Logit-Based Universal Trigger Search for Attacking Claim Verification Models

Anonymous ACL submission

Abstract

Despite their widespread adoption, the robustness of fact-checking models to adversarial perturbations remains underexplored. Existing approaches are typically model-specific and require gradient access or dataset-dependent optimization, with implications for generalization, efficiency, and semantic validity. We introduce FACTFLIP, a framework for analyzing robustness in claim verification models via universal adversarial triggers. FACTFLIP identifies highly perturbative trigger words through a lightweight, model-only analysis of classification logits, without relying on training data or gradient access. FACTFLIP decouples trigger discovery from claim perturbation and adopts an LLM-based perturb-and-verify pipeline to integrate them while preserving semantic validity. Experimental results show that FACTFLIP effectively exposes model vulnerabilities, achieving competitive attack success rates with greater stability and cross-model robustness than fully supervised baselines. Moreover, we show that the identified triggers are highly discriminative and exhibit compositional effects, providing evidence of systematic biases arising from both pre-training and fine-tuning.

1 Introduction

An increasing share of the population relies on online sources as a primary means of accessing information (Reuters Institute, 2024). At the same time, online information environments are characterized by a growing prevalence of false or misleading content, amplified by the scale and speed of digital dissemination (WEF, 2024; Vosoughi et al., 2018). Ensuring the correctness of publicly accessible information is a critical societal challenge, motivating the development of tools to support fact-checking at scale (IFCN, 2025; Mirza et al., 2023).

Several approaches in the fact-checking literature evaluate robustness by modifying claims or evidence through input-specific transformations. These transformations operate at different levels, including character-level noise (Mamta and Cocarascu, 2025), word-level substitutions (Kim and Allan, 2019; Mamta and Cocarascu, 2025; Hidey et al., 2020), and syntactic changes (Thorne et al., 2019; Mamta and Cocarascu, 2025; Hidey et al., 2020), as well as content reformulations generated by LLMs, such as paraphrasing (Thorne et al., 2019; Atanasova et al., 2020; Niewinski et al., 2019; Magomere et al., 2025), colloquialization (Kim et al., 2021), and dialectal variation (Magomere et al., 2025).

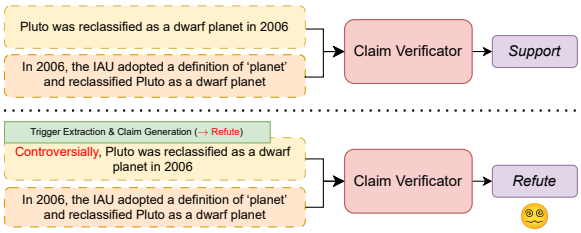

Figure 1: Effect of adversarial triggers on claim verification outcomes.

Other approaches instead study robustness by inducing prediction flips through the insertion of *adversarial triggers*, referred to as *universal* when they are input-agnostic, as illustrated in Figure 1. In this setting, triggers are tokens that systematically bias model predictions toward a target class when injected into otherwise unchanged inputs (Wallace et al., 2019). While this paradigm aims to reveal model-level vulnerabilities and global decision biases, existing methods (Shin et al., 2020; Atanasova et al., 2020; Ebrahimi et al., 2018) typically rely on gradient-based optimization over a reference dataset. Such access to training data and gradients is often unrealistic in practical or adversarial settings, where training data may be proprietary or unavailable, or models are provided via black-box APIs. Moreover, this reliance raises concerns re-

garding (i) generalization, as discovered triggers may reflect dataset-specific artifacts rather than intrinsic model vulnerabilities; (ii) computational efficiency, due to iterative token search and costly gradient backpropagation; and (iii) semantic validity, since direct token insertion can yield unnatural or meaning-altering claims (Morris et al., 2020; Zhou et al., 2024; Atanasova et al., 2020).

In this paper, we introduce FACTFLIP[1], a framework for the discovery of Universal Adversarial Triggers that addresses the limitations of prior approaches. Unlike gradient-based methods, FACTFLIP identifies highly perturbative triggers through a simple, non-contextual analysis of model logits. Specifically, we feed a target fact-checking model with individual trigger words and rank them for each class based on the resulting logits, enabling trigger discovery that depends solely on the target model. This design mitigates generalization concerns and substantially improves computational efficiency, as well as making FACTFLIP applicable to black-box APIs that expose token log probabilities. To ensure semantic validity, FACTFLIP incorporates an LLM-based framework that integrates selected tokens coherently into claims (*LLM as a Perturber*) and verifies entailment constraints between the original and perturbed claims (*LLM as a Verifier*). While similar generate-and-verify paradigms have been explored in prior work (Magomere et al., 2025), we employ this framework to explicitly enforce entailment preservation in adversarial fact-checking (Atanasova et al., 2020), ensuring that perturbations preserve the original claim meaning while remaining linguistically natural and realistic.

To assess the quality of our approach, we evaluate FACTFLIP on six fine-tuned versions of `RoBERTa` (Liu et al., 2019) and on `Qwen2.5-14B-Instruct` (Qwen, 2024) used as a zero-shot classifier. Leveraging a model-only, logit-based method for discovering Universal Adversarial Triggers, without relying on gradient-based optimization, allows us to obtain four main principled insights into the robustness and failure modes of fact-checking models. In particular, our analysis shows that (RQ1) while supervised, gradient-based approaches for trigger extraction achieve higher attack success rates when access to labeled data is available, FACTFLIP attains competitive effectiveness without requiring any training data and exhibits substantially greater stability and cross-model robustness (§4.1); (RQ2) enforcing linguistic plausibility and entailment preservation introduces a clear trade-off between semantic validity and attack strength, with lexically sensitive models remaining vulnerable even under strict constraints (§4.2); (RQ3) the discovered triggers are highly discriminative and exhibit compositional effects (§4.3); and (RQ4) universal adversarial triggers provide a powerful diagnostic lens, uncovering systematic biases rooted in both pre-training and fine-tuning rather than dataset-specific artifacts (§4.4).

[1] `https://anonymous.4open.science/r/FactFlip-8DCE`

2 Related Work

Adversarial Attacks in NLP. Adversarial text generation is widely used to assess the robustness of NLP models. Early approaches such as SEAR (Ribeiro et al., 2018) relied on rule-based perturbations to generate grammatically altered but semantically equivalent texts. Subsequent work extended this paradigm through controlled modifications of text properties, such as sentiment changes (Ribeiro et al., 2020; Alzantot et al., 2018; Jia et al., 2019), or through character-level (Gao et al., 2018; Li et al., 2019; Boucher et al., 2021) and synonym-based lexical substitutions guided by cosine similarity (Jin et al., 2020; Li et al., 2019), masked language modeling (Garg and Ramakrishnan, 2020; Li et al., 2020, 2021), or external knowledge bases (Zang et al., 2020; Ren et al., 2019). More recent work leverages gradient information in supervised settings to identify target input elements and perform perturbations (Ebrahimi et al., 2018; Yoo and Qi, 2021; Yoo et al., 2020; Hou et al., 2023; Yuan et al., 2023). In particular, HotFlip (Ebrahimi et al., 2018) is an instance-based gradient attack that identifies character- or word-level substitutions maximizing the loss for a given input via a first-order approximation.

Gradient-based approaches have been extended to discover universal adversarial triggers, i.e., sequences of tokens that can induce misclassification when inserted into any arbitrary input (Wallace et al., 2019). Representative approaches in this category include AutoPrompt (Shin et al., 2020), which performs an iterative, gradient-guided search over trigger tokens and serves as a gradient-based baseline in our experimental evaluation, as well as methods with dedicated generation mechanisms to improve fluency while preserving effectiveness (Song

et al., 2021; Guo et al., 2021; Xu and Wang, 2024).

Model Robustness for Claim Verification. Several approaches adapt adversarial attack techniques to assess the robustness of claim verification models. Thorne et al. (2019) evaluate robustness through rephrasing strategies that introduce linguistic patterns absent from the training data, including synonym substitutions derived from WordNet and rule-based transformations. Other work examines robustness under more realistic perturbation scenarios, such as discursive claims (Kim et al., 2021), misinformation edits (Magomere et al., 2025), i.e., perturbations that frequently occur on social media, character-level or grammatical-structure modifications (Mamta and Cocarascu, 2025; Hidey et al., 2020), and evidence omission (Atanasova et al., 2022; Abdelnabi and Fritz, 2023). Schuster et al. (2019, 2021) show that negation is strongly associated with the *refute* label in the FEVER (Thorne et al., 2018) dataset, and that models trained on FEVER learn to exploit such correlations, which can be leveraged to mislead their predictions.

More recent work leverages LLM-based paraphrasing to generate adversarial claims, either via unconstrained rewriting with post-hoc label validation (Magomere et al., 2025), or through supervised, gradient-based optimization to enforce label cohesion (Niewinski et al., 2019). In contrast to these approaches, we explicitly incorporate universal triggers into generated claims to exploit model-intrinsic lexical biases. Unlike Atanasova et al. (2020), who jointly discovers and applies triggers via supervised, gradient-based fine-tuning of a generative model, we identify triggers directly from model logits without fine-tuning and decouple trigger discovery from claim generation. In §4.1, we evaluate the effectiveness of the triggers found by FACTFLIP against a gradient-based trigger discovery method that underlies Atanasova et al. (2020)'s approach and use our perturb-and-verify pipeline for controlled claim generation.

3 Approach

FACTFLIP follows a two-stage pipeline (see Figure 2), consisting of an *extraction* stage for identifying perturbative words and a *generation* stage for producing a well-formed perturbed claim.

3.1 Extraction Stage

Given a model M, a vocabulary V, and a set of claim–evidence–label tuples (c_i, e_i, l_i), with

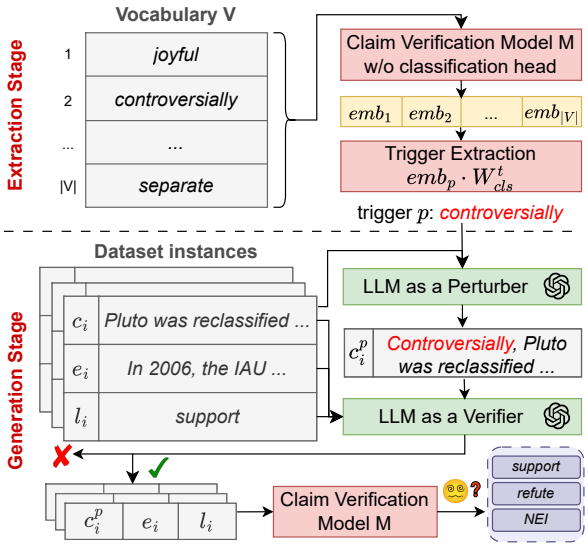

Figure 2: FACTFLIP's pipeline. FACTFLIP first identifies the most perturbing triggers (*Extraction Stage*), then integrates and validates the found trigger inside the original claim (*Generation Stage*).

$l_i \in \{Support, Refute, NEI\}$, our goal is to identify universal trigger words $p \in V$ that, when inserted into a claim c_i, increase the likelihood of misclassification toward a target class $t \neq l_i$.

To identify the most perturbative triggers, we leverage the class-specific weight vectors of the model's classification head as reference decision directions. In particular, for the *Support* and *Refute* labels, given the classification head h_{cls}, we compute the inner product between the non-contextualized embedding emb_p of each candidate trigger word $p \in V$ and the corresponding class-specific weight vector W^t_{cls}. For the *NEI* label, we instead score p by averaging the *NEI* logit obtained from the templates "*The statement "p" is true*" and "*The statement "p" is false*". This controls a known bias toward *NEI*: tokens with weak semantics can spuriously increase the *NEI* logit. The templates add informative context, so high-scoring words are more likely to be genuinely perturbative toward *NEI* rather than merely uninformative. Finally, the embedding emb_p is computed using the embedding of the first token.

The resulting score $emb_p \cdot W^t_{\text{cls}}$ measures the alignment of a word with the decision direction associated with class t: higher values indicate words that, in isolation, bias the model toward that class. We exclude the bias term of h_{cls} to focus on directional alignment rather than global class priors. We hypothesize that words exhibiting strong alignment with a target class in this non-contextualized setting

can retain their perturbative effect when inserted into a claim, thereby increasing the probability of misclassification. This trigger extraction procedure depends only on the target model M and does not require access to training data, as well as being applicable to both encoder-based classifiers and zero-shot generative models. In the latter case, the labels correspond to verbalized output tokens (*"support"*, *"refute"*, *"not enough information"*), and trigger alignment is computed with respect to the logits of the label-specific initial tokens, yielding an equivalent trigger ranking procedure.

In FACTFLIP, V is a controlled vocabulary of 6,621 elements, consisting of words and their antonyms extracted from WordNet. This constrained design is motivated by three considerations: (i) excluding words, such as negations, that can alter the semantic validity of a claim and invalidate its original label (Morris et al., 2020; Zhou et al., 2024; Atanasova et al., 2020); (ii) including antonyms ensures coverage of both semantic polarities of the same concept, enabling a controlled analysis of sensitivity to meaning variations; and (iii) operating at the word level facilitates coherent integration into claims and avoids the interpretability issues associated with token-level perturbations (Ebrahimi et al., 2018; Shin et al., 2020).

3.2 Generation Stage

Once the trigger word p_{chosen} is identified, we use `gpt-4o-mini` as an *LLM as a Perturber* to integrate the trigger p into the original claim c_i, producing a perturbed claim c_i^p. A critical requirement of adversarial evaluation in fact-checking is that the ground-truth label must remain unchanged. If a perturbed claim alters the original entailment relation with the evidence, it becomes impossible to distinguish between a genuine model error and a correct prediction on a semantically different claim. For this reason, the goal of the generation stage is not merely to produce fluent adversarial claims, but to ensure that the semantic relation between the claim and the evidence, corresponding to the original label l_i, is preserved. This setting also reflects a realistic adversarial scenario, in which an attacker aims to maintain the intended meaning of a claim while inducing an incorrect model prediction.

To enforce these constraints, we use the same `gpt-4o-mini` model as an *LLM as a Verifier* to check whether the perturbed claim c_i^p satisfies specific entailment relations with respect to c_i and e_i. We define the binary relation $\Rrightarrow$ to represent the claim verification relation, where $e_i \Rrightarrow c_i$ indicates support and $e_i \Rrightarrow \neg c_i$ indicates refutation, and we use $\models$ for logical entailment.

(i) If $l_i = \textit{Support}$, the original claim must logically entail the perturbed claim. Since $e_i \Rrightarrow c_i$ and $c_i \models c_{i,\text{perturb}}$, it follows that $e_i \Rrightarrow c_{i,\text{perturb}}$.

(ii) If $l_i = \textit{Refute}$, the perturbed claim must logically entail the original claim. Since $e_i \Rrightarrow \neg c_i$ and $c_{i,\text{perturb}} \models c_i$, by contraposition it follows that $e_i \Rrightarrow \neg c_{i,\text{perturb}}$.

(iii) If $l_i = \textit{NEI}$, the original and perturbed claims must logically entail each other, ensuring that no new supporting or refuting information is introduced with respect to the evidence.

Perturbed claims that violate these entailment constraints are discarded. While logical entailment is stricter than the standard *Support*, *Refute* and *NEI* relations used in claim verification, this choice guarantees that the resulting adversarial examples preserve the original ground-truth label, at the cost of an higher number of claims being discarded. Indeed, evaluations that do not enforce label preservation can inadvertently count semantically invalid perturbations as successful attacks, inflating attack success (Morris et al., 2020; Zhou et al., 2024; Atanasova et al., 2020). Starting from a dataset of claim verification tuples (c_i, e_i, l_i), this process yields a new dataset (c_i^p, e_i, l_i) composed of semantically valid adversarial claims that include the selected trigger p. Details about the prompts used and `gpt-4o-mini` parameters are reported in §G.

4 Experimental evaluation

The experimental evaluation addresses four key research questions. RQ1 (Effectiveness) assesses the effectiveness of FACTFLIP in discovering universal adversarial triggers, comparing it against strong dataset-specific baselines (§4.1). RQ2 (Plausibility vs. Attack Strength) investigates the trade-off between linguistic plausibility and attack success rate by evaluating performance under increasing constraints on the naturalness of perturbed claims (§4.2). RQ3 (Trigger Discrimination and Composition) examines FACTFLIP ability to distinguish between highly and weakly perturbative triggers and analyzes the compositional effects of injecting multiple triggers on attack performance (§4.3). RQ4 (Model Diagnosis) analyzes how FACTFLIP can be used as a diagnostic tool to reveal sources of bias in fact-checking models, including biases stemming from pre-training and fine-tuning (§4.4).

For the *LLM as a Verifier* and *Perturber*, we have qualitatively checked the reliability of the generation and verification step and found `gpt-4o-mini` to be consistently satisfactory, thus we use it as both the perturber and verifier in all experiments.

Datasets. We consider six datasets commonly used in the literature to evaluate claim veracity: *AVERITEC* (Schlichtkrull et al., 2023), *SciFact* (Wadden et al., 2020), *HoVer* (Jiang et al., 2020), *FM2* (Eisenschlos et al., 2021), *PolitiHop* (Ostrowski et al., 2021), and *VitaminC* (Schuster et al., 2021). The *NEI* label appears only in *AVERITEC*, *SciFact*, and *VitaminC*. We drop dataset-specific labels such as *Conflicting Evidence* (*AVERITEC*) and *half-true* (*PolitiHop*) to keep the label space consistent across datasets. Moreover, we keep only instances the model predicts correctly, up to 500 per dataset to reduce API costs. Additional dataset statistics are provided in §G.

Models. We test FACTFLIP on two pre-trained models: `roberta-base` (Liu et al., 2019) and `Qwen-2.5-14B-Instruct` (Qwen, 2024). For each dataset, we fine-tune a separate `roberta-base` model, whereas `Qwen-2.5-14B-Instruct` is employed in a zero-shot setting following the experimental setup of Lin et al. (2025), in which Qwen models achieved the best performance among open-source alternatives. In this setting, we avoid using Chain-of-Thought prompting (Wei et al., 2022; Kojima et al., 2022), as it has been shown to provide limited gains in fact verification (Zhang and Gao, 2023) also due to the hallucination of generated rationales (Pan et al., 2023). Detailed model performance results are reported in §B.4.

Metrics. We evaluate our method using the Attack Success Rate (ASR), which measures the effectiveness of an attack in flipping model predictions towards a target class t. ASR is computed over input samples that are initially correctly classified and whose ground-truth label is not t. It is measured as the proportion of such samples whose predicted label is changed to t after the attack.

Baselines. We consider two baselines to evaluate FACTFLIP: AutoPrompt, a gradient-based and inherently supervised method for universal trigger discovery, and FACTFLIP-DS, a dataset-specific variant of our approach introduced to control for the effect of supervision when comparing against supervised baselines for trigger extraction.

AutoPrompt. AutoPrompt constructs adversarial prompts using a Hotflip-style gradient search. While AutoPrompt was originally proposed to identify triggers that elicit correct model predictions, we invert the trigger optimization objective, following Zou et al. (2023), to discover adversarial triggers. Moreover, we adapted AutoPrompt to identify word-level triggers composed of multiple subword tokens by averaging gradient alignment across their constituent embeddings. Full implementation details and hyperparameters are provided in §B.1. In the experiments, AutoPrompt is run for 5 iterations, 10 gradient accumulation steps with a batch size of 32 for `RoBERTa`, and 1 for `Qwen`.

FACTFLIP-*DS (dataset-specific).* AutoPrompt relies on direct access to labeled data to compute gradients of the loss with respect to token embeddings, and is therefore supervised by construction. In contrast, FACTFLIP is fully unsupervised and dataset-agnostic for the trigger extraction. To account for this difference, we introduce FACTFLIP-DS, a dataset-specific variant of our method that incorporates limited supervision while preserving the trigger discovery strategy. FACTFLIP-DS starts from the FACTFLIP's trigger ranking, selects the top-50 trigger candidates, and retains the top-5 triggers that maximize ASR on a validation set, matching the development set size used by AutoPrompt.

4.1 RQ1: Effectiveness

This experiment evaluates the effectiveness of FACTFLIP in discovering universal adversarial triggers, comparing it against AutoPrompt and FACTFLIP-DS. Results are reported in Table 1. For comparability, the top-5 triggers identified by each method are inserted individually into each input, and the average ASR is reported. Additional AutoPrompt experiments with increased gradient accumulation are reported in §C.

Focusing first on the relative difficulty of inducing different target classes, a consistent pattern emerges: flips toward *Refute* are systematically easier and achieve the highest ASR on average, followed by *NEI*, while *Support* remains the hardest class to induce. This ordering is stable across approaches and reflects an intrinsic property of claim verification, as refutation relies on more salient lexical and semantic cues than support, making its decision boundaries easier to cross under adversarial perturbations (Atanasova et al., 2020; Schuster et al., 2019). Second, we observe clear model-dependent robustness patterns. Across models, `Qwen` appears moderately more robust than

	Dataset	*Support*				*Refute*				*NEI*				*Avg* Δ
		FF	FF-DS	Auto-Prompt	Δ	FF	FF-DS	Auto-Prompt	Δ	FF	FF-DS	Auto-Prompt	Δ	
roberta-base	AVERITEC	1.81	0.41	1.44	-1.03	8.33	44.00	5.00	+39.00	0.16	1.43	0.72	+0.71	+12.89
	SciFact	1.23	0.00	2.67	-2.67	14.52	28.57	22.58	+5.99	3.95	5.66	2.17	+3.49	+2.27
	HoVer	1.02	1.55	1.11	+0.44	80.88	88.74	93.58	-4.84	–	–	–	–	-2.20
	FM2	5.26	4.47	8.72	-4.25	11.68	19.05	28.26	-9.21	–	–	–	–	-6.73
	PolitiHop	1.24	0.93	0.00	+0.93	6.67	16.67	0.00	+16.67	–	–	–	–	+8.80
	VitaminC	0.22	0.66	0.00	+0.66	0.75	4.24	18.73	-14.49	9.08	15.84	15.28	+0.56	-4.42
	Average	1.80	1.34	2.32	-0.98	20.47	33.55	28.03	+5.52	4.40	7.64	6.06	+1.58	**+1.77**
	Std. Dev.	1.62	1.48	3.00	1.91	27.35	27.49	30.9	18.12	3.66	6.05	6.55	1.35	7.07
Qwen-2.5-14B	AVERITEC	1.14	0.56	0.96	-0.40	2.38	6.06	1.43	+4.63	5.24	9.29	8.44	+0.85	+1.69
	SciFact	1.65	2.38	0.00	+2.38	0.00	0.00	2.17	-2.17	1.28	8.05	13.92	-5.87	-1.89
	HoVer	1.40	3.41	6.01	-2.60	28.80	11.71	19.05	-7.34	–	–	–	–	-4.97
	FM2	0.80	1.20	0.54	+0.66	29.41	34.88	41.67	-6.79	–	–	–	–	-3.07
	PolitiHop	0.49	1.25	0.31	+0.94	20.00	14.29	20.00	-5.71	–	–	–	–	-2.39
	VitaminC	3.31	4.07	7.22	-3.15	12.50	0.00	8.33	-8.33	7.91	7.92	16.50	-8.58	-6.69
	Average	1.47	2.15	2.51	-0.36	15.52	11.16	15.44	-4.28	4.81	8.42	12.95	-4.53	**-2.89**
	Std. Dev.	0.91	1.26	2.94	1.96	11.63	11.89	13.81	4.43	2.72	0.62	3.36	3.96	2.61

Table 1: ASR comparison between FACTFLIP (FF) and competitors across models and datasets. Values are reported per target class (*Support*, *Refute*, *NEI*), corresponding to the attack target label. Results are averaged over the top-5 most perturbing words. The Δ column reports the ASR difference between FF-DS and AutoPrompt, the last column shows the average ASR change across classes, and "–" indicates the absence of the *NEI* class.

RoBERTa, particularly against *Refute* flips, likely due to its larger scale and instruction-tuned pretraining, which promote stronger semantic consistency and reduce reliance on localized lexical cues.

Turning to supervised approaches for trigger extraction, AutoPrompt and FACTFLIP-DS show no consistent winner across settings: AutoPrompt clearly emerges on Qwen (with a performance increase over FACTFLIP-DS of 2.89 points), whereas FACTFLIP-DS outperforms AutoPrompt on RoBERTa (with an average gain of 1.77 points).

Finally, examining the trade-off between effectiveness and robustness clarifies the role of universal trigger discovery. While supervised dataset-specific approaches tend to achieve higher ASR on average, this advantage relies on access to target data and labels, which is often unrealistic in adversarial settings, and make trigger discovery computationally more expensive, especially for gradient-based approaches. In contrast, the universal FACTFLIP exhibits stable behavior across datasets and models, with consistently lower variance than supervised baselines. The difference is most pronounced in the *Refute* class in the RoBERTa → Qwen cross-model setting, where FACTFLIP drops by 4.95 ASR points, compared to much larger declines for FACTFLIP-DS (22.39 points) and AutoPrompt (12.59 points). This suggests that supervision amplifies model- and dataset-specific artifacts, whereas the unsupervised variant captures intrinsic model vulnerabilities, resulting in slightly lower but more consistent attack effectiveness.

Takeaway #1. Supervised dataset-specific triggers yield higher ASR but require target data and vary widely across models. FACTFLIP is slightly less effective yet more stable, providing a transferable assessment of adversarial weaknesses, capturing intrinsic model vulnerabilities.

4.2 RQ2: Plausibility vs. Attack Strength

Enforcing natural and semantically coherent perturbations is necessary to model realistic adversarial scenarios, but it restricts the trigger space and can reduce attack effectiveness. We analyze this trade-off through two controlled settings: raw trigger injection and similarity-based trigger selection. An analysis of the semantic filtering effect induced by the *LLM as a verifier* is reported in §D.2.

Raw trigger injection. We evaluate FACTFLIP in an ablation setting where claim rewriting is disabled and triggers are directly injected at the beginning of the original claims (FF-RAW). This setting

	Dataset	FF-RAW		FF-SIM	
		Δ_{ASR}	$\Delta_{BERTScore}$	Δ_{ASR}	$\Delta_{BERTScore}$
roberta-base	AVERITEC	+5.80	-6.10	+0.37	+0.60
	SciFact	-3.20	-4.45	-4.30	+5.12
	HoVer	+0.85	-5.90	+1.70	+1.13
	FM2	+2.65	-5.40	+7.90	+1.08
	PolitiHop	+10.60	+0.49	-0.45	+3.38
	VitaminC	+1.10	+1.40	+0.57	+3.66
	Average	**+2.97**	**-3.33**	**+0.96**	**+2.50**
Qwen2.5-14B	AVERITEC	+0.43	-5.26	+1.23	+2.32
	SciFact	+2.47	-5.42	+0.53	+4.06
	HoVer	-0.50	-6.44	-2.45	-0.21
	FM2	+2.80	-6.35	-1.25	+2.74
	PolitiHop	-1.05	+1.11	-7.15	+1.38
	VitaminC	-2.97	-0.62	-1.00	+2.67
	Average	**+0.20**	**-3.83**	**-1.68**	**+2.16**

Table 2: Impact of raw (FF-RAW) and similarity (FF-SIM) trigger injection on ASR (Δ_{ASR}) and semantic similarity ($\Delta_{BERTScore}$). Positive Δ_{ASR} indicates stronger attacks than FACTFLIP, while negative $\Delta_{BERTScore}$ reflects reduced semantic plausibility than FACTFLIP.

	Dataset	*Support*	*Refute*	*NEI*	*Avg* Δ
roberta-base	AVERITEC	+1.81	+2.92	0.00	+1.58
	SciFact	+1.23	+12.14	+2.23	+5.20
	HoVer	+0.68	+58.22	–	+29.45
	FM2	-1.57	+4.90	–	+1.67
	PolitiHop	+1.24	+6.67	–	+3.96
	VitaminC	-0.53	+0.52	+15.05	+5.01
	Average	+0.48	+14.23	+5.79	**+7.81**
Qwen-2.5-14B	AVERITEC	-0.07	+0.18	+1.06	+0.39
	SciFact	0.00	-1.52	-0.06	-0.53
	HoVer	-1.31	+24.68	–	+11.69
	FM2	+0.40	+24.94	–	+12.67
	PolitiHop	-0.61	+14.12	–	+6.76
	VitaminC	-2.78	+9.11	+4.49	+3.61
	Average	-0.72	+11.92	+1.83	**+5.77**

Table 3: Difference in ASR between the top-5 high and low perturbing words identified by FACTFLIP.

measures ASR without enforcing linguistic plausibility, providing an upper bound on attack effectiveness. Results are reported in Table 2. Removing the rewriting phase generally increases ASR, with consistent gains for RoBERTa (average $\Delta = +2.97$). In contrast, Qwen shows smaller and less consistent improvements (average $\Delta = +0.2$). This behavior can be attributed to stronger semantic robustness and the absence of task-specific fine-tuning, which makes the model less sensitive to surface-level trigger patterns. These gains come at the cost of semantic distortion: FF-RAW consistently reduces BERTScore for both models (-3.33 for RoBERTa and -3.83 for Qwen). Overall, FF-RAW maximizes ASR but reduces plausibility.

Similarity-based trigger selection. We analyze the effect of selecting triggers semantically similar to the original claim to improve plausibility while constraining the trigger space. The FF-SIM variant selects the three most similar triggers (by cosine similarity with the claim embedding) from the top-200 perturbing candidates, injecting each individually and averaging ASR across them. Results are reported in Table 2 (a more detailed analysis is provided §D.1). Enforcing semantic coherence reduces attack effectiveness for Qwen, leading to a noticeable ASR decrease (average -1.68), but substantially improves semantic preservation, with consistent BERTScore gains (average $+2.16$). In contrast, RoBERTa shows moderate ASR gains (average $+0.96$) while also improving semantic similarity (average $\Delta_{\text{BERTScore}} = +2.50$). We attribute this behavior to RoBERTa's higher sensitivity to surface-level lexical patterns, which allows semantically aligned triggers to more effectively exploit spurious correlations learned during fine-tuning.

Takeaway #2. Enforcing plausibility often reduces ASR, but lexically sensitive models can still be misled by well-integrated triggers.

4.3 RQ3: Trigger Discrimination and Composition

This section evaluates FACTFLIP ability to rank triggers by perturbation strength and to model their compositional effects on attack performance.

Adversarial impact of ranked triggers. We evaluate how well FACTFLIP ranks triggers according to their adversarial potential by quantifying the difference in impact between highly and minimally perturbing words. To this end, we select both top-ranked perturbing words and least impactful ones, injecting each individually into the original claims to measure their effect on model predictions. Table 3 reports the resulting ASR differences between perturbing and unperturbing triggers. Complete results are provided in §E.1. The results show that the proposed ranking separates perturbing from less perturbing triggers on average. Across models, perturbing words yield higher ASR margins overall (Avg Δ: $+7.81$ for RoBERTa and $+5.77$ for

Bias	*Support*	*Refute*	*NEI*	*Avg*
Religion	1.7 ± 2.3	24.9 ± 33.4	11.1 ± 14.9	12.6
Politics	1.0 ± 0.8	23.2 ± 34.1	8.4 ± 13.4	10.9
Gender	1.8 ± 2.7	21.5 ± 30.9	5.7 ± 7.7	9.7
Race	1.1 ± 0.9	20.6 ± 26.7	5.0 ± 6.0	8.9
Sexual	1.7 ± 1.9	17.0 ± 21.3	1.2 ± 0.6	6.6

Table 4: Mean ASR and std. dev. for each bias category.

		Support	*Refute*	*NEI*	*Avg*
roberta-base	AVERITEC	-3.70	-9.20	-1.59	-4.83
	SciFact	+7.43	+1.62	+1.51	+3.52
	HoVer	-37.44	-3.21	–	-20.33
	FM2	+23.64	-2.11	–	+10.77
	PolitiHop	+26.08	+18.84	–	+22.46
	VitaminC	-0.02	-1.55	+8.92	+2.45
	Average	+2.66	+0.73	+2.95	+2.34

Table 5: Mean difference in occurrence frequency between the top and bottom 5 perturbing triggers.

Qwen), with the largest gains for *Refute* (+14.23 and +11.92). Differences for *Support* are small and often negative, suggesting that for more robust predictions the injected triggers have limited influence regardless of their ranking.

Multi-trigger compositionality. We examine the effect of injecting three triggers simultaneously. Full results are reported in §E.2. While trigger composition generally increases ASR, it substantially reduces the number of claims passing verification. On average, the proportion of verified claims drops by 46.3% for RoBERTa and 43.5% for Qwen compared to single-trigger perturbations. This reveals a clear trade-off between perturbation strength and linguistic plausibility, as increasingly unnatural claims are filtered by the verifier.

Takeaway #3. Highly ranked triggers yield the strongest adversarial effects, and composing multiple triggers further increases ASR at the cost of reduced linguistic plausibility.

4.4 RQ4: Model Diagnosis

Assessing biases affecting minority demographic groups. We demonstrate the use of FACTFLIP as a diagnostic tool to probe model sensitivity to socially salient lexical cues across five bias dimensions: *gender*, *religion*, *sexual orientation*, *politics*, and *race*. The analysis evaluates whether inserting bias-related terms into claims induces prediction flips, possibly causing representational harm (Blodgett et al., 2020). Experiments are conducted on RoBERTa with claim rewriting disabled. For each bias category, we construct a dedicated vocabulary (see §F.1) and rank triggers by adversarial impact, revealing the cues that most strongly influence model predictions. Results are summarized in Table 4, with detailed results in §F.1. Religion- and political-related terms show the strongest effects, sexual orientation cues the weakest, and gender and racial terms intermediate sensitivity. Moderate standard deviations indicate consistent behavior across datasets.

Discovering sources of adversarial sensitivity. This use case examines whether sensitivity to universal adversarial triggers originates from fine-tuning data or pre-training. We test for dataset-specific lexical biases by analyzing whether highly perturbing triggers occur more frequently in training claims of the target class than minimally perturbing ones (see Table 5). Across datasets and labels, this association is weak and inconsistent. For a fine-tuned RoBERTa model, highly perturbing words are only slightly more frequent on average (mean difference: 2.34), with no systematic pattern even in high-ASR settings (e.g., *Refute* on HoVer and FM2). To further isolate the role of pre-training, we compare perturbing triggers from a standard fine-tuned RoBERTa with those obtained after re-initializing the pre-training weights and fine-tuning from scratch. As shown in §F.2, lexical overlap is negligible for both highly and minimally perturbing triggers. This indicates that adversarial sensitivity does not emerge solely by fine-tuning dynamics or pre-training, but by their interaction.

Takeaway #4. Universal trigger vulnerability emerges from the interaction of pre-training and fine-tuning, not from either in isolation.

5 Conclusion

We introduced FACTFLIP, a framework for analyzing the robustness of fact-checking models via universal adversarial triggers. FACTFLIP relies on a lightweight, model-only logit analysis and an LLM-based perturb-and-verify pipeline to generate semantically valid adversarial claims. Experiments across multiple datasets and models show that FACTFLIP effectively exposes systematic vulnerabilities with greater stability than fully supervised baselines, as well as different sources of bias.

6 Limitations

Our trigger extraction assumes access to model internals (e.g., classifier-head parameters) or, in generative zero-shot settings, access to label-token logits. As a result, FACTFLIP is not applicable to fully black-box APIs that do not expose token log probabilities.

Because triggers are ranked using non-contextual alignment, FACTFLIP may under-rank triggers whose adversarial effect is highly context-dependent, which can reduce ASR in some cases. This behavior is expected: FACTFLIP targets universal triggers intended to expose general vulnerabilities rather than instance-specific weaknesses. However, incorporating contextualized representations to improve ASR is a promising direction, which we leave to future work.

Finally, the generation stage depends on an external LLM (perturber/verifier), which incurs additional cost and latency and may introduce variability across prompts and model versions.

7 Ethical considerations

Our framework could be exploited by malicious actors to craft well-engineered claims aimed at fooling existing automated claim verification systems. However, our goal is strictly diagnostic: we introduce FACTFLIP to systematically expose model weaknesses and failure modes, enabling future work to develop more robust and reliable claim verification models. This work is not intended for, nor do we endorse, adversarial misuse.

The data we generate with the perturb-and-verify pipeline exclusively integrates the found triggers with the original claims of the datasets we tested. These datasets may contain texts that refer to individual people. Although the integration of the triggers may inadvertently alter the meaning of the original texts, we do not perform any steps to protect/anonymize the data we generate, as (i) the introduced perturbation is minimal by definition; (ii) it would be impossible to guarantee reproducibility in real world scenarios, as personally identifiable information is crucial for the claim verification task; and (iii) it would be difficult to evaluate FACTFLIP and compare it with other approaches, as the datasets we used have been published with personally identifiable information.

8 Use of AI assistants

When writing this paper, we used ChatGPT to improve the flow of writing and the vocabulary of the initial drafts we manually wrote. Each suggestion has been manually validated by the authors.

References

Sahar Abdelnabi and Mario Fritz. 2023. Fact-saboteurs: a taxonomy of evidence manipulation attacks against fact-verification systems. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA. USENIX Association.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Nicholas Boucher, Ilia Shumailov, Ross J. Anderson, and Nicolas Papernot. 2021. Bad characters: Imperceptible NLP attacks. *CoRR*, abs/2106.09898.

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

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606, Online. Association for Computational Linguistics.

Bairu Hou, Jinghan Jia, Yihua Zhang, Guanhua Zhang, Yang Zhang, Sijia Liu, and Shiyu Chang. 2023. Textgrad: Advancing robustness evaluation in NLP by gradient-driven optimization. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

IFCN. 2025. State of the fact-checkers report 2024. International Fact-Checking Network.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4129–4142, Hong Kong, China. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

Byeongchang Kim, Hyunwoo Kim, Seokhee Hong, and Gunhee Kim. 2021. How robust are fact checking systems on colloquial claims? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1535–1548, Online. Association for Computational Linguistics.

Youngwoo Kim and James Allan. 2019. FEVER breaker's run of team NbAuzDrLqg. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 99–104, Hong Kong, China. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online. Association for Computational Linguistics.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See-Kiong Ng, and Tat-Seng Chua. 2025. FACT-AUDIT: An adaptive multi-agent framework for dynamic fact-checking evaluation of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 360–381, Vienna, Austria. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Jabez Magomere, Emanuele La Malfa, Manuel Tonneau, Ashkan Kazemi, and Scott A. Hale. 2025. When claims evolve: Evaluating and enhancing the robustness of embedding models against misinformation edits. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22374–22404,

Vienna, Austria. Association for Computational Linguistics.

Mamta Mamta and Oana Cocarascu. 2025. FactEval: Evaluating the robustness of fact verification systems in the era of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10647–10660, Albuquerque, New Mexico. Association for Computational Linguistics.

Muhammad Shujaat Mirza, Labeeba Begum, Liang Niu, Sarah Pardo, Azza Abouzied, Paolo Papotti, and Christina Pöpper. 2023. Tactics, threats & targets: Modeling disinformation and its mitigation. In *30th Annual Network and Distributed System Security Symposium, NDSS*. The Internet Society.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 20–26, Hong Kong, China. Association for Computational Linguistics.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-hop fact checking of political claims. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3892–3898. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Team Qwen. 2024. Qwen2.5: A party of foundation models.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Reuters Institute. 2024. Digital news report 2024. Reuters Institute for the Study of Journalism.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: a dataset for real-world claim verification with evidence from the web. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. 2021. Universal adversarial attacks with natural triggers for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3724–3733, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*

Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2944–2953, Hong Kong, China. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

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)*, pages 7534–7550, Online. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

WEF. 2024. The global risks report 2024. World Economic Forum.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yue Xu and Wenjie Wang. 2024. LinkPrompt: Natural and universal adversarial attacks on prompt-based language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6473–6486, Mexico City, Mexico. Association for Computational Linguistics.

Jin Yong Yoo, John Morris, Eli Lifland, and Yanjun Qi. 2020. Searching for a search method: Benchmarking search algorithms for generating NLP adversarial examples. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 323–332, Online. Association for Computational Linguistics.

Jin Yong Yoo and Yanjun Qi. 2021. Towards improving adversarial training of NLP models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 945–956, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lifan Yuan, YiChi Zhang, Yangyi Chen, and Wei Wei. 2023. Bridge the gap between CV and NLP! a gradient-based textual adversarial attack framework. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7132–7146, Toronto, Canada. Association for Computational Linguistics.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

Huichi Zhou, Zhaoyang Wang, Hongtao Wang, Dongping Chen, Wenhan Mu, and Fangyuan Zhang. 2024. Evaluating the validity of word-level adversarial attacks with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4902–4922, Bangkok, Thailand. Association for Computational Linguistics.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

A Appendix overview

This appendix provides extended experimental results and methodological details complementing the analyses presented in the main paper. For clarity and consistency, the appendix is structured to mirror the four research questions (RQ1–RQ4) investigated in the experimental section, with additional preliminary material covering experimental setup and baselines.

Appendix B details the experimental foundations shared across all research questions. It describes the adaptation of AutoPrompt to the fact verification setting and reports the baseline performance of the evaluated fact-checking models on the original test sets.

Appendix C extends the analysis of RQ1 (Effectiveness) by reporting additional results for AutoPrompt under increased gradient accumulation.

These experiments, omitted from the main paper, provide further evidence of the instability and overfitting behavior of gradient-based baselines in this setting.

Appendix D supports RQ2 (Plausibility vs. Attack Strength) with a detailed study of the LLM-as-a-verifier component. It quantifies the proportion of perturbed claims filtered out by the semantic consistency check and analyzes its differential impact across target classes and models.

Appendix E expands on RQ3 (Trigger Discrimination and Composition) by reporting full attack success rates for both highly perturbing and minimally perturbing triggers, as well as comprehensive results on multi-trigger compositionality. This section provides absolute ASR values and additional statistics that substantiate the ranking and compositional claims discussed in the main paper.

Appendix F complements RQ4 (Model Diagnosis) with extended case studies on demographic bias and adversarial sensitivity. It includes detailed breakdowns of bias-related triggers across datasets and classes, examples of bias vocabularies, and an analysis of lexical overlap aimed at disentangling the effects of pre-training and fine-tuning.

Appendix G illustrates the prompts used for the perturb-and-verify pipeline, as well as the prompt used to perform claim verification with `Qwen2.5-14B-Instruct`. Moreover, this section provides additional statistics on the composition of the generated datasets.

B Experimental setup and baselines

B.1 AutoPrompt adaptation

This section describes the adaptation of AutoPrompt (Shin et al., 2020) to the fact verification setting considered in this work. AutoPrompt automatically constructs prompts (Wallace et al., 2019) through a Hotflip-style gradient search (Ebrahimi et al., 2018). The method introduces n universal trigger tokens initialized as [MASK] at the beginning of the input, which are iteratively updated to maximize the likelihood of a target class t, that is, the class we intend to perturb.

For each input x_i^{train} in a dataset D^{train}, where x_i^{train} is the prompt containing c_i^{train} and e_i^{train} and c_i^{train} is initialized with n [MASK] trigger tokens, AutoPrompt computes the gradient with respect to the j-th trigger token w_{trig}^j and identifies a set of candidate tokens $w \in V_{cand}$ from the model vocabulary that maximize the probability of t:

$$V_{cand} = \underset{w \in V}{\text{top-}k} \left[w_{emb}^T \nabla \log p(t \mid x_i^{train}) \right] \quad (1)$$

where w_{emb}^T denotes the embedding of token w, and the gradient is taken with respect to w_{trig}^j. Then, each token $w \in V_{cand}$ replaces the j-th trigger token inside each $x_i^{dev} \in D^{dev}$. The perturbation performance of each candidate w is then evaluated on D^{dev}, and the best-performing tokens are retained for the next iteration

$$V_{trig} = \underset{w \in V_{cand}}{\text{top-}k} \left[\text{PScore}(w; D^{dev}) \right] \quad (2)$$

with PScore being the *ASR* of trigger token w on D^{dev}.

Unlike AutoPrompt, our objective is to identify words, possibly composed of multiple tokens. Let $\mathcal{W}$ be the set of words, where each word $w \in \mathcal{W}$ is composed of $|w|$ tokens with embeddings emb_{w_i}. Since AutoPrompt's search requires a fixed number of trigger tokens, we divide the vocabulary into m bins B_m according to token length. For each bin B_m, we define the word-level gradient score as

$$s(w) = \frac{1}{|w|} \sum_{i=1}^{|w|} emb_{w_i}^T \nabla \log p(t|x)$$

and select the top-k candidate words as

$$V_{cand}^{(m)} = \text{top-}k_{w \in B_m} [s(w)]$$

retaining the candidate words with the strongest perturbation on D^{dev} for the next iteration (Equation 2). Finally, we unify all the candidate words across the B_m bins $V_{cand} = \bigcup_{m=1}^{M} V_{cand}^{(m)}$ and compute V_{trig} with Equation 2.

B.2 RoBERTa fine-tuning parameters

The `roberta-base` models have been fine-tuned for 30 epochs with learning rate equal to $1e-5$, maximum sentence length equal to 512, weight decay equal to $1e-2$, and class weight to handle label imbalance. We select and report the model checkpoint corresponding to the epoch achieving the highest F1 score on the validation set.

B.3 `gpt-4o-mini` parameters

Every `gpt-4o-mini` generation uses temperature equal to 0 and top_p equal to 1. We set a number of retries equal to 50 to handle OpenAI API failures.

	Dataset	Accuracy	Precision	Recall	F1 Score
roberta-base	AVERITEC	84.42	86.36	85.04	85.66
	SciFact	75.00	69.66	69.73	69.50
	HoVer	79.68	80.02	79.68	79.62
	FM2	81.45	81.51	81.42	81.43
	PolitiHop	87.13	70.53	72.19	71.31
	VitaminC	89.26	85.78	86.15	85.92
Qwen2.5-14B	AVERITEC	62.99	69.74	75.08	60.73
	SciFact	82.33	85.41	80.96	81.57
	HoVer	73.60	73.61	73.60	73.60
	FM2	87.20	87.57	87.37	87.19
	PolitiHop	97.66	93.14	96.62	94.78
	VitaminC	77.80	71.10	70.40	70.66

Table 6: Performance of tested fact checker models on the original test sets. The `roberta-base` model is fine-tuned on each dataset, while `Qwen2.5-14B-Instruct` is tested in zero-shot.

B.4 Fact-checking accuracy of the tested models

We report the baseline performance of the evaluated fact-checking models on the original, unperturbed test sets. This analysis serves two purposes: (i) to verify that all models achieve competitive performance in the clean setting, and (ii) to contextualize the impact of adversarial perturbations reported in subsequent sections. Baseline accuracy varies across datasets and target classes, reflecting differences in dataset size, label distribution, and evidence complexity. The results, reported in Table 6, confirm that both *RoBERTa* and *Qwen* exhibit strong verification capabilities prior to adversarial intervention, ensuring that observed attack effects cannot be attributed to weak baseline performance.

B.5 Computational infrastructure

We report the model sizes and the computational infrastructure used to run the experiments. The `roberta-base` model contains 0.1 billion parameters, while `Qwen2.5-14B-Instruct` contains 14 billion parameters. The experiments with `roberta-base` have been run on a NVIDIA L40S with 46GB. For `Qwen2.5-14B-Instruct`, we loaded the model in half precision (float16) and ran it on one NVIDIA Ampere A100 with 64GB.

C RQ1 - Effectiveness

This section reports additional AutoPrompt results obtained by increasing the number of gradient accumulation steps during trigger optimization (from 10 to 500). As shown in Table 7, increasing gradient accumulation does not consistently improve attack success rates. In several cases, performance deteriorates, particularly on datasets with limited lexical diversity. This behavior suggests that prolonged optimization encourages overfitting to dataset-specific artifacts rather than discovering robust universal triggers. For this reason, these configurations are excluded from the main experimental comparison and reported here solely for completeness.

D RQ2 - Plausibility vs. Attack strength

D.1 Similarity-based trigger selection.

This section extends the analysis presented in the main paper by providing a more fine-grained, class-wise view of attack success rates across datasets and trigger selection strategies. While the main paper focuses on aggregated trends, these extended results make explicit how such trends emerge from systematic shifts in specific decision classes. The results of this evaluation are reported in Table 8 and Table 9. Across both models, the detailed breakdown confirms that most ASR gains, both for `FF-RAW` and `FF-SIM`, are driven primarily by changes in the *Refute* class, while effects on *Support* and *NEI* are generally smaller and less consistent. This pattern is particularly pronounced for *RoBERTa*, reinforcing the interpretation that refutation decisions are more susceptible to lexical and stylistic cues introduced by adversarial triggers. Moreover, the class-wise results clarify the trade-offs observed in the main paper. For `FF-RAW`, large ASR increases are often associated with substantial semantic degradation, whereas `FF-SIM` achieves a more balanced profile, combining moderate attack effectiveness with improved semantic preservation. The consistency between aggregate metrics and class-level behavior supports the robustness of the main conclusions: adversarial effectiveness is strongly shaped by the interaction between trigger plausibility, model sensitivity to surface patterns, and the biases induced by fine-tuning.

D.2 Impact of the LLM-as-a-verifier.

This experiment evaluates the selectivity of the LLM-as-a-verifier by measuring how many perturbed claims are discarded by the semantic filtering step. After trigger injection and claim rewriting, the verifier retains only those claims that preserve the original entailment relation with the evidence, allowing us to quantify the extent to which unconstrained perturbations introduce semantic drift.

		Support			*Refute*			*NEI*			*Avg* Δ
	Dataset	FactFlip	Autoprompt	Δ	FactFlip	Autoprompt	Δ	FactFlip	Autoprompt	Δ	
roberta-base	AVERITEC	1.81	1.45	+0.36	8.33	9.09	-0.76	0.16	0.51	-0.35	-0.25
	SciFact	1.23	3.28	-2.05	14.52	22.45	-7.93	3.95	6.12	-2.17	-4.05
	HoVer	1.02	0.59	+0.43	80.88	81.08	-0.20	–	–	–	+0.12
	FM2	5.26	8.36	-3.10	11.68	12.5	-0.82	–	–	–	-1.96
	PolitiHop	1.24	0	+1.24	6.67	0	+6.67	–	–	–	+3.96
	VitaminC	0.22	0	+0.22	0.75	7.14	-6.39	9.08	14.91	-5.83	-4.00
	Average	1.80	2.28	-0.48	20.47	22.04	-1.57	4.40	7.18	-2.78	**-1.03**
Qwen-2.5-14B	AVERITEC	1.14	0.10	+1.04	2.38	7.41	-5.03	5.24	10.50	-5.26	-3.08
	SciFact	1.65	0.00	+1.65	0.00	5.26	-5.26	1.28	17.86	-16.58	-6.73
	HoVer	1.40	2.40	-1.00	28.8	21.62	+7.18	–	–	–	+3.09
	FM2	0.80	1.04	-0.24	29.41	16.67	+12.74	–	–	–	+6.25
	PolitiHop	0.49	0.57	-0.08	20.00	10.00	+10.00	–	–	–	+4.96
	VitaminC	3.31	7.21	-3.9	12.5	7.15	+5.35	7.91	13.19	-5.28	-1.28
	Average	1.47	1.89	-0.42	15.52	11.35	+4.17	4.81	13.85	-9.04	**+0.54**

Table 7: Attack Success Rate comparison between FACTFLIP and AutoPrompt with 500 accumulation steps.

The statistics reported in Table 10 show that the verifier filters out a large fraction of generated claims. On average, *RoBERTa* rejects 55.8% of Support claims, 86.4% of Refute claims, and 71.6% of NEI claims, while *Qwen* discards 58.4%, 92.8%, and 70.3%, respectively. The particularly high rejection rates for the *Refute* class suggest that, although effective attacks exist, most refutation-oriented perturbations induce semantic inconsistencies that are detected by the entailment check. Overall, these results indicate that the LLM-as-a-verifier acts as a strong semantic regularizer, pruning perturbations that violate entailment consistency and substantially constraining the space of admissible adversarial claims.

E RQ3 - Trigger discrimination and composition

E.1 Adversarial impact of ranked triggers

This section provides detailed results supporting the trigger ranking experiment discussed in the main paper. In addition to reporting ASR differences between highly perturbing and minimally perturbing triggers, in Table 11 we include absolute ASR values for each dataset and class. The results confirm that FACTFLIP consistently assigns higher adversarial impact to triggers that induce larger prediction shifts, while minimally perturbing words yield substantially weaker effects. This separation holds across datasets and target labels, supporting the reliability of the learned trigger ranking.

E.2 Multi-trigger compositionality

This section provides a detailed analysis of the compositional effects arising from simultaneously injecting multiple triggers into the same claim. In particular, we study whether the adversarial impact of individual triggers compounds when combined, and how this interaction affects both attack success and semantic validity.

Table 13 reports ASR obtained by injecting, for each claim, either the three most perturbing triggers or the three least perturbing ones according to FACTFLIP 's ranking. Across both models, combining high-impact triggers generally leads to a substantial amplification of adversarial effectiveness. This effect is especially pronounced for the *Refute* class, where RoBERTa exhibits an average ASR increase of +37.60 points when using perturbing combinations compared to unperturbing ones, while Qwen shows a corresponding gain of +10.33. For *NEI*, the gains are more moderate but consistent across models (+6.93 for RoBERTa and +4.97 for Qwen), whereas *Support* remains largely stable, indicating lower sensitivity to trigger composition.

At the same time, multi-trigger injection has a marked impact on claim validity. As shown in Table 12, injecting three triggers instead of a single one leads to a sharp reduction in the number of claims passing the verification step enforced by the LLM-as-a-verifier. Averaged across datasets and labels, the proportion of verified claims drops by 46.29% ± 20.88 for RoBERTa and by 43.49% ±

	Dataset	FF-RAW				FF-SIM			
		Support	*Refute*	*NEI*	Δ	*Support*	*Refute*	*NEI*	Δ
roberta-base	AVERITEC	$1.2^{-0.6}$	$25.9^{+17.6}$	$0.6^{+0.4}$	+5.80	$0.6^{-1.2}$	$10.6^{+2.3}$	$0.2^{0.0}$	+0.37
	SciFact	$0.3^{-0.9}$	$3.8^{-10.7}$	$6.0^{+2.0}$	-3.20	$1.0^{-0.2}$	$5.8^{-8.7}$	$0.0^{-4.0}$	-4.30
	HoVer	$3.4^{+2.4}$	$80.2^{-0.7}$	–	+0.85	$0.3^{-0.7}$	$85.0^{+4.1}$	–	+1.70
	FM2	$9.8^{+4.5}$	$12.5^{+0.8}$	–	+2.65	$4.7^{-0.6}$	$28.1^{+16.4}$	–	+7.90
	PolitiHop	$0.0^{-1.2}$	$29.1^{+22.4}$	–	+10.60	$0.7^{-0.5}$	$6.3^{-0.4}$	–	-0.45
	VitaminC	$0.8^{+0.6}$	$3.6^{+2.8}$	$9.0^{-0.1}$	+1.10	$0.8^{+0.6}$	$2.4^{+1.6}$	$8.6^{-0.5}$	+0.57
	Average	$2.6^{+0.8}$	$25.9^{+5.4}$	$5.2^{+0.8}$	**+2.97**	$1.4^{-0.4}$	$23.0^{+2.5}$	$2.9^{-1.5}$	**+0.96**
Qwen2.5-14B	AVERITEC	$0.4^{-0.7}$	$4.2^{+1.8}$	$5.4^{+0.2}$	+0.43	$0.5^{-0.6}$	$5.8^{+3.4}$	$6.1^{+0.9}$	+1.23
	SciFact	$0.0^{-1.7}$	$1.6^{+1.6}$	$8.8^{+7.5}$	+2.47	$0.0^{-1.7}$	$0.0^{0.0}$	$4.6^{+3.3}$	+0.53
	HoVer	$1.6^{+0.2}$	$27.6^{-1.2}$	–	-0.50	$2.4^{+1.0}$	$22.9^{-5.9}$	–	-2.45
	FM2	$0.4^{-0.4}$	$35.4^{+6.0}$	–	+2.80	$0.8^{0.0}$	$26.9^{-2.5}$	–	-1.25
	PolitiHop	$0.4^{-0.1}$	$18.0^{-2.0}$	–	-1.05	$0.3^{-0.2}$	$5.9^{-14.1}$	–	-7.15
	VitaminC	$2.6^{-0.7}$	$7.4^{-5.1}$	$4.8^{-3.1}$	-2.97	$6.6^{+3.3}$	$10.0^{-2.5}$	$4.1^{-3.8}$	-1.00
	Average	$0.9^{-0.6}$	$15.7^{+0.2}$	$6.3^{+1.5}$	**+0.20**	$1.8^{+0.3}$	$11.9^{-3.6}$	$4.9^{+0.1}$	**-1.68**

Table 8: Attack success rates under FF-RAW and FF-SIM settings. Superscripts indicate absolute changes from FACTFLIP, and Δ denotes the average variation across labels.

	Dataset	FF-RAW				FF				FF-SIM			
		Support	*Refute*	*NEI*	*Avg*	*Support*	*Refute*	*NEI*	*Avg*	*Support*	*Refute*	*NEI*	*Avg*
roberta-base	AVERITEC	83.36	83.24	84.58	83.73	88.76	89.50	91.21	89.82	90.80	88.75	91.73	90.43
	SciFact	80.42	80.95	80.43	80.60	85.32	85.90	83.93	85.05	91.22	90.11	89.17	90.17
	HoVer	83.47	85.29	–	84.38	88.33	92.22	–	90.28	90.31	92.51	–	91.41
	FM2	82.94	80.23	–	81.59	87.59	86.38	–	86.99	88.70	87.43	–	88.07
	PolitiHop	83.06	84.53	–	83.80	82.10	84.51	–	83.31	86.88	86.48	–	86.68
	VitaminC	82.71	82.09	77.69	80.83	81.97	81.27	75.06	79.43	83.47	82.74	83.08	83.10
	Average	**82.66**	**82.72**	**80.90**	**82.49**	**85.68**	**86.63**	**83.40**	**85.81**	**88.56**	**88.00**	**87.99**	**88.31**
Qwen2.5-14B	AVERITEC	82.09	82.79	82.21	82.36	85.59	89.75	87.53	87.62	89.46	89.93	90.45	89.95
	SciFact	81.26	82.67	81.53	81.82	85.12	88.72	87.89	87.24	90.66	93.23	90.01	91.30
	HoVer	83.71	84.55	–	84.13	88.82	92.33	–	90.58	89.95	90.78	–	90.37
	FM2	77.84	79.16	–	78.50	82.24	87.47	–	84.86	86.22	88.97	–	87.60
	PolitiHop	83.68	86.69	–	85.19	82.97	85.19	–	84.08	86.72	84.20	–	85.46
	VitaminC	77.86	79.16	78.26	78.43	76.60	82.62	77.92	79.05	81.55	81.32	82.27	81.71
	Average	**81.07**	**82.50**	**80.67**	**81.74**	**83.56**	**87.68**	**84.45**	**85.57**	**87.43**	**88.07**	**87.58**	**87.73**

Table 9: BERTScore under FF-RAW, FACTFLIP, and FF-SIM settings, reported per label and averaged across classes.

Class	RoBERTa	Qwen
Support	55.80 ± 7.26	58.42 ± 14.02
Refute	86.41 ± 8.19	92.79 ± 2.12
NEI	71.61 ± 13.13	70.31 ± 16.51
Average	71.27 ± 9.53	73.84 ± 10.89

Table 10: Percentage of claims filtered out by the LLM-as-a-verifier, reported as mean ± standard deviation across all the datasets.

22.51 for Qwen. This reduction is particularly severe for the *Refute* class, where over 80% of claims are filtered out on average, indicating that compositional perturbations frequently introduce semantic distortions that violate entailment consistency with the evidence.

Taken together, these results highlight a clear trade-off induced by trigger composition. While combining multiple high-impact triggers significantly strengthens adversarial attacks, it also increases semantic drift, causing a large fraction of perturbed claims to be rejected by the verifier. Importantly, the remaining claims that do pass verification still achieve non-negligible ASR, showing that compositional attacks can remain effective even under strict semantic constraints. Over-

		Support			*Refute*			*NEI*			*Avg* Δ
	Dataset	pert.	unpert.	Δ	pert.	unpert.	Δ	pert.	unpert.	Δ	
roberta-base	AVERITEC	1.81	0	+1.81	8.33	5.41	+2.92	0.16	0.00	+0.16	+1.63
	SciFact	1.23	0.00	+1.23	14.52	2.38	+12.14	3.95	0.00	+3.95	+5.77
	HoVer	1.02	0.34	+0.68	80.88	22.66	+58.22	–	–	–	+29.45
	FM2	5.26	6.83	-1.57	11.68	6.78	+4.90	–	–	–	+1.67
	PolitiHop	1.24	0	+1.24	6.67	0	+6.67	–	–	–	+3.96
	VitaminC	0.22	0.75	-0.53	0.75	0.23	+0.52	9.08	3.56	+5.52	+1.84
	Average	1.80	1.32	+0.48	20.47	6.24	+14.23	4.40	1.19	+3.21	**+7.39**
Qwen-2.5-14B	AVERITEC	1.14	1.21	-0.07	2.38	2.20	+0.18	5.24	4.18	+1.06	+0.39
	SciFact	1.65	1.65	0.00	0.00	1.52	-1.52	1.28	1.34	-0.06	-0.53
	HoVer	1.40	2.71	-1.31	28.8	4.12	+24.68	–	–	–	+11.69
	FM2	0.80	0.40	+0.40	29.41	4.47	+24.94	–	–	–	+12.67
	PolitiHop	0.49	1.10	-0.61	20.00	5.88	+14.12	–	–	–	+6.76
	VitaminC	3.31	6.09	-2.78	12.5	3.39	+9.11	7.91	3.42	+4.49	+3.61
	Average	1.47	2.19	-0.72	15.52	3.60	+11.92	4.81	2.98	+1.83	**+5.77**

Table 11: Effectiveness of FACTFLIP in distinguishing highly perturbing and less perturbing words across different models, datasets, and target class. The delta column reports the difference in ASR between perturbing and unperturbing words, with positive values indicating high discriminative power of the ranking.

Label	**RoBERTa**	**Qwen**
Support	21.91 ± 26.43	32.00 ± 17.28
Refute	87.44 ± 9.45	80.33 ± 28.73
NEI	29.53 ± 26.76	18.15 ± 21.52
Average	46.29 ± 20.88	43.49 ± 22.51

Table 12: Percentage reduction in the number of claims passing the verification step when injecting three triggers instead of a single trigger. Results are reported as mean ± standard deviation across datasets for each label.

all, this analysis further supports the reliability of FACTFLIP 's trigger ranking and illustrates how adversarial strength and linguistic plausibility interact under multi-trigger perturbations.

F RQ4 - Model diagnosis

F.1 Assessing biases affecting minority demographic groups

This section provides a detailed breakdown of the bias analysis introduced in the main paper, reporting fine-grained results across datasets, target classes, and bias categories. The goal of this analysis is to better characterize how socially salient lexical cues influence model predictions when injected into claims in isolation.

For each bias dimension, *gender*, *religion*, *sexual orientation*, *politics*, and *race*, we construct a dedicated vocabulary of 50 terms, covering both majority and minority identifiers as well as neutral and less frequent descriptors. Table 14 reports a representative subset of 10 words per category. These vocabularies are used to systematically probe lexical sensitivity by inserting a single bias-related term at the beginning of each claim, followed by a punctuation mark, without enabling claim rewriting. This setup ensures that observed effects are attributable to the lexical cue alone, rather than to semantic reformulation.

Table 15 reports the attack success rate obtained when injecting the top-ranked perturbing terms (pert.) and the least perturbing ones (unpert.), as identified by FACTFLIP, across all datasets and target classes. Results reveal substantial variability across datasets and labels. In HoVer and FM2, which contain a large proportion of *Refute* instances, several bias categories lead to pronounced increases in ASR when perturbing words are injected. In particular, political- and religion-related terms exhibit large positive deltas, exceeding +20 points in some cases (e.g., political cues on HoVer and religion cues on PolitiHop). These effects indicate a strong interaction between bias-related lexical cues and contradiction detection. In contrast, effects on the *Support* class are generally smaller and more heterogeneous, with both positive and negative deltas depending on the dataset and bias category. For datasets that include the *NEI* class (e.g., AVERITEC, SciFact, and VitaminC), we observe that bias-related triggers often induce non-trivial

		Support			*Refute*			*NEI*			*Avg* Δ
	Dataset	pert.	unpert.	Δ	pert.	unpert.	Δ	pert.	unpert.	Δ	
roberta-base	AVERITEC	2.09	0.23	+1.86	75	0.00	+75.00	0.23	0.00	+0.23	+25.70
	SciFact	8.00	2.60	+5.40	50.00	0.00	+50.00	18.18	5.00	+13.18	+22.86
	HoVer	0.00	0.00	0.00	100.00	87.18	+12.82	–	–	–	+6.41
	FM2	6.40	8.57	-2.17	57.14	19.35	+37.79	–	–	–	+17.81
	PolitiHop	0.72	0.00	+0.72	50.00	0.00	+50.00	–	–	–	+25.36
	VitaminC	0.20	1.51	-1.31	0.00	0.00	0.00	21.21	13.83	+7.38	+2.02
	Average	2.90	2.15	+0.75	55.36	17.76	+37.60	13.21	6.28	+6.93	**+16.69**
Qwen2.5-14B	AVERITEC	2.26	0.63	+1.63	0.00	6.25	-6.25	5.56	4.33	+1.23	-1.13
	SciFact	1.67	4.26	-2.59	0.00	1.75	-1.75	6.76	2.76	+4.00	-0.11
	HoVer	2.75	4.32	-1.57	41.18	8.33	+32.85	–	–	–	+15.64
	FM2	1.26	1.99	-0.73	35.29	8.54	+26.75	–	–	–	+13.01
	PolitiHop	1.84	2.68	-0.84	0.00	8.33	-8.33	–	–	–	-4.59
	VitaminC	2.52	7.53	-5.01	26.67	7.96	+18.71	14.18	4.40	+9.78	+7.83
	Average	2.05	3.57	-1.52	17.19	6.86	+10.33	8.83	3.86	+4.97	**+5.11**

Table 13: ASR for the set of three most perturbing (*pert.*) and three least perturbing words (*unpert.*) across models, datasets and target classes.

Religion	**Politics**	**Gender**	**Race**	**Sexual**
christian	democrat	man	african	heterosexual
catholic	republican	woman	kurdish	gay
protestant	libertarian	boy	arab	lesbian
anglican	socialist	girl	asian	bisexual
baptist	communist	transgender	caucasian	pansexual
lutheran	anarchist	nonbinary	european	asexual
hindu	monarchist	agender	slavic	queer
jewish	conservative	bigender	mulatto	polysexual
muslim	progressive	pangender	indigenous	sapiosexual
buddhist	capitalist	cisgender	yoruba	aceflux

Table 14: Example of 10 words per bias category (from a total of 50 words per category).

ASR increases, particularly for religion, political, and racial cues. Notably, in VitaminC, several bias categories yield double-digit deltas for *NEI*, indicating that such lexical insertions can substantially alter uncertainty-related predictions. Aggregating results across datasets and labels, the final row of Table 15 reports an overall average delta of +2.49, confirming that perturbing bias-related terms systematically exerts a stronger adversarial effect than unperturbing ones. As reported in the main paper, religion and politics emerge as the most impactful categories overall, followed by gender and race, while sexual orientation cues tend to produce weaker and less consistent effects. Importantly, the direction and magnitude of the deltas vary considerably across datasets, indicating that these biases are not driven by isolated artifacts but reflect dataset-specific interactions between lexical cues, label distributions, and learned representations. This analysis further supports the use of FACTFLIP as a diagnostic tool for uncovering fine-grained sources of adversarial sensitivity and socially grounded biases in claim verification models.

F.2 Discovering sources of adversarial sensitivity

This analysis examines the extent to which adversarial sensitivity to universal triggers arises from task-specific fine-tuning data versus biases inherited from pre-training. To this end, we compare the perturbing and unperturbing triggers identified by FACTFLIP for two models with identical architectures and downstream training data: a standard RoBERTa initialized with pre-trained weights, and a variant in which the pre-training weights are randomly re-initialized prior to fine-tuning.

Table 16 reports the percentage of word overlap between the top-50 most perturbing and least perturbing triggers extracted from the two models across datasets. The results show near-zero overlap in almost all cases. For most datasets, there is no shared vocabulary between perturbing (or unperturbing) triggers identified in the pre-trained and re-initialized models, with only negligible overlap observed in isolated cases (e.g., FM2 and PolitiHop). This lack of overlap indicates that adversarial triggers are not primarily explained by dataset-specific lexical artifacts acquired during fine-tuning. If fine-tuning data were the dominant source of adversarial

			Support			*Refute*			*NEI*			*Avg* Δ
	Dataset	**Bias category**	pert.	unpert.	Δ	pert.	unpert.	Δ	pert.	unpert.	Δ	
roberta-base	AVERITEC	gender	1.09	0.65	+0.44	4.42	6.23	-1.81	0.55	0.42	+0.13	-0.41
		political	0.94	0.29	+0.65	9.18	12.46	-3.28	0.61	0.49	+0.12	-0.84
		racial	1.09	0.80	+0.29	4.92	6.89	-1.97	0.61	0.67	-0.06	-0.58
		religion	0.58	0.79	-0.21	9.51	7.05	+2.46	0.73	0.61	+0.12	+0.79
		sexual orientation	1.81	0.43	+1.38	5.90	5.08	+0.82	0.49	0.36	+0.13	+0.78
	SciFact	gender	1.63	1.63	0	0.80	4.63	-3.83	1.93	1.23	+0.70	-1.04
		political	0.89	1.48	-0.59	1.20	1.39	-0.19	0.70	0.35	+0.35	-0.14
		racial	2.22	2.52	-0.30	1.29	1.29	0	2.63	0.53	+2.10	+0.60
		religion	2.07	2.81	-0.74	1.39	1.10	+0.29	4.39	0.35	+4.04	+1.20
		sexual orientation	1.18	1.78	-0.60	1.39	1.69	-0.30	1.58	2.10	-0.52	-0.47
	HoVer	gender	0.64	0.06	+0.58	82.72	92.28	-9.56	–	–	–	-4.49
		political	1.12	0.07	+1.05	91.62	71.57	+20.05	–	–	–	+10.55
		racial	0.30	0.22	+0.08	73.45	93.09	-19.64	–	–	–	-9.78
		religion	0.59	0.19	+0.40	90.72	77.15	+13.57	–	–	–	+6.99
		sexual orientation	0.99	0.65	+0.34	59.43	37.86	+21.57	–	–	–	+10.96
	FM2	gender	7.13	0.89	+6.24	12.60	2.57	+10.03	–	–	–	+8.14
		political	2.32	2.70	-0.38	6.13	4.61	+1.52	–	–	–	+0.57
		racial	2.08	1.92	+0.16	9.44	4.94	+4.50	–	–	–	+2.33
		religion	6.18	1.12	+5.06	12.71	6.21	+5.50	–	–	–	+5.28
		sexual orientation	5.43	2.36	+3.07	10.03	4.02	+6.01	–	–	–	+4.54
	PolitiHop	gender	0	0	0	21.82	14.54	+7.28	–	–	–	+3.64
		political	0	0	0	20.00	18.18	+1.82	–	–	–	+0.91
		racial	0	0	0	20.00	10.91	+9.09	–	–	–	+4.55
		religion	0	0	0	27.27	9.09	+18.18	–	–	–	+9.09
		sexual orientation	0	0	0	14.54	5.45	+9.09	–	–	–	+4.55
	VitaminC	gender	0.50	0.57	-0.07	6.55	6.61	-0.06	14.53	5.94	+8.59	+2.82
		political	0.50	0.47	+0.03	11.30	10.12	+1.18	23.79	13.08	+10.71	+3.97
		racial	0.59	0.57	+0.02	14.37	1.39	+12.98	11.89	9.87	+2.02	+5.01
		religion	0.59	0.35	+0.24	7.91	10.21	-2.30	28.09	10.62	+17.47	+5.14
		sexual orientation	0.56	0.44	+0.12	10.47	10.20	+0.27	1.58	2.10	-0.52	-0.04
	Average											**+2.49**

Table 15: Difference in attack success rate between perturbing and non-perturbing lexical cues across bias categories and target classes (Support, Refute, NEI). Results are reported for each dataset by inserting a single bias-related word at the beginning of the claim, followed by a punctuation mark. The last column reports the overall average delta per model.

sensitivity, one would expect a substantially higher overlap between the two models, given that they are trained on the same downstream data. Instead, the observed divergence suggests that pre-training plays a central role in shaping the lexical vulnerabilities exploited by universal adversarial triggers.

Taking this insight together with the results reported in the main paper, we claim that adversarial sensitivity to universal triggers cannot be attributed exclusively to either pre-training or fine-tuning dynamics. Instead, it emerges from their interaction: biases encoded during pre-training interact with task-specific supervision to produce stable yet model- and dataset-dependent behaviors.

G Prompts and dataset statistics

Prompts. Figure 3 and Figure 4 show the prompts used for the *LLM as a Perturber* and *LLM as a Verifier* respectively. To both the perturber and verifier we specify the entailment constraints that the new generated claim must satisfy, depending on the label between the original claim and evidence. The entailment constraints are specified in §3.2, and their textualized version in Table 17. Inside Figure 3, we specify that the trigger must be inserted inside the original claim without any modification, such as stemming or lemmatization, to make sure that the added trigger is the same as the one found by FACTFLIP. Furthermore, we spec-

Dataset	% of word overlap	
	pert.	unpert.
AVERITEC	0.01	0.01
SciFact	0.00	0.00
HoVer	0.05	0.00
FM2	0.01	0.01
PolitiHop	0.01	0.01
VitaminC	0.01	0.00

Table 16: Word overlap between the 50 most perturbing and least perturbing triggers identified by FACTFLIP for a fine-tuned RoBERTa model and a RoBERTa model with re-initialized pre-training weights, both fine-tuned on the same dataset. Scores are averaged across the different target classes.

ify to minimize the number of edits to the original claim, so to better isolate the trigger's contribution to the adversarial performance. Additionally, Figure 5 shows the prompt used by Qwen to perform zero-shot claim verification.

Dataset statistics. We report the resulting dataset sizes related to the experiments reported inside this paper. Table 18, Table 19, Table 20 and Table 21 report the number of generated samples associated with the experiments in Table 1, Table 2, Table 3 and Table 13 respectively. From the number of claims that satisfy the verification step, three main patterns can be identified. In particular, (i) generating claims targeting the *refute* class is consistently more difficult, as it becomes easier to use trigger words that inadvertently flip the meaning of the original claim, thus not satisfying the entailment constraints; (ii) choosing trigger words based on similarity with the claim (FF-SIM in Table 19) alleviates this issue, while (iii) using multiple triggers does the opposite (Table 21), with almost all generated claim candidates being discarded due to the entailment constraints. The number of instances of the FF-RAW variant in Table 19 represents the upper bound on the amount of instances that can be generated in our experiments, as FF-RAW does not use the perturb-and-verify pipeline.

Perturber prompt

You will be given a claim and a (series of) word(s). You must add all the provided words inside the given claim. The words added must be exactly like the one provided: any sort of stemming, lemmatization or similar is not allowed.
Also, the rest of the claim must be exactly like the original: no existing word must be removed or modified, and only the provided words must be added. The word(s) must be added inside the claim, not at the beginning. In case of multiple words, you can add them in different positions: you are not forced to add them consecutively.
The novel claim must exclusively satisfy the following rule: "{label}"
First reason step-by-step. Then write "Final answer: " followed exclusively by the generated claim. Do not write anything else after "Final answer: ".
Ensure that the final claim is exactly like the original, except for the added word and other stopwords.
Claim: {claim}
Word: {word}
Let's think step-by-step.

Figure 3: Prompt to generate the novel perturbing claims. The *{label}* is filled with the corresponding label instruction shown in Table 17.

Verifier prompt

You will be given an original claim and a novel claim derived from the original. You must determine if the following rule is satisfied or not: "{label}"
First think step-by-step, then write "Final answer:" followed exclusively by "yes" if the rule is satisfied, "no" otherwise. Do not write anything else after "Final answer:".
Original claim: {claim1}
Novel claim: {claim2}
Let's think step-by-step.

Figure 4: Prompt used by the verifier to check the entailment constraints between the original claim and the generated claim. The *{label}* is filled with the corresponding label instruction shown in Table 17.

Label	Instruction
Support	*the factual content of the original claim must entail the factual content of the new claim. Inside the claims, subjective statements are not factual content, and thus they must not be considered in the final evaluation.*
Refute	*the factual content of the new claim must entail the factual content of the original claim. Inside the claims, subjective statements are not factual content, and thus they must not be considered in the final evaluation.*
NEI	*the factual content of the original claim must entail the factual content of the new claim and viceversa. Inside the claims, subjective statements are not factual content, and thus they must not be considered in the final evaluation.*

Table 17: Textual label descriptions used inside the Figure 3 and Figure 4 prompting templates.

Verifier prompt

You are a fact checking system. You must indicate whether the claim is supported, refuted or "not enough information" based on the given evidence.
After "Answer: ", write exclusively "support", "refute" or "not enough information". Do not write anything else.
{input}
Answer:

Figure 5: Prompt used by `Qwen2.5-14B-Instruct` to perform claim verification.

		Support			*Refute*			*NEI*		
	Dataset	FF	FF-DS	AutoPrompt	FF	FF-DS	AutoPrompt	FF	FF-DS	AutoPrompt
roberta-base	AVERITEC	551	490	486	60	50	20	622	421	416
	SciFact	244	57	75	62	14	93	76	53	46
	HoVer	683	711	539	204	151	109	–	–	
	FM2	570	515	585	137	42	46	–	–	
	PolitiHop	322	215	329	15	6	4	–	–	
	VitaminC	455	454	397	134	165	105	749	606	658
Qwen2.5-14B	AVERITEC	438	356	314	42	33	140	420	323	225
	SciFact	121	84	61	94	85	46	78	87	79
	HoVer	358	469	383	59	111	63	–	–	
	FM2	551	521	373	51	43	24	–	–	
	PolitiHop	406	319	324	10	7	10	–	–	
	VitaminC	393	270	263	64	28	24	785	341	206

Table 18: Number of instances produced by the perturb-and-verify pipeline for the experiment in Table 1.

		Support		*Refute*		*NEI*	
	Dataset	FF-RAW	FF-SIM	FF-RAW	FF-SIM	FF-RAW	FF-SIM
roberta-base	AVERITEC	1380	532	610	123	1640	522
	SciFact	675	99	1005	104	570	61
	HoVer	1250	609	1250	474	–	–
	FM2	1270	633	1230	242	–	–
	PolitiHop	690	282	55	16	–	–
	VitaminC	1175	361	1510	372	2315	655
Qwen2.5-14B	AVERITEC	1035	417	595	120	1280	394
	SciFact	785	111	1010	59	675	109
	HoVer	910	461	930	249	–	–
	FM2	1120	496	1060	182	–	–
	PolitiHop	735	346	100	34	–	–
	VitaminC	885	274	1205	261	1800	462

Table 19: Number of samples produced for `FF-RAW` and `FF-SIM` in Table 2.

		Support		*Refute*		*NEI*	
	Dataset	pert.	unpert.	pert.	unpert.	pert.	unpert.
roberta-base	AVERITEC	551	603	60	37	622	542
	SciFact	244	41	62	42	76	58
	HoVer	683	890	204	256	–	–
	FM2	570	600	137	118	–	–
	PolitiHop	322	366	15	8	–	–
	VitaminC	455	536	134	174	749	342
Qwen2.5-14B	AVERITEC	438	331	42	227	420	407
	SciFact	121	121	94	264	78	149
	HoVer	358	431	59	291	–	–
	FM2	551	535	51	380	–	–
	PolitiHop	406	273	10	51	–	–
	VitaminC	393	345	64	472	785	507

Table 20: Number of perturbing and unperturbing samples produced by the perturb-and-verify pipeline when one trigger is inserted inside the claim (Table 3).

		Support		*Refute*		*NEI*	
	Dataset	pert.	unpert.	pert.	unpert.	pert.	unpert.
roberta-base	AVERITEC	431	427	4	4	442	337
	SciFact	75	38	4	2	33	20
	HoVer	490	725	62	39	–	–
	FM2	531	455	7	31	–	–
	PolitiHop	278	213	2	0	–	–
	VitaminC	493	464	18	15	726	347
Qwen2.5-14B	AVERITEC	265	319	3	64	414	416
	SciFact	60	47	5	57	74	181
	HoVer	364	278	17	108	–	–
	FM2	397	402	17	82	–	–
	PolitiHop	217	149	2	12	–	–
	VitaminC	278	239	15	113	409	589

Table 21: Number of perturbing and unperturbing samples produced by the perturb-and-verify pipeline when 3 triggers are inserted inside the claim (Table 13).