Preprint:
Leite, J.A., Razuvayevskaya, O., Bontcheva, K. et al. (Submitted: 2026) LLM-based adversarial persuasion attacks on fact-checking systems. [Preprint - arXiv] (Submitted)

https://doi.org/10.48550/arXiv.2601.16890

LLM-Based Adversarial Persuasion Attacks on Fact-Checking Systems

João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva and Carolina Scarton
Department of Computer Science, University of Sheffield, UK
{j.leite,o.razuvayevskaya,k.bontcheva,c.scarton}@sheffield.ac.uk

Abstract

Automated fact-checking (AFC) systems are susceptible to adversarial attacks, enabling false claims to evade detection. Existing adversarial frameworks typically rely on injecting noise or altering semantics, yet no existing framework exploits the adversarial potential of persuasion techniques, which are widely used in disinformation campaigns to manipulate audiences. In this paper, we introduce a novel class of persuasive adversarial attacks on AFCs by employing a generative LLM to rephrase claims using persuasion techniques. Considering 15 techniques grouped into 6 categories, we study the effects of persuasion on both claim verification and evidence retrieval using a decoupled evaluation strategy. Experiments on the FEVER and FEVEROUS benchmarks show that persuasion attacks can substantially degrade both verification performance and evidence retrieval. Our analysis identifies persuasion techniques as a potent class of adversarial attacks, highlighting the need for more robust AFC systems.

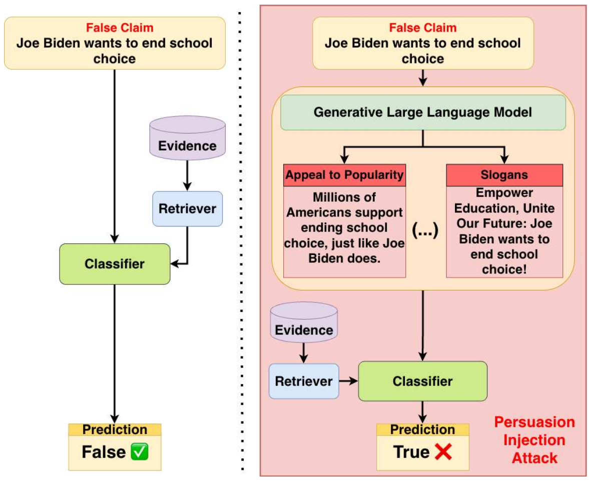

Figure 1: **Overview of a persuasion injection attack.** An automated fact-checking (AFC) system correctly classifies the claim as *False* (left). Using our novel attack method (right), a generative LLM injects persuasion techniques into the claim; the same AFC system now incorrectly predicts *True*.

1 Introduction

Automated Fact-Checking (AFC) systems are playing an increasingly important role in countering disinformation, as they promptly identify check-worthy claims, retrieve relevant evidence, and verify the accuracy of those claims using the retrieved evidence (Nakov et al., 2021). AFC systems are, however, not immune to adversarial attacks affecting their reliability and usefulness in real-world settings (Liu et al., 2025).

Early work on adversarial attacks against AFC systems has primarily conceptualised robustness as invariance to surface-level adversarial perturbations such as typos, synonym substitutions, or character noise (Thorne et al., 2019; Mamta and Cocarascu, 2025), whilst recent approaches introduce semantic perturbations generated by Large Language Models (LLMs) (Atanasova et al., 2020; Huang et al., 2023; Chen et al., 2025a).

While valuable, these efforts leave an unaddressed gap: disinformation in-the-wild often employs persuasion techniques such as manipulative wording, simplifications, and distractions to influence audiences (Benkler et al., 2018; Guess and Lyons, 2020). With generative LLMs capable of generating highly persuasive text at scale (Breum et al., 2024; Hackenburg et al., 2025), malicious actors can now systematically weaponise these techniques to craft disinformation that is fluent and rhetorically optimised to evade detection. Yet, despite this imminent threat to the reliability of AFCs, no existing work has investigated their susceptibility to this form of adversarial manipulation.

To address this gap, **we introduce persuasion injection attacks (Figure 1), the first adversarial framework that systematically weaponises persuasive rhetoric against AFC systems**. Our method uses a generative LLM to rewrite claims using a taxonomy of 15 persuasion techniques across 6 groups drawn from Piskorski et al. (2023).

This novel adversarial framework is used to attack a typical AFC pipeline, consisting of *an evidence retriever* that identifies relevant information, and *a classifier* that predicts binary claim veracity based on supporting evidence. Since our adversarial framework injects persuasion into the input claim, it acts as an upstream perturbation that targets both stages simultaneously. To disentangle these effects, we adopt a decoupled evaluation strategy. First, we experiment with the **veracity classifier** under two conditions: without any supporting evidence (relying on surface form) and with gold-evidence (simulating a perfect retriever). Next, we evaluate the **evidence retriever** to determine if persuasion disrupts the semantic alignment necessary for retrieving the correct evidence.

Our findings reveal that **persuasion injection attacks pose a major threat to AFC pipelines, reducing accuracy more than twice as much as the previously studied adversarial attacks**, such as synonym substitutions and character perturbations. Under an optimised attacker that always selects the most damaging technique for a given claim, **accuracy collapses to near-zero, even when the classifier is provided with gold-evidence**. In particular, techniques in the *Manipulative Wording* category emerge as the most damaging, since they remove concrete information and introduce ambiguity, thus simultaneously degrading evidence retrieval and classification performance.

The key contributions of this work are:

- **Persuasion injection attacks**, a class of adversarial attacks that exploit persuasion techniques to induce failures in AFC systems.
- The **evaluation of AFC systems' robustness to persuasion-based attacks** under controlled settings that isolate evidence retrieval from veracity classification.
- The identification of **persuasion techniques that simultaneously degrade evidence retrieval and veracity classification**, causing a complete AFC pipeline failure.

We make the code openly available to support future work on strengthening AFC systems.[1]

[1]Code available at `https://www.github.com/TBA`.

2 Related Work

2.1 Adversarial Attacks on AFC Systems

AFC systems are vulnerable to a wide range of adversarial attacks that target claims, evidence, or their pairing (Liu et al., 2025). In this section, we focus specifically on attacks that target claims.

Early work revealed that AFC models are brittle to **surface-level transformations**. Rule-based edits such as synonym substitution, negation, date manipulation, or subset-based reasoning degrade both evidence retrieval and verdict prediction (Thorne et al., 2019; Hidey et al., 2020). Character-level perturbations (typos, homoglyphs, invisible control characters, and "leet" substitutions) corrupt tokenisation and reduce prediction accuracy while preserving visual similarity (Boucher et al., 2022). Short appended n-grams (Atanasova et al., 2020) and synonym substitution (Alzantot et al., 2018) can flip entailment and veracity labels while preserving meaning. The FactEval benchmark (Mamta and Cocarascu, 2025) further demonstrates that modern LLM-based AFC systems remain vulnerable to these perturbations.

Beyond surface noise, semantic claim manipulations pose deeper challenges. Style-transfer attacks (e.g., transforming formal claims into colloquial phrasing) substantially degrade evidence retrieval (Kim et al., 2021). Recent progress enabled a new class of adversarial attacks built around **LLM-driven claim generation**. Early approaches used smaller LMs such as GPT-2 to produce lexically informed paraphrases or natural-language triggers that subtly alter semantics while maintaining fluency (Atanasova et al., 2020). More recent systems, including Grover (Zellers et al., 2019) and topic-guided generators (Huang et al., 2023), can generate fluent and contextually plausible false claims that reliably evade AFC systems. LLMs have also been adopted as optimisation engines for crafting adversarial claims: methods such as GEM (Niewinski et al., 2019) generate inputs explicitly tuned to mislead AFC models, achieving higher attack success than rule-based or lexical perturbations. Similarly, Chen et al. (2025a) used GPT-4o to iteratively induce factual errors (e.g., swapping entities or dates) guided by feedback from an AFC system, achieving significant performance degradation.

2.2 Persuasion in NLP

Persuasion techniques (also referred to as propaganda techniques) encompass rhetorical strategies intended to influence a reader's judgment (Da San Martino et al., 2019). The SemEval propaganda detection tasks (Da San Martino et al., 2020; Dimitrov et al., 2021) established the earliest datasets, leading to recent iterations covering 23 persuasion techniques across 6 categories in multiple languages (Piskorski et al., 2023). Further datasets explored persuasion in multimodal data (Wu et al., 2022), code-switched text (Salman et al., 2023), and rationale-annotated data (Hasanain et al., 2025). Recent studies demonstrate that generative LLMs are highly proficient in producing persuasive language (Breum et al., 2024; Pauli et al., 2025) and can change individuals' beliefs on various topics (Hackenburg et al., 2025). Together, these efforts have established persuasion detection as a mature NLP task with well-defined taxonomies and benchmark datasets (Martino et al., 2020; Srba et al., 2025).

A growing body of work studies how persuasion is weaponised in disinformation. Huang et al. (2023) used persuasion techniques to perform data augmentation, reporting improvements in detecting disinformation. Moreover, Nikolaidis et al. (2024) demonstrated that simple distributions of persuasion techniques can act as features for classifiers trained to reliably distinguish propaganda, conspiracy theories, and hyperpartisan content. Leite et al. (2025) found that persuasion techniques are heavily employed in different disinformation topics, and while techniques such as *Doubt* and *Loaded Language* are ubiquitous, others show strong domain specificity (e.g., *Appeal to Authority* in climate change disinformation, or *Flag Waving* in the Russia-Ukraine War). Together, these findings demonstrate that persuasion techniques can play a key role in making disinformation more believable.

To the best of our knowledge however **no prior work has studied the adversarial potential of persuasion techniques. This is the key novel contribution of this paper, as it weaponises persuasion techniques to attack AFC systems**.

3 Methodology

Persuasion Techniques. We initially consider a taxonomy of 23 persuasion techniques grouped into **6 high-level categories** derived from definitions in computational social science from Piskorski et al. (2023): *Attack on Reputation*, *Justification*, *Distraction*, *Simplification*, *Call*, and *Manipulative Wording*. The full list of persuasion techniques and their definitions is provided in Appendix A.

Variant Generation. For each claim x_i, we generate a set of adversarial variants $\mathcal{V}_i=\{x'_{(i,1)},\ldots,x'_{(i,23)}\}$ using the 23 persuasion techniques. We use QWEN2.5-7B-INSTRUCT as the generative model to produce the adversarial variants due to its strong instruction-following performance and modest parameter scale (7 billion), while being fully open-weight (Qwen et al., 2025). To verify that our results are not specific to this generator, we additionally compare results against LLAMA-3-8B-INSTRUCT in Appendix C.

We construct prompts containing: (1) the definition of a specific persuasion technique $t \in \mathcal{T}$, (2) the original claim x_i, (3) constraints to ensure the output is well-formatted, and the claim label (*True*/*False*) is preserved, and (4) two few-shot examples. All prompt templates used in the experiments are available in Appendix A.

Label-Invariance Validation. Persuasion techniques typically operate by altering emphasis, framing, or discourse focus rather than by modifying the underlying factual content (e.g., *Red Herring*: "Divert attention from the claim by introducing an unrelated topic"). Accordingly, we do not enforce strict semantic equivalence between the original and persuasive claims. Automated semantic similarity metrics such as BERTScore (Zhang et al., 2019) are ill-suited as filters in this setting as they frequently assign low similarity scores to valid attacks that preserve the original veracity label.[2] Instead, we validate *label-invariance*, requiring that each persuasive variant retains the same veracity label (*True* or *False*) as its source claim.

We conducted a manual validation study in which a sample of 690 persuasive claims (30 per technique and stratified by dataset FEVER/FEVEROUS) was annotated for factual correctness given gold-evidence. The original claim text was not shown to prevent anchoring bias and ensure independent verification. A persuasive claim was considered label-invariant if the assigned verdict matched the original ground-truth

[2]For example, the *Slogan* variant "Step aside for a new era: Horsford hands the baton in 2009!" (see Table 9) receives a low BERTScore (0.45) despite preserving the same *False* label as the original claim.

label. Based on this validation, we excluded 8 persuasion techniques with label preservation rates $\leq 80\%$ from all experiments. Among the 15 retained techniques, **less than 1.5% of adversarial claims flipped the original label** (*True*↔*False*). Full details of the validation protocol are reported in Appendix B.

Attacker Capabilities. We assume a binary classification task of claim verification where the target label $y_i \in \{\text{TRUE}, \text{FALSE}\}$. Let $f(x, E)$ denote the classifier's predicted probability that the input claim x with supporting evidence E belongs to the class TRUE. We evaluate two distinct adversary capabilities:

- **The Blind Attacker (Average Case):** Simulates a naive adversary who applies a persuasion technique at random without knowledge of the target system's sensitivities. We compute metrics on the full pooled set of variants, effectively calculating the expected performance degradation: $\mathbb{E}_{t\sim\mathcal{T}}[\mathcal{M}(f(x_{i,t}, E), y_i)]$, where $\mathcal{T}$ is the uniform distribution over the persuasion techniques, and $\mathcal{M}$ is a performance metric.
- **The Oracle Attacker (Worst-Case)** simulates an adversary with query access to the classifier, selecting the variant $x^* \in \mathcal{V}_i$ that maximises the probability of the incorrect class. Formally, $x_i^* = \text{argmax}_{x' \in \mathcal{V}_i} \mathcal{L}(x', E)$, where $\mathcal{L} = (1 - f(x', E))$ if $y_i = \text{TRUE}$ and $f(x', E)$ if $y_i = \text{FALSE}$.

4 Experimental Setup

4.1 Datasets

We use two widely adopted fact-checking benchmarks in English: **FEVER** (Thorne et al., 2018) and **FEVEROUS** (Aly et al., 2021). FEVER represents the task of verifying short claims against unstructured textual evidence (sentences from Wikipedia), whereas FEVEROUS consists of longer claims, with structured evidence (tables and lists, also from Wikipedia), and a higher number of pieces of evidence per claim (Table 1).

Given that our focus is on assessing AFC systems' vulnerability to persuasion on verifiable facts, for both datasets, we use the subset of claims where the ground-truth label is either TRUE or FALSE, excluding NOTENOUGHINFO. For FEVER, we use the official data splits. For FEVEROUS, since gold-evidence is not publicly available for the official test split, we repurpose the development set as our held-out **Test** set. A validation split (**Dev** set) is also created for model checkpointing and hyperparameter tuning by randomly sampling a stratified subset of the official training set. The remaining samples are left as our **Train** set.

Train and **Dev** splits are used exclusively to train and validate the AFC classifiers, respectively. All adversarial attacks are applied solely to the held-out **Test** set. This ensures that no adversarial examples are observed during training and validation. Table 1 summarises dataset statistics.

Dataset / Split	*True* Claims	*False* Claims	Tokens ($\text{avg}_{\text{claim}}$)	$\text{Evidence}_{\text{gold}}$ ($\text{avg}_{\text{claim}}$)
FEVER train	80,035 (72.9%)	29,775 (27.1%)	8.1	1.8
FEVER dev	3,333 (50.0%)	3,333 (50.0%)	8.1	1.7
FEVER test	3,333 (50.0%)	3,333 (50.0%)	8.4	1.6
FEVEROUS train	35,567 (60.6%)	23,116 (39.4%)	25.2	4.5
FEVEROUS dev	6,268 (60.5%)	4,099 (39.5%)	25.3	4.6
FEVEROUS test	3,908 (52.9%)	3,481 (47.1%)	24.9	4.0

Table 1: **Dataset statistics.**

4.2 AFC Classifiers

Following prior work (Kim et al., 2021; Aly et al., 2021; Abdelnabi and Fritz, 2023), we use ROBERTA-BASE (Liu et al., 2019) as our veracity classifier. While generative LLMs are increasingly applied to verification tasks (Augenstein et al., 2024), they frequently exhibit hallucinations (Huang et al., 2025), sensitivity to prompt phrasing (Errica et al., 2025), and risk data contamination from popular benchmark datasets such as FEVER and FEVEROUS (Chen et al., 2025b). These factors introduce substantial confounders to our analysis; therefore, we adopt a strong, standard encoder-based baseline to ensure controlled evaluation and comparability with prior work.

To compare the adversarial robustness of surface-level versus evidence-grounded classification on adversarial vulnerability, we consider two AFC settings:

Claim-Only Model (No Evidence). This model relies exclusively on the linguistic surface forms of the claim. Given a claim c, the model predicts a probability distribution over class labels y: $P(y|c) = \text{softmax}(W \cdot \text{Encoder}(c))$, where W is a learnable classification head. While real-world AFC systems incorporate evidence, this serves as a baseline to measure susceptibility to stylistic forms without evidence grounding.

Gold-Evidence Model. We concatenate the claim c and the set of gold-evidence snippets $E_{gold}=\{e_1, ..., e_k\}$ into a single sequence, using the special separator token [SEP] to delimit distinct pieces of evidence. The input sequence is formatted as: $x = \texttt{[CLS]}\ c\ \texttt{[SEP]}\ e_1\ \texttt{[SEP]} \ldots\ \texttt{[SEP]}\ e_k\ \texttt{[SEP]}$. By providing the model with gold-evidence, we simulate a perfect retriever and therefore decouple the impact of the attack from retrieval vulnerabilities.

Dev Set Performance. Our veracity classifiers achieve strong performance on the **Dev** set. The *Gold-Evidence* models reach **0.953**±0.001 and **0.910**±0.002 Macro F1 on FEVER and FEVEROUS, respectively, significantly outperforming the *Claim-Only* setting (FEVER: **0.826**±0.002; FEVEROUS: **0.661**±0.002). Full training details and hyperparameters are provided in Appendix D.

4.3 Evidence Retrieval Analysis

While our primary experiments condition the veracity classifier on gold-evidence to isolate reasoning vulnerabilities, we additionally evaluate whether persuasion attacks disrupt the evidence collection stage itself (see subsection 5.4). For this analysis, we employ a standard sparse retrieval setup using the **BM25** algorithm implemented via Pyserini (Lin et al., 2021). We construct separate indices for the FEVER (June 2017 dump) and FEVEROUS (December 2020 dump) Wikipedia snapshots. Retrieval is performed at the page level, where each document corresponds to a Wikipedia article identified by its title, and each claim serves as the query.

4.4 Baseline Adversarial Attacks

We compare our novel adversarial attacks against four standard baselines commonly used in prior work on adversarial attacks against AFC systems (Thorne et al., 2019; Jin et al., 2020; Boucher et al., 2022; Liu et al., 2025):

- **Synonym Substitution.** We replace tokens with WordNet synonyms while matching part-of-speech. To achieve this, the text is POS-tagged, and eligible tokens (alphabetic, length >2, available synonyms) are sampled at a fixed rate (10%).
- **Word Swap.** Given a claim with at least two tokens, we compute swaps proportional to claim length. We calculate the number of swaps by taking 10% of the number of words in the claim, rounding down, and ensuring at least one swap for short claims. Each swap selects two distinct positions and exchanges their tokens.
- **Character Perturbations.** To simulate natural typos, we apply small character-level edits to randomly selected tokens (10% of tokens, minimum length $\geq$3 characters). Each chosen word receives at least one edit. For longer words ($>$8 characters), we apply two edits. The types of operations include swaps, deletions, insertions, or substitutions of characters, with perturbations applied uniformly.
- **Paraphrase.** We use the same LLM as in our persuasion injection attack (QWEN2.5-7B-INSTRUCT) to produce paraphrases of claims. Prompts used can be found in Appendix A.

4.5 Metrics

Classification Metrics. We report standard classification metrics: **Accuracy**, **Macro F1**, and **ROC AUC**. Additionally, we report **Attack Success Rate (ASR)**, which quantifies how often an adversarial claim causes a misclassification relative to the model's performance on the corresponding original claim:

$$\text{ASR} = \frac{\sum_{i \in C_{correct}} \mathbb{I}(\hat{y}(x_i) \neq y_i)}{|C_{correct}|} \quad (1)$$

where $C_{correct}$ is the subset of claims correctly predicted in the original setting. $\mathbb{I}$ is the indicator function and x_i is the attack variant. We further distinguish between **Evasion** ASR (False→True) and **Sabotage** ASR (True→False).

Retrieval Metrics. For the retrieval analysis (see subsection 4.3), we measure performance using **Recall@k**, i.e., the proportion of gold-evidence retrieved within the top-k candidates.

5 Results

5.1 Effectiveness of Persuasion Attacks

Table 2 shows that standard adversarial baselines–Synonym Substitution, Character Perturbations, Word Swap, and Paraphrasing–cause modest accuracy degradation. On FEVER, these attacks reduce claim-only accuracy by at most 6 points and evidence-based accuracy by 3 to 10 points, while on FEVEROUS the impact is smaller, with accuracy drops below 3.2 points in both settings.

In contrast, persuasion attacks substantially disrupt veracity prediction. Under the *Blind* setting (random technique per claim), claim-only accuracy drops by 13.2 on FEVER, and 6.6 on FEVEROUS, while gold-evidence accuracy decreases by

	FEVER								FEVEROUS							
	Claim-Only				Gold-Evidence				Claim-Only				Gold-Evidence			
Adversarial Attack	$F1_{Mac}\uparrow$	AUC ↑	Acc ↑	$\Delta_{Acc}\downarrow$	$F1_{Mac}\uparrow$	AUC ↑	Acc ↑	$\Delta_{Acc}\downarrow$	$F1_{Mac}\uparrow$	AUC ↑	Acc ↑	$\Delta_{Acc}\downarrow$	$F1_{Mac}\uparrow$	AUC ↑	Acc ↑	$\Delta_{Acc}\downarrow$
None (Original Claims)	0.814	0.893	0.814	–	0.943	0.979	0.943	–	0.627	0.686	0.641	–	0.894	0.964	0.894	–
Synonym Substitution	0.764	0.848	0.764	−0.050	0.843	0.927	0.845	−0.098	0.618	0.666	0.623	−0.018	0.863	0.948	0.863	−0.031
Character Perturbations	0.762	0.846	0.762	−0.052	0.868	0.938	0.869	−0.074	0.619	0.675	0.632	−0.009	0.877	0.953	0.878	−0.016
Word Swap	0.755	0.847	0.757	−0.057	0.897	0.957	0.897	−0.046	0.608	0.662	0.609	−0.032	0.883	0.955	0.884	−0.010
Paraphrasing	0.755	0.837	0.757	−0.057	0.912	0.953	0.912	−0.031	0.610	0.669	0.629	−0.012	0.881	0.954	0.882	−0.012
Persuasion (Blind)	0.678	0.745	0.682	**−0.132**	0.764	0.862	0.764	**−0.179**	0.567	0.597	0.575	**−0.066**	0.722	0.870	0.726	**−0.168**
Persuasion (Oracle)	0.042	0.004	0.043	**−0.771**	0.142	0.029	0.164	**−0.779**	0.010	0.001	0.010	**−0.631**	0.203	0.129	0.250	**−0.644**

Table 2: **Performance degradation of fact-checking models across different adversarial attacks.** Persuasion-based attacks induce substantially larger accuracy drops than lexical or paraphrastic perturbations.

roughly 17 points on both datasets. Under the *Oracle* setting (worst-case optimisation), performance collapses: claim-only accuracy falls to 0.043 on FEVER and 0.010 on FEVEROUS, implying that **for nearly every claim there exists at least one persuasive variant that flips the model's prediction**. Even in the gold-evidence setting, accuracy is still severely degraded (0.164 on FEVER; 0.250 on FEVEROUS).

5.2 Error Analysis: Evasion vs. Sabotage

Table 3 displays the Attack Success Rate (ASR) for Evasion and Sabotage (see subsection 4.5).

	FEVER		FEVEROUS	
Metric	**Claim**	**+E_{gold}**	**Claim**	**+E_{gold}**
Evasion *(False→True)*				
Blind (Avg)	0.306	**0.070**	0.318	**0.039**
Oracle (Worst)	0.996	**0.759**	0.997	**0.434**
Sabotage *(True→False)*				
Blind (Avg)	0.217	**0.392**	0.299	**0.449**
Oracle (Worst)	0.998	**1.000**	1.000	**0.999**

Table 3: **Attack success rates (ASR) for persuasion attacks.** Higher scores indicate more effective attacks.

Models with access to gold-evidence are substantially more robust against evasion. In the *Blind* setting, *Claim-Only* models are highly exploitable, with an Evasion ASR exceeding 30% on both datasets. In contrast, the *Gold-Evidence* model sharply suppresses these attacks, with evasion rates of only 7% on FEVER and 4% on FEVEROUS. However, under the *Oracle* setting, a sophisticated attacker can still force evasion at substantial rates of 75.9% on FEVER and 43.4% on FEVEROUS, representing over a 10× increase relative to Blind attacks. This pattern shows that **persuasive *False* claims easily flip evidence-free predictions, but rarely do so when gold-evidence is available, *unless* the attacker can identify the most damaging persuasive variant per claim.**

***Gold-Evidence* models are more vulnerable to sabotage**. In the *Blind* attack, Sabotage ASR rises from below 22% in *Claim-Only* models to above 44% for *Gold-Evidence* models, while in the *Oracle* attack, Sabotage ASR stays at nearly 100% in both settings. These results indicate that persuasion injection can induce claim-evidence misalignment, biasing evidence-grounded models mostly toward *False* predictions.

5.3 Technique-Level Vulnerability

We now aim to identify which specific persuasion techniques are most effective at causing evasion (FALSE→TRUE), as it is the fundamental goal of a real-world adversarial attack on AFC systems (i.e., to disrupt the system's ability to detect *False* claims). Table 4 splits the techniques into *Effective*, *Partially Effective*, and *Neutralised* based on whether they achieve Evasion ASR ≥5% in the *Gold-Evidence* setting. To illustrate how *Effective* techniques cause evasion, we present concrete examples in Appendix E.

	FEVER		FEVEROUS	
Persuasion Technique	**Claim**	**+E_{gold}**	**Claim**	**+E_{gold}**
Effective *(+Evd. EASR ≥ 0.05 in both datasets)*				
Obfuscation **(MW)**	0.73	**0.38**	0.65	**0.15**
Repetition **(MW)**	0.33	**0.25**	0.25	**0.10**
Slogan **(C)**	0.74*	**0.07**	0.66*	**0.18**
Flag Waving **(J)**	0.51	**0.10**	0.59	**0.06**
Appeal to Popularity **(J)**	0.30	**0.05**	0.45	**0.06**
Partially Effective *(+Evd. EASR ≥ 0.05 in one dataset only)*				
Appeal to Values **(J)**	0.62	**0.11**	0.59	0.04
Appeal to Authority **(J)**	0.35	**0.07**	0.40	0.04
Neutralised *(+Evd. EASR < 0.05 in both datasets)*				
Loaded Language **(MW)**	0.27	0.04	0.31	0.03
Guilt by Association **(AoR)**	0.31	0.04	0.22	0.02
Name Calling / Labelling **(AoR)**	0.12	0.02	0.11	0.03
Whataboutism **(D)**	0.21	0.02	0.21	0.02
Casting Doubt **(AoR)**	0.18	0.01	0.27	0.03
Red Herring **(D)**	0.22	0.01	0.29	0.01
Appeal to Hypocrisy **(AoR)**	0.18	0.02	0.08	0.01
Conversation Killer **(C)**	0.03	0.01	0.09	0.01

Table 4: **Effective and neutralised persuasion techniques.** Evasion ASR scores for Claim-only and Gold-evidence models. Higher scores indicate more effective attacks. Coarse persuasion categories are shown in parentheses: **(MW)**=Manipulative Wording, **(C)**=Call, **(J)**=Justification, **(S)**=Simplification, **(AoR)**=Attack on Reputation, **(D)**=Distraction. Best score in the Claim-only setting is denoted with *. Scores with +E_{gold} EASR≥0.05 are in **bold**.

Surface-Form Vulnerabilities. In the claim-only setting, performance reflects the verifier's reliance on surface-level linguistic cues. All techniques (except *Conversation Killer* in FEVER) induce non-trivial evasion rates above 5%. Techniques such as *Slogan* and *Obfuscation* are particularly effective, achieving evasion rates above 65%. *Slogan* compresses claims into short, punchy formulations that resemble news headlines ("*Step aside for a new era: Horsford hands the baton in 2009!*"; see Table 9), which can increase perceived credibility. *Obfuscation* reduces specificity (e.g., "*(...) approximately one and a half months*"), making it less apparent that the claim asserts a precise, falsifiable statement, leading to high evasion rates despite no change in the underlying veracity.

Vulnerabilities Under Evidence-Grounding. Several high-impact techniques achieve substantial evasion power even for models grounded in gold-evidence, forming a class of effective threats spanning three coarse categories: ***Manipulative Wording***, ***Calls***, and ***Justification***.

Among these, ***Manipulative Wording* remains the most damaging group** (except *Loaded Language*). *Obfuscation* attacks replace concrete entities with vague formulations: "*50 days*"→"*approximately one and a half months*", and "*Thanksgiving*"→"*holiday season*"; See Table 9), making it harder for evidence to explicitly contradict claims. This technique's high effectiveness aligns with Glockner et al. (2024), which found that incomplete or vague formulations elicit substantial disagreement even among human fact-checkers. Next, *Repetition* can reiterate irrelevant or partially true elements of the claim, diverting the focus away from the component that is actually false: "*GZMB is a serine protease, a serine protease with* {*false predicate*}".

Justification techniques also show high effectiveness in FEVER, with all techniques exceeding the 5% threshold in this dataset; notably, *Flag Waving* and *Appeal to Popularity* maintain effectiveness in both datasets. Both techniques aim to increase the plausibility of *False* information by framing it to be widely accepted: "*True Sri Lankans must support* {*false predicate*}", and "*Everyone is celebrating* {*false predicate*}", for *Flag Waving* and *Appeal to Authority*, respectively.

In contrast, techniques in groups related to ***Attack on Reputation*, *Distraction*, and *Simplification***, which redirect the focus away from the claim (towards the speaker, an external issue, or a simplified framing), **collapse to near-zero evasion rates on models grounded on gold-evidence**.

5.4 Impact on Evidence Retrieval

So far, we have examined persuasion attacks under gold-evidence to isolate vulnerabilities in the veracity classifier. However, AFC pipelines rely on retrieval systems to access supporting evidence. In this section, we shift focus to the evidence retriever, examining whether persuasion attacks degrade retrieval quality and expose failure modes specific to the retrieval stage.

Retrieval Vulnerability. Figure 2 illustrates the retrieval performance (Recall@k) across varying depths (k=$\{3, 5, 7, 10\}$) and under different types of adversarial attacks. Standard lexical attacks[3] cause modest retrieval degradation. The strongest lexical attack (*Character Noise*) drops $Recall@5$ by 0.105 and 0.023 points with respect to the *Baseline*, on FEVER and FEVEROUS, respectively. Persuasion attacks cause deeper failures. Under the *Blind* setting, persuasion injection degrades $Recall@5$ by 0.197 points on FEVER and 0.215 on FEVEROUS, in comparison to the *Baseline*. The *Oracle* setting again shows a catastrophic power, collapsing $Recall@k$ to near-zero for all values of k. These results confirm that **persuasion techniques not only make claims harder to verify, but also decouple the claim from its supporting evidence, causing a complete pipeline failure in AFC systems.**

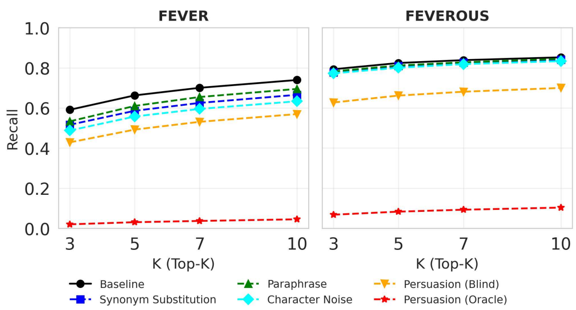

Figure 2: **BM25 retrieval performance**. Lower scores indicate more effective attacks.

Retrieval vs. Classification Vulnerabilities. We analyse the extent to which techniques that successfully evade classification also impair retrieval. Fig-

[3] BM25 is invariant to token permutation. Therefore, results for the *Word Swap* attack are identical to the baseline, and thus we do not include it in this analysis.

ure 3 plots *Retrieval Degradation* ($|\Delta$Recall@5$|$) against *Classification Degradation* under E_{gold} (Evasion ASR) for individual persuasion techniques, with scores averaged across FEVER and FEVEROUS.

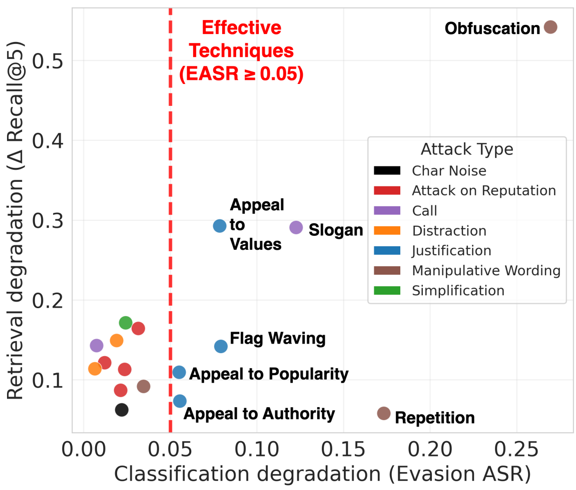

Figure 3: **Retrieval vs. classification vulnerabilities.** Scores are averaged between FEVER and FEVEROUS. Evasion ASR computed under access to gold-evidence. Higher scores indicate more effective attacks.

Not all techniques that are effective in degrading classification also substantially harm retrieval. *Repetition*, for example, achieves the second-highest evasion rate (17.3%) yet reduces Recall@5 by only 0.05–less than *Char Noise*, with 0.06. This occurs because *Repetition* mostly adds tokens that are already present in the claim, which has a negligible impact on the retriever. Similarly, three out of five *Justification* techniques (*Flag Waving*, *Appeal to Popularity*, and *Appeal to Authority*) form a small cluster of techniques with low effectiveness in attacking the retriever, with up to 0.14 $|\Delta Recall@5|$.

Appeal to Values and *Slogan* are effective in degrading both classification and retrieval performance, with Evasion ASR of 7.8% and 12.2%, respectively, and $|\Delta Recall@5|$ of 0.29. More importantly, the single most harmful technique substantially degrades both components: ***Obfuscation* stands out as an outlier attack** (top-right in Figure 3), with an average Evasion ASR of 27% and a $|\Delta Recall@5|$ of 0.54. By replacing concrete entities with vague generalisations (e.g., changing "*50 days*" to "*approximately one and a half months*" or "*Thanksgiving*" to "*holiday season*"; see Table 9), ***Obfuscation* denies the retriever the ability to match concrete keywords needed to find evidence, while simultaneously creating ambiguity that confuses the veracity classifier**.

6 Discussion

These results demonstrate that adversarial persuasion is a potent attack mechanism capable of catastrophically disrupting AFC pipelines. *Manipulative Wording* techniques, especially, can substantially degrade retrieval performance (Recall@5 drops by 0.54 points), decoupling the claim from its supporting evidence. Even if evidence retrieval is perfect, *Manipulative Wording* can still mislead the classifier, resulting in evasion rates of up to 27%.

The efficacy of these techniques aligns with recent findings from Glockner et al. (2024), who showed that claims that are vague or encourage inferences elicit substantial disagreement even between human fact-checkers. Consequently, *Manipulative Wording* attacks could be countered by reducing ambiguity through grounding claims in richer, structured contexts that limit the scope for rhetorical interpretation. Our findings support this intuition in two ways. First, models without access to evidence exhibit substantially higher evasion rates (e.g., *Obfuscation*: $>$65% compared to $<$38% with evidence), which underscores the need to ground claims in explicit context. Second, providing denser and more structured evidence improves robustness, e.g. *Obfuscation* evasion rates drop from 38% on FEVER to 15% on FEVEROUS, as the latter provides structured evidence and $\approx 2\times$ more evidence per claim.

7 Conclusion & Future Work

This paper introduced Persuasion Injection Attacks, the first semantic adversarial framework that weaponises persuasion techniques against AFC systems. Experiments on two benchmark datasets showed that AFC pipelines fail to disentangle persuasive rhetoric from factual content and optimised persuasion attacks can lead to collapse in both retrieval and classification (even when gold-evidence is provided), demonstrating that persuasive rhetoric constitutes a potent adversarial strategy.

The findings motivate our future work on making AFCs more robust, in particular to *Manipulative Wording* attacks. We also plan to investigate whether this vulnerability persists across languages and topics, which is motivated by evidence that different persuasion techniques tend to be employed in different disinformation topics (Leite et al., 2025).

Acknowledgments

This work is supported by the Engineering and Physical Sciences Research Council (EPSRC) under grant UKRI3352: Longitudinal, Multilingual, and Multi-format Investigation and Detection of LLM-Generated Disinformation. João A. Leite is supported by a University of Sheffield EPSRC Doctoral Training Partnership Scholarship.

Ethical Considerations

Potential Misuse By developing a systematic framework for injecting persuasion into claims, there is a risk that malicious actors could use these techniques to craft more deceptive disinformation that evades automated detection. However, by documenting and exposing these vulnerabilities, we enable the community to develop more robust defence mechanisms. Furthermore, we do not introduce new persuasion techniques in this work. Instead, we use techniques that are well-documented in prior work on persuasion and propaganda (Martino et al., 2020; Piskorski et al., 2023).

Data and Artifacts We use the FEVER (`https://fever.ai/dataset/fever.html`) and FEVEROUS (`https://fever.ai/dataset/feverous.html`) datasets, which are publicly available under CC-BY-SA 3.0 licenses. Our use of these artifacts is consistent with their intended use as benchmarks for verifying factual claims. We release our attack code under an open-source MIT license to facilitate reproducibility and allow the community to freely reuse, modify, and build upon our adversarial framework for defensive research purposes.

Limitations

This study is restricted to English, and the findings may not directly transfer to other languages without adaptation to account for linguistic and cultural variation.

The experiments rely on Wikipedia-based benchmarks (FEVER and FEVEROUS). While these offer controlled conditions for evaluating factual reasoning, they may differ from disinformation in social media posts or news articles. Thus persuasion attacks may behave differently in these genres.

Finally, our adversarial variants were generated using modest-sized LLMs (around 7B parameters). The strength of individual persuasion techniques likely depends on the generator's capabilities, and more capable models might enhance techniques that appeared less effective in these experiments.

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, pages 6719–6736, USA. USENIX Association.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

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, 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.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press.

Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad Characters: Imperceptible NLP Attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1987–2004.

Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The Persuasive Power of Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:152–163.

Sanxing Chen, Yukun Huang, and Bhuwan Dhingra. 2025a. Real-time Factuality Assessment from Adversarial Feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1610–1630, Vienna, Austria. Association for Computational Linguistics.

Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, and Baishakhi Ray. 2025b. Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10091–10109, Suzhou, China. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. 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 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. 2025. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. 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 1543–1558, Albuquerque, New Mexico. Association for Computational Linguistics.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-Checking Ambiguous Claims with Evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Andrew M. Guess and Benjamin A. Lyons. 2020. *Misinformation, Disinformation, and Online Propaganda*, page 10–33. SSRC Anxieties of Democracy. Cambridge University Press.

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. 2025. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884.

Maram Hasanain, Md Arid Hasan, Mohamed Bayan Kmainasi, Elisa Sartori, Ali Ezzat Shahroor, Giovanni Da San Martino, and Firoj Alam. 2025. PropXplain: Can LLMs enable explainable propaganda detection? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23855–23863, Suzhou, China. 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.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking Fake News for Real Fake News Detection: Propaganda-Loaded Training Data Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

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.

João A. Leite, Olesya Razuvayevskaya, Carolina Scarton, and Kalina Bontcheva. 2025. A Cross-Domain Study of the Use of Persuasion Techniques in Online Disinformation. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pages 1100–1103, New York, NY, USA. Association for Computing Machinery.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira.

2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Fanzhen Liu, Sharif Abuadbba, Kristen Moore, Surya Nepal, Cecile Paris, Jia Wu, Jian Yang, and Quan Z. Sheng. 2025. Adversarial Attacks Against Automated Fact-Checking: A Survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22979–23001, Suzhou, China. 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.

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.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A Survey on Computational Propaganda Detection. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, volume 5, pages 4826–4832.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*.

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.

Nikolaos Nikolaidis, Jakub Piskorski, and Nicolas Stefanovitch. 2024. Exploring the Usability of Persuasion Techniques for Downstream Misinformation-related Classification Tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6992–7006, Torino, Italia. ELRA and ICCL.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2025. Measuring and Benchmarking Large Language Models' Capabilities to Generate Persuasive Language. 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 10056–10075, Albuquerque, New Mexico. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Muhammad Salman, Asif Hanif, Shady Shehata, and Preslav Nakov. 2023. Detecting Propaganda Techniques in Code-Switched Social Media Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16794–16812, Singapore. Association for Computational Linguistics.

Ivan Srba, Olesya Razuvayevskaya, João A. Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, Carolina Scarton, Kalina Bontcheva, and Maria Bielikova. 2025. A survey on automatic credibility assessment using textual credibility signals in the era of large language models. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

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.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The Fact Extraction and VERification (FEVER) Shared Task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Hanqian Wu, Xinwei Li, Lu Li, and Qipeng Wang. 2022. Propaganda Techniques Detection in Low-Resource Memes with Multi-Modal Prompt Tuning. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and

Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Definition of Persuasion Techniques & Prompt Templates

This appendix documents the persuasion taxonomy and prompt templates used to generate claim variants throughout our experiments. We provide (i) the definitions of all persuasion techniques considered, and (ii) the exact prompt templates used for adversarial generation and paraphrasing, to support transparency and reproducibility.

Persuasion Taxonomy. Table 5 lists the full set of 23 persuasion techniques considered in this work, grouped into 6 high-level categories. Technique definitions follow the taxonomy introduced by Piskorski et al. (2023).

Adversarial Generation Prompt. Figure 4 shows the universal prompt template used to generate persuasive claim variants. For each persuasion technique, the template is instantiated with definitions drawn from the taxonomy in Table 5.

The prompt explicitly constrains the generator to preserve the original factual content, ensuring that generated variants differ in persuasive form rather than underlying veracity.

Paraphrase Prompt. Figure 5 presents the prompt used to generate neutral paraphrases of claims. This prompt instructs the model to rewrite the input claim while preserving its original meaning and tone.

Paraphrased claims are used as one of the baseline transformations in our experimental evaluation (see subsection 4.4).

B Manual Validation of Label Preservation

To ensure that persuasive claim variants preserve the ground-truth veracity label of the original claim, we performed a manual validation study to rule out semantic drift as a confounding factor in the observed attack effectiveness.

Annotation Protocol. We sampled a random subset of 690 adversarial claims: 30 per persuasion technique, stratified by dataset (FEVER/FEVEROUS) and ground-truth label (*True*/*False*).

Each instance was annotated by a trained annotator who was shown the persuasive claim together with the corresponding gold-evidence. The annotator assigned one of three verdicts: TRUE, FALSE, or NOT ENOUGH INFO (NEI).

Decision Procedure. Annotation followed a fixed three-step procedure designed to decouple rhetorical form from factual content:

1. **Core Fact Identification:** isolate the central factual assertion expressed by the claim.
2. **Evidence Verification:** determine whether the gold-evidence explicitly supports or refutes this assertion.
3. **Ambiguity Resolution:** if the claim becomes subjective, evaluative, or underspecified such that it cannot be verified given the evidence, assign *NEI*.

Label Preservation Criterion. An adversarial claim was considered *label-preserving* if the annotator's verdict matched the original ground-truth label. Claims annotated as *NEI* or receiving a verdict opposite to the original label were counted as non-preserving.

Technique Selection and Filtering. Table 6 reports label preservation, label flip, and ambiguity rates for all 23 persuasion techniques.

Based on this validation, **all techniques with label preservation rates $\leq 80\%$ were excluded from the experiments.** This filtering step removed seven techniques. The remaining sixteen retained techniques show a label preservation rate of $\mathbf{90.22}\%$. For the claims that failed the preservation check, less than $\mathbf{1.5}\%$ correspond to label flips (i.e., *False*→*True* or *True*→*False*).

C Generator Robustness

To verify that the susceptibility of AFC systems to Persuasion Injection reflects a property of the fact-checking pipeline rather than an artefact of a particular generative model, we replicate the *Blind* attack using LLAMA-3-8B-INSTRUCT.

We conduct this robustness check on **FEVEROUS**, which is more challenging to attack overall

Instruction: Rewrite the claim to inject persuasive wording of type: "{TECHNIQUE_NAME}".
Definition: {DEFINITION}

Goal: Keep the original factual content identical (no new facts, numbers, or named entities). Modify wording only to introduce the specified persuasion technique.

Constraints:
- Do NOT change the factuality of the claim.
- Keep length within ±20% of original.
- Produce a single-sentence output.
- If the claim is already clearly using this technique, produce a stronger, clearer instance of it (without adding facts).

Few-shot outputs that demonstrate the tone for "{TECHNIQUE_NAME}":
{EXAMPLES}

Now rewrite the following claim into a version that injects "{TECHNIQUE_NAME}":
CLAIM: "{CLAIM}"

IMPORTANT: Output ONLY the rewritten claim text. Do not include any prefixes (like "Here is the rewritten claim:"), explanations, or commentary. Start directly with the claim text.

Figure 4: **Prompt template for persuasion injection attacks.**

Instruction: Produce a paraphrase of the provided claim. Keep the same length as much as possible.

Example:
Input: "The senator voted for the amendment in 2019."
Output: "In 2019, the senator supported the amendment."

Claim: {CLAIM}
Paraphrase:

IMPORTANT: Output ONLY the paraphrased claim text. Do not include any prefixes, explanations, or commentary.

Figure 5: **Prompt template for the paraphrase attack.**

(see subsection 5.1). For consistency, both generators are evaluated on the same subset of 15 **validated persuasion techniques** used in the main experiments (see Table 6).

As shown in Table 7, attacks generated by LLAMA-3-8B-INSTRUCT induce nearly identical accuracy degradation to those generated by QWEN2.5-7B under both target settings (claim-only and evidence-based). This consistency indicates that the observed vulnerabilities are not driven by generator-specific behaviour.

D Training Details & Hyperparameters

We fine-tune RoBERTa-base classifiers on the FEVER-style fact-checking splits, treating each claim concatenated with its supporting evidence sentences (or claim-only in the ablation) as a single sequence. Inputs are truncated and padded to 512 tokens. Models use standard Hugging Face training. We use an effective batch size of 256 (128 per device × 2 gradient accumulation steps). We optimise for five epochs with weight decay 0.01, linear warmup over 10% of steps, gradient norm clipping at 1.0, and early stopping with patience one based on dev AUC ROC. We mitigate class imbalance with inverse-frequency loss weights computed from the training data.

We perform a grid search over learning rates {2e-5, 3e-5, 5e-5}, and select the best by dev AUC ROC. Then, we fine-tune the model using the best learning rate using three different random seeds. For each experiment, we aggregate dev metrics across seeds (mean ± sd) and keep the best-performing checkpoint for downstream use. Table 8 displays the classification results on the dev set.

All experiments were conducted using a single NVIDIA-V100 (40GB VRAM) GPU. We estimate approximately 5 GPU hours to fully reproduce all experiments, including generating the adversarial texts, fine-tuning the veracity classifiers, and performing inference, for both FEVER and FEVEROUS.

E Examples of Evasion

Table 9 presents representative examples of adversarial claims for the persuasion techniques identified as persistent threats in RQ2 (see subsection 5.3). For each technique, we show one example from **FEVER** and one from **FEVEROUS**, illustrating how persuasive reframings can induce evasion even when the underlying factual content is contradicted by gold-evidence.

Across techniques, the examples reveal a com-

Category	Persuasion Technique	Definition
Attack on Reputation	Name Calling / Labelling	Attach a loaded, insulting, or demeaning label to a person or group; targets identity rather than argument.
	Guilt by Association	Associate the target with a negatively perceived person, group, or event to discredit them.
	Casting Doubt	Undermine credibility by questioning competence, motives, or facts about the target.
	Appeal to Hypocrisy	Accuse the opponent of hypocrisy to deflect criticism.
	Questioning the Reputation	Attack the reputation of the target by making strong negative claims about it.
Justification	Flag Waving	Justify a claim by appealing to group pride, patriotism, or shared values.
	Appeal to Authority	Support a claim by invoking an authority's endorsement to lend weight.
	Appeal to Popularity	Justify a claim because many people support it or do it.
	Appeal to Values	Link the claim to widely held positive values (tradition, religion, ethics).
	Appeal to Fear	Promote or reject an idea by evoking fear about its consequences.
Distraction	Strawman	Misrepresent the claim into a weaker or extreme version that is easier to refute.
	Red Herring	Divert attention from the claim by bringing up an unrelated topic.
	Whataboutism	Deflect criticism by pointing to different wrongdoing elsewhere.
Simplification	Causal Oversimplification	Blame a complex outcome on a single cause, ignoring other factors.
	False Dilemma	Present only two options when more exist; force a black-and-white choice.
	Consequential Oversimplification	Argue that one action will inevitably lead to extreme consequences.
Call	Slogan	Short, catchy phrase intended to persuade emotionally.
	Conversation Killer	Phrases that shut down discussion and discourage critical thinking.
	Appeal to Time	Argue urgency or timeliness to force immediate action.
Manipulative Wording	Loaded Language	Use emotionally charged words or phrases to sway feelings.
	Obfuscation	Use unclear or vague wording to hide meaning or confuse the audience.
	Exaggeration or Minimisation	Overstate or downplay importance to skew perception.
	Repetition	Repeat words or phrases to increase persuasion through familiarity.

Table 5: **Taxonomy of persuasion techniques.** The 23 techniques are grouped into 6 high-level categories following the hierarchy defined by Piskorski et al. (2023).

mon pattern: persuasive modifications do not introduce new factual claims, but instead alter emphasis, framing, or salience in ways that reduce the model's sensitivity to explicit contradictions in the evidence. *Obfuscation* replaces concrete predicates with vaguer temporal or semantic expressions; *Repetition* reiterates misleading elements to distract from the false predicate; and *Slogan* and *Flag Waving* inject rhetorical or identity-based language that shifts the claim away from a strictly factual register. *Appeal to Popularity* further amplify this effect by introducing evaluative or social cues that the model appears to treat as non-factual context.

In all cases, the evidence remains sufficient to refute the claim, yet the evidence-based AFC model flips from a correct *False* prediction on the original claim to an incorrect *True* prediction on the adversarial variant.

Persuasion Technique	Preservation	Flip	Ambiguity
INCLUDED (>80% Preservation)			
Repetition	96.67	3.33	0.00
Obfuscation	93.33	0.00	6.67
Flag Waving	93.33	0.00	6.67
Casting Doubt	93.33	0.00	6.67
Conversation Killer	93.33	6.67	0.00
Appeal to Values	93.33	0.00	6.67
Slogan	90.00	0.00	10.00
Loaded Language	90.00	0.00	10.00
Appeal to Authority	90.00	0.00	10.00
Appeal to Popularity	90.00	0.00	10.00
Appeal to Hypocrisy	86.67	3.33	10.00
Red Herring	86.67	0.00	13.33
Guilt by Association	86.67	3.33	10.00
Whataboutism	86.67	3.33	10.00
Name Calling / Labelling	83.33	3.33	13.33
Overall (Micro Avg.)	**90.22**	**1.56**	**8.22**
EXCLUDED (≤80% Preservation)			
Appeal to Time	80.00	0.00	20.00
Causal Oversimplification	76.67	6.67	16.67
False Dilemma	56.67	6.67	36.67
Exaggeration / Minimisation	36.67	3.33	60.00
Consequential Oversimplification	30.00	3.33	66.67
Questioning the Reputation	26.67	0.00	73.33
Strawman	16.67	6.67	76.67
Appeal to Fear / Prejudice	6.67	0.00	93.33

Table 6: **Label preservation analysis per persuasion technique (%).** *Preservation* denotes the proportion of persuasive claims whose annotated verdict matches the original ground-truth label. *Flip* denotes cases where the annotated verdict contradicts the original label (TRUE↔FALSE). *Ambiguity* denotes claims judged as *Not Enough Info*, i.e., unverifiable given the gold-evidence.

Target Setting	Baseline (No Attack)	Blind Attack Accuracy Qwen2.5-7B	Blind Attack Accuracy Llama-3-8B
Claim-Only	0.641	0.575	0.568
Evidence-Based	0.894	0.726	0.727

Table 7: **Robustness across generators (FEVEROUS).** AFC accuracy under blind attacks generated by Qwen2.5-7B and Llama-3-8B. Both generators yield comparable performance degradation, indicating generator-agnostic attack effectiveness.

Dataset	Model	LR	Dev Macro F1	Dev ROC AUC
FEVER	Claim-Only	2e-5	0.826 ± 0.002	0.913 ± 0.001
	+ Evidence	5e-5	0.953 ± 0.001	0.985 ± 0.001
FEVEROUS	Claim-Only	3e-5	0.661 ± 0.002	0.719 ± 0.001
	+ Evidence	3e-5	0.910 ± 0.002	0.972 ± 0.002

Table 8: **Development set performance.** Best checkpoint performance (mean±std) and optimal learning rates (LR).

Persuasion Technique	**Dataset**	**Original Claim** *(predicted False)*	**Adversarial Claim** *(predicted True)*	**Gold-Evidence**
Obfuscation	FEVER	Hannah and Her Sisters' ending is **unrelated to Thanksgiving**.	The conclusion of "Hannah and Her Sisters" has a nuanced **disconnect from the holiday season**.	Hannah and Her Sisters is a 1986 American comedy-drama film which tells the intertwined stories of an extended family over two years that begins and ends with a family Thanksgiving dinner.
	FEVEROUS	The Malaysian plover lays eggs on beaches which are **incubated for 50 days**.	The species incubates its eggs in coastal areas for a duration of **approximately one and a half months**.	The eggs are incubated by both the male and female for about 30 days, and then both parents care for the precocial chicks until they can fly after about 30 more days.
Repetition	FEVER	The Hunger Games are series of **three films**.	The Hunger Games, The Hunger Games, The Hunger Games–are series of **three films**.	The Hunger Games film series consists of four science fiction dystopian adventure films based on The Hunger Games trilogy of novels, by the American author Suzanne Collins.
	FEVEROUS	GZMB is a serine protease, an enzyme with **no responsibility for cleaving peptide bonds in proteins**.	GZMB is a serine protease, a serine protease with **no responsibility for cleaving peptide bonds in proteins**.	**1.** Granzyme B is a serine protease that in humans is encoded by the GZMB gene. **2.** Serine proteases (or serine endopeptidases) are enzymes that cleave peptide bonds in proteins, in which serine serves as the nucleophilic amino acid at the (enzyme's) active site.
Slogan	FEVER	The Nottingham Forest F.C. is **based in Lincolnshire, England**.	"Nottingham Forest F.C.: **Rooted in Lincolnshire**, Shining in Every Corner!"	Nottingham Forest Football Club is a professional association football club based in Nottingham, England.
	FEVEROUS	Steven Horsford **retired from the position** of Nevada Senate Majority Leader in 2009 and appointed a senator to lead the caucus election efforts during for the 2012 election cycle.	"**Step aside** for a new era: Horsford **Hands the Baton** in 2009!"	In February 2009, he assumed the position of Nevada Senate Majority Leader.
Flag Waving	FEVER	Honeymoon is the **third** major-label record by Lana Del Rey.	True fans of American music must recognize Honeymoon as the **third** major-label record by Lana Del Rey.	Honeymoon is the fourth studio album and third major-label record by American singer and songwriter Lana Del Rey.
	FEVEROUS	The Galle Polling Division is in the Galle Electoral District in Sri Lanka, with a 2012 total population of 101,749 covering 11,860 square miles; and Galle matched the Parliamentary Election results of the country **five out of seven times**.	True Sri Lankans must support the accurate representation of their home in the Galle Polling Division, which matched the Parliamentary Election results **five out of seven times**.	The winner of Galle has matched the final country result 5 out of 8 times.
Appeal to Popularity	FEVER	One Dance was Drake's first number one single in **34 countries**.	Everyone is celebrating One Dance as Drake's first number one single in **34 countries**.	"One Dance" reached number one in **15 countries**, including Australia...
	FEVEROUS	The First Question Award (All tracks are written by Cornelius) released in February 25, 1994 is the debut studio album by Cornelius, which peaked at **number one** on the Oricon Albums Chart.	The First Question Award, released in February 25, 1994, is the debut studio album by Cornelius—everyone is still celebrating its success at **number one** on the Oricon Albums Chart.	**1.** Japanese Albums (Oricon): 4. **2.** The First Question Award peaked at number four on the Oricon Albums Chart.

Table 9: **Examples of effective persuasion techniques.** Examples of adversarial claims for persuasion techniques identified as effective threats (see subsection 5.3). Each example alters the claim to cause evasion: the evidence-based AFC model flips from a correct prediction (*False*) in the original claim to an incorrect prediction (*True*) in the adversarial claim. *False* information is in **bold**, and direct counter evidence is underscored.