 summarises all models. Figure 3b–c shows the two outputs that are supposed to respond to the labels. In the main setting, when the swap lowers trust, every model becomes less confident and asks for more search; when the swap raises trust, every model becomes more confident and asks for less search. The 4B GRPO baseline, for example, lowers its confidence by 7.3 pp and raises its

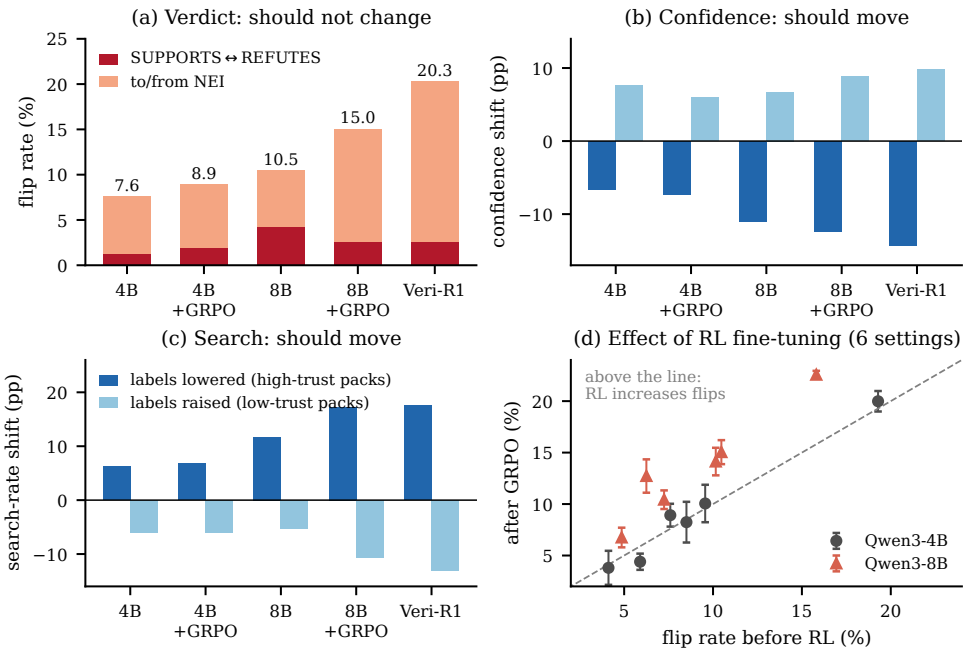

Figure 3: TRUSTSWAP label-swap results. **(a)–(c)** Main setting (AVeriTeC, neutral prompt): (a) flip rate, split into reversals and switches to or from NEI; (b, c) mean change in confidence and search requests where labels are lowered (dark) or raised (light), both in the intended direction. **(d)** All six settings: flip rate before and after GRPO fine-tuning (mean ± s.d., five seeds); points above the diagonal mean that training increased flips.

search rate by 6.7 pp on high-trust packs, and moves both by about 6 pp in the opposite direction on low-trust packs. The pattern holds in 49 of 50 comparisons for each output (five models, both pack types, all datasets and prompts; Appendix C); the single exception is Veri-R1 on 10 low-trust CONFACT pairs. The models therefore use the labels for confidence and search as the contract intends.

The verdict does not follow the contract. The same swap also changes the verdict (Figure 3a). In the main setting, the flip rate is 7.6% for untrained Qwen3-4B, 8.9% for its GRPO baseline, 10.5% for untrained Qwen3-8B, 15.0% for its GRPO baseline, and 20.3% for Veri-R1. Across all six settings, it ranges from 3.8% to 22.6% for the Qwen3 models and from 15.8% to 50.0% for Veri-R1. Most flips are abstention changes (78% for the 4B baseline): 73% of these are switches *to* NEI, and 95% of those occur on high-trust packs, where the swap lowers trust (Figure 2a). A weaker contract could accept such abstentions, but reversals cannot be justified in this way, and they are not rare: they make up 13–40% of flips in the main setting. For the 4B baseline, 74% of reversals occur on high-trust packs, where the swap lowers every HIGH source together and leaves the evidence text unchanged, so no reliability-weighted reading of the evidence explains a switch between SUPPORTS and REFUTES. Figure 2b shows a reversal caused by raising a single label.

Stating the rule in the prompt does not help. The principled prompt, which states the contract, raises the flip rate in 14 of 15 model–dataset pairs, for example from 8.9% to 20.0% for the 4B baseline on AVeriTeC. Almost all of this is abstention (18.6 of 20.0 points), in line with the prompt's advice not to commit when only low-trust evidence is available. Prompting alone is therefore not enough, which motivates a training-time fix.

RL fine-tuning can amplify the shortcut. Figure 3d compares each GRPO baseline with its untrained model. At 8B, training raises the flip rate in every setting, by 1.9–6.8 pp, and all 30 seed–setting runs lie above the untrained value; in the main setting the rate rises from 10.5% to 15.0%,

while accuracy stays close (67.0% before and $65.1 \pm 0.6\%$ after). The increase comes from abstention flips, which double from 6.3% to 12.5% of confident pairs, while reversals fall from 4.2% to 2.6%. At 4B the change is small and mixed (-1.5 to $+1.3$ pp across settings). The reward contains no trust term, so this behaviour is learned indirectly; an outcome reward alone therefore does not protect against the shortcut.

5 TRUST-SWAP AUGMENTATION

GRPO baseline. We fine-tune each Qwen3 model with Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and LoRA of rank 32 (Hu et al., 2022) for 25 steps, with 16 claims per step and G=8 rollouts per claim. The reward is $R = R_{\text{format}} + R_{\text{label}} + 0.5\,R_{\text{evidence}}$: a format term, a label-correctness term, and an evidence-citation term, none of which refers to trust labels. Training uses 311 claims from CONFACT and SciFact, excluding evaluation claims by identifier (REFUTES 67.8%, SUPPORTS 17.7%, NEI 14.5%, resampled to 40/30/30). One CONFACT claim text appears in both sets. Source labels and class labels are correlated in this pool: every training claim with a LOW-trust source is labelled REFUTES. The learning rate is 4×10^{-5} at 4B and 10^{-5} at 8B (§5.1); Appendix A lists all settings. This short run is a controlled instance of RL fine-tuning for fact-checking, not an attempt to maximise accuracy; AVeriTeC is out of domain for the training claims.

Trust-swap augmentation (TSA). The diagnosis shows that the model should become *invariant* to the labels in its verdict but stay *sensitive* to them in its confidence and search. TSA gives the verdict a direct training signal for this invariance (Figure 1). For every training claim whose pack contains a HIGH or LOW label, we add the swapped pack $\phi(E)$ with the same gold verdict. The original and swapped versions form two separate GRPO groups of G rollouts each, so each version's advantages are normalised within its own group and the label-correctness reward pushes both versions toward the same gold verdict. The reward is unchanged, and no term constrains confidence or search, which remain free to respond to the labels. TSA doubles the number of rollouts per claim but changes nothing else in the baseline recipe.

Optional consistency reward. We also test an explicit link between the two versions. For matched rollouts on E and $\phi(E)$ with verdicts v and v' and gold verdict g, the consistency reward is $R_{\text{cons}} = +1$ if $v = v' = g$, -1 if $v \neq v'$ and both are well-formed, and 0 otherwise; the zero case stops the model from earning the reward by answering NEI on both. *TSA + consistency* adds λR_{cons} to the reward of both groups.

Baselines. Besides the GRPO baseline, we compare with two families that also target trust behaviour. *Trust-aware reward shaping* adds four reward terms that ask confidence and search to track source trust, confidence to match correctness, and the verdict to agree with an NLI model when the NLI signal is clear, with an extra penalty for disagreeing with the NLI model on packs that contain LOW sources; a second variant adds a denser reward for abstaining on weakly supported claims. The *consistency reward* alone also samples rollouts on both E and $\phi(E)$, but places them in one joint group of $2G$ rollouts and adds λR_{cons}, for $\lambda \in \{0.1, 0.2, 0.4\}$. It therefore differs from TSA + consistency only in using a joint rather than a separate group per version, and it samples as many rollouts per claim as TSA, so comparisons with TSA are matched in compute.

5.1 EVALUATION PROTOCOL

Seeds and statistics. We train five seeds for the GRPO baseline and trust-aware rewards, and three for the other arms at 4B; at 8B, all arms have five matched seeds. Because single-seed RL results are often unreliable (Henderson et al., 2018; Bouthillier et al., 2021), we report differences in percentage points with 95% bootstrap intervals over seeds (10,000 resamples) (Agarwal et al., 2021) and mark with $\star$ intervals that exclude zero. Comparisons against the 4B baseline resample seeds independently within each arm; comparisons between two arms that share seed indices resample the per-seed differences. With three seeds an interval excludes zero essentially only when all per-seed differences share a sign (one-sided sign-test $p = 0.125$), so we confirm 4B results with the unseen perturbations and the 8B replication. The main setting was fixed before the experiments; we do not correct for multiple comparisons, and results in the other five settings serve as supporting evidence.

Table 1: Change in flip rate (pp) relative to the 4B GRPO baseline in all six settings (N/P: neutral/principled prompt); first row: baseline flip rate (%, mean ± s.d., five seeds). Negative is better. ⋆: 95% interval excludes zero. Last row: TSA + consistency (λ=0.4) minus TSA on matched seeds.

Method	AVeriTeC N	AVeriTeC P	SciFact N	SciFact P	CONFACT N	CONFACT P
GRPO baseline (%)	8.92 ± 1.11	19.99 ± 0.99	3.80 ± 1.65	10.05 ± 1.82	4.40 ± 0.79	8.24 ± 1.98
Trust-aware reward	+0.76	+0.23	−0.32	+1.28	+0.05	+1.05
Trust-aware + abstention	−0.12	+0.28	+0.77	−0.36	+0.44	−0.52
Consistency, λ=0.1	+0.46	−0.12	+0.84	−2.08	−0.86⋆	−0.55
Consistency, λ=0.2	−0.19	−1.14	+0.28	−2.22	−0.67	−0.37
Consistency, λ=0.4	+0.79	−0.61	−0.25	−1.28	−1.09⋆	−1.69
TSA (ours)	−1.12⋆	−1.49⋆	−0.56	+0.44	−1.52⋆	−1.66⋆
TSA + consistency, λ=0.2	−1.15⋆	−0.43	+0.50	−0.48	−0.32	−0.75
TSA + consistency, λ=0.4	−1.60⋆	−0.99⋆	−0.51	−1.00	−1.03	−0.93
TSA + consistency (λ=0.4) − TSA	−0.48	+0.50	+0.05	−1.44⋆	+0.50	+0.74

8B training. The 4B learning rate made 8B training unstable (mean accuracy fell to 51–59% across arms; Appendix H). We therefore re-selected the learning rate on the baseline arm only, requiring accuracy within 2 pp of the untrained model and an NEI rate of at most 30%, without looking at any TRUSTSWAP metric. Of three candidates (2×10^{-5} and 10^{-5} for 25 steps, 4×10^{-5} for 10 steps), only 10^{-5} passed. We moved to 8B rather than adding 4B seeds because the hosting service withdrew the 4B base model from training during the study.

6 RESULTS

TSA reduces verdict flips. Table 1 reports every method in all six settings. TSA lowers the flip rate significantly in four settings: by 1.12 pp (13% relative) in the main setting, 1.49 pp (7%) on AVeriTeC with the principled prompt, and 1.52 pp (35%) and 1.66 pp (20%) on CONFACT. Accuracy is maintained: TSA's accuracy is equal to or slightly higher than the baseline's in all six settings (+0.2 to +1.2 pp; Table 5 in Appendix D). TSA + consistency (λ=0.4) gives the largest reduction in the main setting (−1.60 pp, 18% relative) and a lower mean flip rate than the baseline in all six settings, two of them significantly. The result does not hinge on the 0.7 confidence threshold: TSA's estimate is negative at every threshold from 0 to 0.9 and significant from 0.5 to 0.7 (Appendix F).

TSA outperforms reward-based alternatives. Neither trust-aware reward changes the flip rate in the main setting (+0.76 and −0.12 pp, both intervals include zero), and the consistency reward alone does not reduce it at any weight. Direct comparisons confirm the gap: in the main setting, TSA's flip rate is lower than that of the trust-aware reward by 1.86 pp⋆, of the trust-aware reward with abstention by 0.99 pp⋆, and of the consistency reward by 1.58 pp⋆ (λ=0.1) and 1.91 pp⋆ (λ=0.4; at λ=0.2 the difference, −0.92 pp, is not significant). Across the six settings, TSA is significantly better than the trust-aware reward in five. TSA is not better everywhere: on SciFact with the principled prompt, the consistency reward alone gives a lower flip rate. Overall, supervising the verdict directly on both versions of a claim is more effective than rewarding trust-appropriate behaviour.

TSA keeps the trust response that the contract asks for. A method could reduce flips by making the model ignore the labels altogether, which would also remove the useful confidence and search responses. Table 2 shows that TSA does not do this. Its confidence response (−0.67) and search response (−0.36) do not change significantly, and accuracy (62.3% vs. 61.7%) and the NEI rate (30.0% vs. 31.0%) stay close to the baseline. The reduction is concentrated on high-trust packs (−1.29 pp⋆), where a verdict change is hardest to justify, with a smaller reduction on low-trust packs (−1.03 pp) and none on mixed packs (−0.12 pp). In contrast, TSA + consistency (λ=0.4) weakens the search response (−1.00⋆), another reason to prefer plain TSA.

TSA generalizes to unseen perturbations. A method trained on the label swap might only learn to ignore that particular change. Figure 4 evaluates TSA on the two perturbations that it never saw. Under label removal, TSA lowers the flip rate by 1.35 pp⋆ ($[-2.01, -0.72]$), slightly more than on the swap it was trained on. Under the downshift, the reduction keeps its direction (−0.57 pp, $[-1.52, +0.36]$). TSA therefore meets a strict criterion for generalization: a reduction under both

Table 2: All outputs at 4B in the main setting (AVeriTeC, neutral prompt). Δ columns are differences from the GRPO baseline. The verdict columns should decrease: flip rate and its two parts, reversals and abstention flips (pp). The trust-response columns should *not* decrease: the confidence response and the search response measure how strongly confidence and search move in the intended direction under the swap (pp; baseline values in the first row). ⋆: 95% interval excludes zero.

	Verdict (should not change)				Trust response (should stay)		Quality	
Method	Flip (%)	Δ flip [95% CI]	Δ rev.	Δ abst.	Δ conf.	Δ search	Acc. (%)	NEI (%)
GRPO baseline	8.92 ± 1.11	–	–	–	7.03	6.62	61.7 ± 2.3	31.0 ± 5.0
Trust-aware reward	9.67 ± 1.31	$+0.76$ [-0.61, $+2.06$]	$+0.44$	$+0.32$	$+0.26$	$+0.44$	62.5 ± 2.5	30.5 ± 5.8
Trust-aware + abstention	8.80 ± 0.32	-0.12 [-1.06, $+0.74$]	$+0.17$	-0.29	$+0.19$	-0.21	63.4 ± 2.2	29.1 ± 4.5
Consistency, λ=0.1	9.39 ± 1.60	$+0.46$ [-1.10, $+2.25$]	$+0.70^\star$	-0.24	$+1.07^\star$	$+0.36$	61.8 ± 0.2	30.9 ± 0.4
Consistency, λ=0.2	8.73 ± 1.05	-0.19 [-1.53, $+1.03$]	$+0.14$	-0.34	$+0.23$	-0.19	61.4 ± 1.8	31.0 ± 4.0
Consistency, λ=0.4	9.71 ± 1.17	$+0.79$ [-0.63, $+2.10$]	$+0.76$	$+0.03$	$+0.94^\star$	-0.20	61.8 ± 2.0	30.3 ± 4.5
TSA (ours)	7.80 ± 0.15	$-1.12^\star$[-2.04, -0.32]	-0.01	$-1.11^\star$	-0.67	-0.36	62.3 ± 0.7	30.0 ± 1.7
TSA + consistency, λ=0.2	7.77 ± 0.82	$-1.15^\star$[-2.32, -0.05]	-0.25	-0.90	$+0.25$	-0.27	61.8 ± 1.2	33.4 ± 3.1
TSA + consistency, λ=0.4	7.32 ± 1.13	$-1.60^\star$[-2.90, -0.18]	-0.22	$-1.38^\star$	-0.76	$-1.00^\star$	61.2 ± 1.0	32.4 ± 3.2

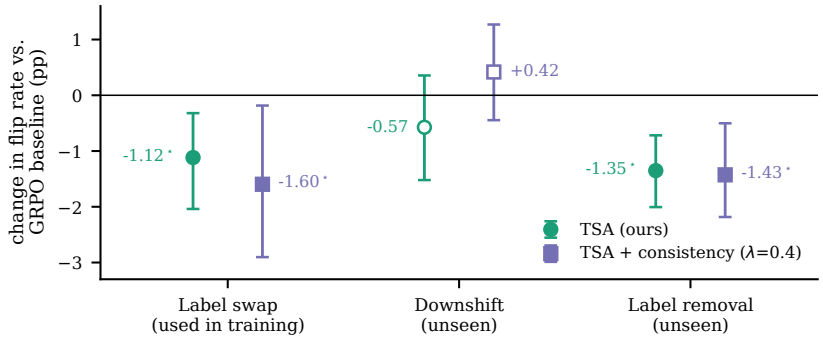

Figure 4: Change in flip rate relative to the 4B GRPO baseline (main setting) under the label swap used in training and under two label perturbations never seen in training. Filled markers: 95% interval excludes zero. TSA's gain carries over to label removal and keeps its direction under the downshift.

unseen perturbations, significant for at least one (details in Appendix G). TSA + consistency also carries over to label removal (-1.43 pp*) but not to the downshift, where it is 0.99 pp* worse than TSA ($[+0.74, +1.37]$, three of three seeds). The explicit consistency term thus buys a larger gain on the trained-on perturbation at the cost of robustness to new ones, which is why we recommend plain TSA.

What each component changes. The flip-type columns of Table 2 show how the methods differ. TSA removes abstention flips (-1.11 pp*, 16% relative) while its confidence response stays unchanged: when HIGH sources become LOW, the TSA model more often keeps its verdict and still lowers its confidence, moving the expression of doubt from the verdict to the confidence as the contract asks. Reversals are unchanged (-0.01 pp); no method in our study significantly reduces reversals under the swap. Nor does TSA fix the persistent failures: the 29 claims (3%) whose verdict flips in at least four of five baseline runs still flip under TSA, 23 of them in all three runs, so its gain comes from claims whose verdict flips in only some runs (Appendix E). The consistency reward acts mainly on confidence: it strengthens the confidence response by 15% ($+1.07^\star$ at λ=0.1) and 13% ($+0.94^\star$ at λ=0.4), while slightly increasing reversals at λ=0.1 ($+0.70$ pp*). Group structure matters: with the same consistency reward and compute, separate groups instead of one joint group lower the flip rate by 2.38 pp* at λ=0.4 (-0.96 pp, not significant, at λ=0.2).

Scaling to 8B. Table 3 repeats the main comparison on Qwen3-8B. The diagnosis carries over and becomes stronger: the 8B models have a larger confidence response than the 4B models (10.2–12.3 vs. about 7), and GRPO training raises the flip rate from 10.5% to 15.0%, entirely through abstention flips. TSA keeps the confidence response and accuracy at 8B, but its reduction in flips

Table 3: Qwen3-8B in the main setting (five matched seeds; learning rate re-selected on the baseline, §5.1). Δ flip compares each method with the 8B GRPO baseline on matched seeds. All six settings are in Appendix H.

Qwen3-8B	Flip rate (%)	Δ flip [95% CI]	Rev. / abst. (%)	Conf. resp.	Acc. (%)	NEI (%)
Untrained	10.5	–	4.2 / 6.3	10.2	67.0	16.0
GRPO baseline	15.0 ± 1.2	–	2.6 / 12.5	11.8	65.1 ± 0.6	26.0 ± 1.0
TSA (ours)	14.7 ± 1.6	-0.36 [-2.21, $+1.33$]	3.0 / 11.7	12.3	64.2 ± 0.8	24.7 ± 1.7
TSA + consistency, λ=0.4	14.3 ± 1.3	-0.74 [-2.49, $+0.58$]	3.2 / 11.1	11.8	64.4 ± 0.3	24.1 ± 1.3

is not detectable (-0.36 pp, $[-2.21, +1.33]$; TSA + consistency -0.74 pp, $[-2.49, +0.58]$) in any setting (Appendix H), and both methods slightly increase reversals ($+0.39$ pp* and $+0.66$ pp*). At this scale, the amplification caused by GRPO itself ($+4.6$ pp) is larger than the effect of either mitigation at our training budget. One untested explanation is that the 8B runs use a four times smaller learning rate for the same 25 steps, which weakens the extra signal from the swapped groups.

7 Discussion

Recommendations for evaluation. Measuring each output separately showed that the shortcut is confined to the verdict and that the consistency reward changes confidence rather than the verdict. Testing on unseen perturbations separated the method that generalizes (TSA) from the one that mainly fits the training perturbation (TSA + consistency). This result motivates evaluation on perturbations not used during training. Contrast sets provide a related approach to testing model behaviour beyond the original test set (Gardner et al., 2020).

Choosing a contract. Whether abstaining under low trust is acceptable depends on the deployment. A system that forwards abstentions to human fact-checkers may prefer the weaker contract, in which only reversals are errors; a system whose verdicts are published directly, or whose labels can be manipulated, needs the stricter one. Because TrustSwap reports reversals and abstentions separately, it supports both choices. Under the stricter contract, TSA removes a sixth of label-driven abstentions at 4B; under either contract, reversals remain the open problem.

Toward larger models. At 8B, GRPO itself amplifies the shortcut and TSA does not offset it within our 25-step, 311-claim budget. Natural next steps are longer training on more claims, training data in which labels and verdicts are not correlated, and designs that keep labels out of the verdict computation while passing them to the confidence and search outputs.

Limitations. We study one model family at two sizes, one RL algorithm with short LoRA training, and a single-pass pipeline in which the search flag is not executed. The unseen perturbations were run at 4B in the main setting only, and new 4B arms have three seeds. The 8B evaluation used the hosted sampling service and the 4B evaluation used local inference, so we compare methods only within a model size. Trust labels come from heuristic domain ratings, and the training pool correlates LOW labels with REFUTES. Veri-R1 is evaluated with our prompt format rather than its own pipeline, so its numbers show that the shortcut exists, not how it performs in deployment. Our metrics are behavioural and have not been validated against human judgements. Finally, the contract is a design choice, which is why we report reversals separately.

8 Conclusion

TrustSwap shows that fact-checking models use source-reliability labels correctly for confidence and search, but also let them decide the verdict, and that GRPO fine-tuning can amplify this. Trust-swap augmentation reduces the shortcut at 4B without hurting accuracy or the useful trust responses, outperforms reward-based alternatives, and generalizes to an unseen perturbation. Extending these gains to larger models and to verdict reversals is the main open problem that TrustSwap makes measurable.

REFERENCES

Sahar Abdelnabi, Kai Greshake, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90, 2023. doi: 10.1145/3605764.3623985. URL https://doi.org/10.1145/3605764.3623985.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f514cec81cb148559cf475e7426eed5e-Paper.pdf.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. FEVEROUS: Fact extraction and VERification over unstructured and structured information. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/68d30a9594728bc39aa24be94b319d21-Abstract-round1.html.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25f7be9694d7b32d5cc670927b8091e1-Abstract-Conference.html.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/file/0184b0cd3cfb185989f858a1d9f5c1eb-Paper.pdf.

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. CrAM: Credibility-aware attention modification in LLMs for combating misinformation in RAG. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23760–23768, 2025. doi: 10.1609/aaai.v39i22.34547. URL https://ojs.aaai.org/index.php/AAAI/article/view/34547.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pp. 1307–1323. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.117. URL https://doi.org/10.18653/v1/2020.findings-emnlp.117.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor (eds.), *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pp. 219–226. ACM, 2019. doi: 10.1145/3306618.3317950. URL https://doi.org/10.1145/3306618.3317950.

Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented LLMs. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9656–9664, 2025. doi: 10.24963/ijcai.2025/1073. URL https://www.ijcai.org/proceedings/2025/1073.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206, 2022. doi: 10.1162/TACL_A_00454. URL https://doi.org/10.1162/tacl_a_00454.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Qi He, Cheng Qian, Xiusi Chen, Bingxiang He, Yi R. Fung, and Heng Ji. Veri-R1: Toward precise and faithful claim verification via online reinforcement learning, 2025. URL https://arxiv.org/abs/2510.01932.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. doi: 10.1609/aaai.v32i1.11694. URL https://doi.org/10.1609/aaai.v32i1.11694.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Kumar Singh, and Mohit Bansal. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3441–3460. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.309. URL https://aclanthology.org/2020.findings-emnlp.309/.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09516.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc.,

2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Media Bias/Fact Check. Media bias/fact check: Source ratings, n.d. URL https://mediabiasfactcheck.com/. Ratings obtained through the CONFACT repository, https://github.com/zoeyyes/CONFACT.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. Not all contexts are equal: Teaching LLMs credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19844–19863, 2024. doi: 10.18653/v1/2024.emnlp-main.1109. URL https://aclanthology.org/2024.emnlp-main.1109/.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with checklist. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.442. URL https://doi.org/10.18653/v1/2020.acl-main.442.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 65128–65167. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cd86a30526cd1aff61d6f89f107634e4-Paper-Datasets_and_Benchmarks.pdf.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16196–16208, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8710ef761bbb29a6f9d12e4ef8e4379c-Abstract.html.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL https://aclanthology.org/P17-2067/.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/99261adc8a6356b38bcf999bba9a26dc-Abstract-Conference.html.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

ETHICS STATEMENT

This work studies how fact-checking models respond to source-reliability labels, with the goal of making such systems harder to manipulate. The label perturbations could in principle guide an attacker who controls labels; we report aggregate behaviour only and release no attack tooling beyond the perturbation scripts needed for replication. All datasets are public research datasets. Source-reliability ratings encode judgements that can themselves be contested, which is one reason we argue that they should not decide verdicts.

REPRODUCIBILITY STATEMENT

The anonymous repository contains the diagnostic construction scripts, the label perturbations, inference and metric code, reward and training code (including TSA and the consistency reward), prompts, and unit tests. Appendix A lists all hyperparameters. All numbers and figures are computed from per-example prediction files by stdlib-only scripts. We will release all trained LoRA adapters, prediction files, and analysis scripts with the paper. This matters because the 4B base model is no longer available for training on the hosted service we used, so released adapters and predictions are the stable way to reproduce the 4B results.

AI USE STATEMENT

We used AI assistants for writing support only, namely grammar correction, wording, and proofreading. The authors reviewed all text and take full responsibility for the content of the paper.

A EXPERIMENTAL DETAILS

Training. All arms use LoRA of rank 32, GRPO with 16 claims per step and $G{=}8$ rollouts per group, 25 steps, a KL penalty of 0.1 towards the base model, sampling temperature 1.3, a maximum response length of 256 tokens, and class resampling of 40/30/30 over REFUTES/SUPPORTS/NEI. The learning rate is 4×10^{-5} at 4B and 1×10^{-5} at 8B. The base reward is $R = R_{\text{format}} + R_{\text{label}} + 0.5\, R_{\text{evidence}}$. The trust-aware reward adds a trust-behaviour term (weight 0.3; confidence, search, and conflict handling should track source trust and evidence sufficiency), a calibration term (0.2; $1 - |p - \mathbb{1}[\text{correct}]|$), an entailment term (0.4; the verdict should match an NLI model's prediction, strictly when the NLI signal is clear and loosely otherwise), and a penalty (0.3) when the verdict disagrees with the NLI prediction on packs that contain LOW-trust sources. None of these terms uses the swapped pack. The abstention variant adds a denser reward for abstaining on claims whose evidence is weak according to the NLI model. In TSA, each swapped version is an additional GRPO group with its own advantage normalisation. In TSA + consistency, the two groups of a claim exchange R_{cons} after both have been sampled. The consistency reward alone uses one joint group of $2G$ rollouts per claim. Training ran on the Tinker hosted service; a 25-step 4B run cost about US$15.

Evaluation. 4B models were evaluated with local Hugging Face inference on one GPU; 8B models with the hosted sampling API. Both use greedy decoding with at most 256 new tokens. For 8B, we checked that the hosted prompt renderer produces byte-identical prompts to the local chat template with thinking disabled. None of the 478,325 prediction records in our evaluation files lacks a valid verdict.

Training data. The 311 training claims are 211 CONFACT claims (all REFUTES) and 100 SciFact claims, selected to exclude every evaluation claim by identifier (one CONFACT claim text appears in both sets). Every training claim with a LOW-trust source is labelled REFUTES, and every SUPPORTS and NEI claim has only HIGH-trust sources.

B DATA CONSTRUCTION AND PROMPTS

Trust labels are assigned per source domain: a manual list of about 150 domains takes precedence (peer-reviewed journals, government, and wire services as HIGH; tabloids and social media as LOW; mainstream news as MEDIUM), followed by Media Bias/Fact Check credibility ratings (about 6,100 domains, distributed with the CONFACT repository) and top-level-domain rules for unlisted `.gov` and `.edu` domains. Other domains are UNKNOWN. The AVeriTeC evidence is built from the question–answer evidence of each claim, keeping claims with at least two evidence items and at least one labelled source. The perturbed packs are identical to the original except for the labels.

Neutral prompt. `You are a fact-checking assistant. Given a claim and a list of evidence pieces (with their source domains and a 'trust' label per source), determine whether the claim is SUPPORTED by the evidence, REFUTED by it, or has Not Enough Information (NEI). [claim] [evidence] Respond using EXACTLY this format: verdict / confidence / needs_more_search / evidence_ids_used / reasoning.`

Principled prompt adds: `The verdict is determined by the EVIDENCE CONTENT. The 'trust' label is NOT a truth signal. A HIGH-trust source can be wrong; a LOW-trust source can be right. Do not let the trust label flip the verdict. Use the 'trust' label to CALIBRATE your confidence ... Use the 'trust' label to decide needs_more_search: if the only evidence supporting/refuting the claim comes from low-trust sources, set needs_more_search: yes rather than committing to a definitive verdict. If evidence is weak or absent regardless of trust, prefer NEI over SUPPORTS/REFUTES.`

C FULL DIAGNOSIS

Table 4: Label-swap results for all models and settings. Flip: flip rate (%); Rev.: reversal rate (% of confident pairs); Δp and Δm: confidence and search shift (pp) on high-trust (H) and low-trust (L) packs. GRPO rows are means over five seeds. SciFact has only high-trust packs.

Setting	Model	Flip	Rev.	Δp H	Δp L	Δm H	Δm L
AVeriTeC N	Qwen3-4B	7.6	1.2	−6.6	+7.6	+6.2	−6.1
	Qwen3-4B + GRPO	8.9	1.9	−7.3	+6.0	+6.7	−6.1
	Qwen3-8B	10.5	4.2	−11.0	+6.7	+11.7	−5.3
	Qwen3-8B + GRPO	15.0	2.6	−12.4	+8.9	+17.3	−10.6
	Veri-R1	20.3	2.6	−14.3	+9.8	+17.6	−13.1
AVeriTeC P	Qwen3-4B	19.3	1.6	−24.5	+22.8	+19.6	−20.6
	Qwen3-4B + GRPO	20.0	1.4	−25.1	+22.8	+22.2	−22.2
	Qwen3-8B	15.8	2.3	−27.1	+25.2	+73.6	−73.6
	Qwen3-8B + GRPO	22.6	2.3	−27.6	+26.3	+73.7	−75.8
	Veri-R1	25.0	6.0	−13.3	+14.2	+18.5	−20.6
SciFact N	Qwen3-4B	4.1	0.0	−5.5	–	+5.9	–
	Qwen3-4B + GRPO	3.8	0.6	−4.2	–	+3.4	–
	Qwen3-8B	6.2	3.1	−5.8	–	+3.1	–
	Qwen3-8B + GRPO	12.7	1.9	−7.9	–	+10.8	–
	Veri-R1	15.8	4.8	−11.1	–	+11.0	–
SciFact P	Qwen3-4B	9.6	0.7	−19.7	–	+8.8	–
	Qwen3-4B + GRPO	10.1	0.8	−21.3	–	+10.5	–
	Qwen3-8B	7.2	1.4	−28.2	–	+60.9	–
	Qwen3-8B + GRPO	10.4	0.5	−27.3	–	+57.2	–
	Veri-R1	22.3	5.4	−13.2	–	+16.9	–
CONFACT N	Qwen3-4B	5.9	3.1	−3.0	+2.0	+2.6	−2.9
	Qwen3-4B + GRPO	4.4	1.7	−2.8	+0.8	+2.1	−0.6
	Qwen3-8B	4.9	3.1	−4.6	+2.4	+3.1	−2.9
	Qwen3-8B + GRPO	6.7	2.5	−4.7	+2.4	+3.9	−2.9
	Veri-R1	34.6	2.2	−23.0	+16.8	+33.1	−17.6
CONFACT P	Qwen3-4B	8.5	1.8	−13.0	+2.3	+13.7	−3.4
	Qwen3-4B + GRPO	8.2	1.8	−13.7	+3.2	+18.0	−1.5
	Qwen3-8B	10.2	2.5	−23.6	+6.7	+62.0	−23.1
	Qwen3-8B + GRPO	14.1	1.7	−24.0	+8.4	+60.7	−29.1
	Veri-R1	50.0	7.3	−26.8	−11.0	+41.9	+20.0

D ACCURACY IN ALL SETTINGS AND PILOT RESULTS

Table 5: Label accuracy (%) on the original evidence at 4B, mean over seeds, in all six settings.

Method	AVeriTeC N	AVeriTeC P	SciFact N	SciFact P	CONFACT N	CONFACT P
GRPO baseline	61.7	57.1	71.9	69.5	71.6	64.5
Trust-aware reward	62.5	56.8	71.5	69.1	71.5	63.8
Trust-aware + abstention	63.4	58.4	72.1	68.7	71.8	64.7
Consistency, λ=0.1	61.8	56.4	72.3	70.2	71.9	65.6
Consistency, λ=0.2	61.4	56.9	71.8	69.8	72.1	65.0
Consistency, λ=0.4	61.8	57.6	72.5	69.8	72.0	67.0
TSA (ours)	62.3	57.4	72.3	69.7	72.1	65.7
TSA + consistency, λ=0.2	61.8	56.3	71.5	70.0	71.4	64.8
TSA + consistency, λ=0.4	61.2	56.8	71.8	69.3	71.7	65.9

Table 6: Per-class accuracy (%, mean ± s.d. over seeds) at 4B in the main setting, on the original and the swapped packs. No class collapses under TSA.

	Original pack			Swapped pack		
Method	SUPPORTS	REFUTES	NEI	SUPPORTS	REFUTES	NEI
GRPO 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
TSA (ours)	50.0 ± 1.5	70.5 ± 1.5	65.1 ± 1.6	46.5 ± 1.6	70.3 ± 1.9	66.9 ± 2.0
TSA + consistency, 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

Flip counts. TSA has slightly more confident pairs than the baseline (842 vs. 807 on average), so its lower flip rate also corresponds to fewer flips in absolute terms (65.7 vs. 71.8 per run). We report rates because the number of confident pairs varies across seeds (739–916 for the baseline).

Trust-aware rewards in single-seed pilots. In our own single-seed pilot runs, the two trust-aware rewards appeared to reduce label-driven verdict changes and, for the abstention variant, confident verdicts on weakly supported claims (by more than 10 pp in one seed). With five seeds, neither effect holds: Table 2 shows no change in the flip rate, and the abstention variant changes the rate of confident verdicts on weakly supported claims by $+2.0$ pp ($[-6.8, +8.6]$) on AVeriTeC and $+2.9$ pp ($[-3.4, +7.9]$) on SciFact relative to the trust-aware reward. This is why all comparisons in this paper use multiple seeds.

E Persistent failures

We group the 984 AVeriTeC claims (main setting) by how many of the five baseline runs flip their verdict. Of the 29 claims that flip in at least four baseline runs, 19 are high-trust packs, 5 are low-trust packs, and 5 are mixed packs; 24 flip mainly through abstention and 5 through reversal. TSA flips 23 of them in all three runs and 6 in one or two runs; TSA + consistency flips 15 in all three runs, 13 in one or two, and 1 in none. Because the total number of flips falls under TSA (65.7 vs. 71.8 per run) while these persistent claims keep flipping, TSA's gain comes from claims whose verdict flips in only some runs. A per-group comparison of flip counts would be biased by regression to the mean, since claims are grouped using the baseline runs, so we report only this aggregate conclusion.

F Sensitivity to the confidence threshold

Table 7: Change in flip rate (pp) relative to the 4B GRPO baseline in the main setting when confident pairs are defined with different thresholds on $\max(p, p')$ (0: all pairs). Pairs: mean number of pairs kept for the baseline. $\star$: 95% interval excludes zero.

Threshold	Pairs	Baseline (%)	TSA	TSA + cons. (λ=0.4)	Trust-aware reward	Consistency (λ=0.4)
0.0	984	7.32	−0.65	−0.91*	+0.35	−0.07
0.5	841	8.56	−1.07*	−1.44*	+0.59	+0.98
0.6	828	8.70	−1.03*	−1.54*	+0.80	+0.92
0.7	807	8.92	−1.12*	−1.60*	+0.76	+0.79
0.8	683	10.06	−1.21	−1.51*	+0.48	−0.55
0.9	622	8.32	−0.38	−0.60	+1.02	−0.12

TSA's estimate is negative at every threshold; at 0.8 and 0.9 fewer pairs remain and the intervals widen. Unlabeled packs, which the swap does not change, contribute no flips; excluding them raises the baseline flip rate from 8.92% to 9.20% and leaves TSA's reduction unchanged (-1.14 pp*).

G UNSEEN PERTURBATIONS IN DETAIL

Table 8: Flip rates and flip types at 4B in the main setting under the trained-on swap and the two unseen perturbations. Reversal and abstention are percentages of confident pairs. $\star$: 95% interval excludes zero.

Perturbation	Method	Pairs	Flip (%)	Reversal (%)	Abstention (%)	Δ flip vs. baseline
Label swap (trained on)	GRPO baseline	807	8.92 ± 1.11	1.93	6.99	–
	TSA (ours)	842	7.80 ± 0.15	1.91	5.89	$-1.12^{\star}[-2.04, -0.32]$
	TSA + consistency, λ=0.4	865	7.32 ± 1.13	1.71	5.61	$-1.60^{\star}[-2.90, -0.18]$
Downshift (unseen)	GRPO baseline	795	4.55 ± 0.99	1.37	3.18	–
	TSA (ours)	829	3.98 ± 0.57	1.25	2.73	-0.57 $[-1.52, +0.36]$
	TSA + consistency, λ=0.4	854	4.97 ± 0.38	1.24	3.73	$+0.42$ $[-0.45, +1.27]$
Label removal (unseen)	GRPO baseline	810	7.86 ± 0.59	1.99	5.87	–
	TSA (ours)	845	6.51 ± 0.56	1.68	4.83	$-1.35^{\star}[-2.01, -0.72]$
	TSA + consistency, λ=0.4	874	6.43 ± 0.79	1.53	4.90	$-1.43^{\star}[-2.18, -0.50]$

Under both unseen perturbations, TSA has fewer reversals and fewer abstention flips than the baseline (point estimates). The downshift changes fewer labels than the swap, so all flip rates are lower. The extra flips of TSA + consistency under the downshift are abstention flips (3.73% vs. 3.18% for the baseline).

H QWEN3-8B DETAILS

Table 9: Qwen3-8B in the main setting. Top: the unchanged 4B recipe (learning rate 4×10^{-5}) makes training unstable, so we re-selected the learning rate on the baseline arm only. Bottom: the selected recipe (10^{-5}), five matched seeds per arm.

Arm	Accuracy (%)	NEI rate (%)	Flip rate (%)	Δ vs. baseline [95% CI]	Reversal / abstention (%)
Unchanged 4B recipe (learning rate 4×10^{-5})					
GRPO baseline	59.1 ± 5.1	32.4 ± 13.8	16.6 ± 5.5	–	–
TSA	51.2 ± 6.0	45.9 ± 11.0	27.0 ± 5.4	–	–
TSA + consistency	52.5 ± 5.2	45.7 ± 8.0	27.1 ± 3.4	–	–
Selected recipe (learning rate 10^{-5}), five matched seeds					
Untrained Qwen3-8B	67.0	16.0	10.5	–	4.2 / 6.3
GRPO baseline	65.1 ± 0.6	26.0 ± 1.0	15.0 ± 1.2	–	2.6 / 12.5
TSA	64.2 ± 0.8	24.7 ± 1.7	14.7 ± 1.6	-0.36 $[-2.21, +1.33]$	3.0 / 11.7
TSA + consistency	64.4 ± 0.3	24.1 ± 1.3	14.3 ± 1.3	-0.74 $[-2.49, +0.58]$	3.2 / 11.1

Table 10: Qwen3-8B in all six settings (selected recipe). Flip rates in %; Δ columns are matched-seed differences in pp. GRPO training raises the flip rate above the untrained model in every setting (all five seeds). Neither intervention changes the flip rate significantly in any setting. The only significant method-versus-method contrast is TSA + consistency minus TSA on CONFACT with the principled prompt (-1.19, $[-1.90, -0.43]$), a different setting from the one significant contrast at 4B; we treat it as exploratory.

Setting	Untrained	GRPO baseline	TSA	TSA + cons.	Δ TSA	Δ TSA + cons.
AVeriTeC N	10.5	15.0 ± 1.2	14.7 ± 1.6	14.3 ± 1.3	-0.36	-0.74
AVeriTeC P	15.8	22.6 ± 0.4	22.6 ± 0.8	22.3 ± 0.8	$+0.05$	-0.30
SciFact N	6.2	12.7 ± 1.6	13.1 ± 1.1	13.4 ± 1.0	$+0.33$	$+0.64$
SciFact P	7.2	10.4 ± 0.9	10.7 ± 2.2	10.2 ± 2.4	$+0.30$	-0.18
CONFACT N	4.9	6.7 ± 1.0	6.8 ± 0.5	7.1 ± 0.7	$+0.07$	$+0.39$
CONFACT P	10.2	14.1 ± 1.3	14.7 ± 0.9	13.5 ± 0.6	$+0.57$	-0.62

Confidence response at 8B. At 8B, TSA does not weaken the confidence response relative to the baseline ($+0.54$, $[-0.11, +1.16]$), while TSA + consistency has a weaker response than TSA ($-0.50^{\star}$, $[-0.83, -0.24]$, five of five seeds), in line with our recommendation of plain TSA.