



Fact-Saboteurs: A Taxonomy of Evidence Manipulation Attacks against Fact-Verification Systems

Sahar Abdelnabi and Mario Fritz, *CISPA Helmholtz Center for Information Security*

<https://www.usenix.org/conference/usenixsecurity23/presentation/abdelnabi>

This paper is included in the Proceedings of the
32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Open access to the Proceedings of the
32nd USENIX Security Symposium
is sponsored by USENIX.

Fact-Saboteurs: A Taxonomy of Evidence Manipulation Attacks against Fact-Verification Systems

Sahar Abdelnabi and Mario Fritz
CISPA Helmholtz Center for Information Security

Abstract

Mis- and disinformation are a substantial global threat to our security and safety. To cope with the scale of online misinformation, researchers have been working on automating fact-checking by retrieving and verifying against relevant evidence. However, despite many advances, a comprehensive evaluation of the possible attack vectors against such systems is still lacking. Particularly, the automated fact-verification process might be vulnerable to the exact disinformation campaigns it is trying to combat. In this work, we assume an adversary that automatically tampers with the online evidence in order to disrupt the fact-checking model via *camouflaging* the relevant evidence or *planting* a misleading one. We first propose an exploratory taxonomy that spans these two targets and the different threat model dimensions. Guided by this, we design and propose several potential attack methods. We show that it is possible to subtly modify claim-salient snippets in the evidence and generate diverse and claim-aligned evidence. Thus, we highly degrade the fact-checking performance under many different permutations of the taxonomy's dimensions. The attacks are also robust against post-hoc modifications of the claim. Our analysis further hints at potential limitations in models' inference when faced with contradicting evidence. We emphasize that these attacks can have harmful implications on the inspectable and human-in-the-loop usage scenarios of such models, and we conclude by discussing challenges and directions for future defenses.

1 Introduction

Disinformation and misinformation have recently raised much-deserved global and societal concerns [71]. They can have major harmful consequences on our core democratic values (e.g., polarizing the public's opinions and affecting elections [4]), individuals' lives (e.g., spreading hurtful rumors and false accusations [21]), and society's health and security (e.g., spreading non-scientific claims about pandemics [16]), to name a few. To face such dangers, fact-checking and verification (used interchangeably [79]) is essential to debunk

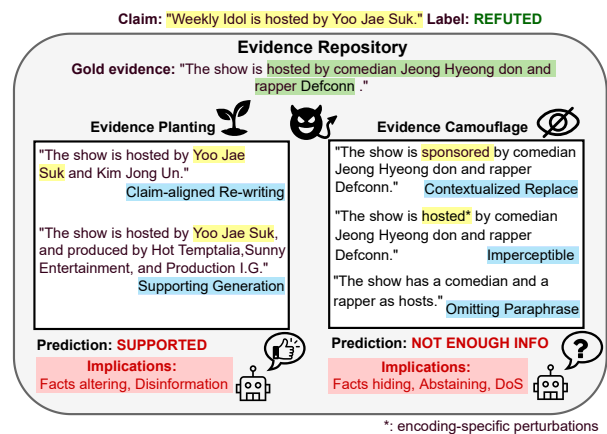


Figure 1: We propose a taxonomy and several evidence manipulation attacks against fact-verification models. The taxonomy includes the attacks' target: **Camouflaging** (to hide the relevant evidence) and **Planting** (to introduce a deceiving one). The attacks might negatively affect the inspectability and humans in the loop.

false claims and limit their dissemination; it is a strategy now employed by many platforms [46, 81] and an established common practice in journalism [66].

A Need for Automation. However, manual fact-verification is time-consuming [27]. Given the proliferation of online misinformation and its rapid spread, human fact-checkers can find it burdensome and challenging to keep up [28]. This motivated an active research area within the Natural Language Processing (NLP) community to automate the evidence-based claim verification task [79, 64, 61, 54, 29, 76]. One of the largest and most popular frameworks in this domain is Fact Extraction and Verification (FEVER) [79], which aims to verify human-written claims against Wikipedia as a relatively credible source.

Besides academic interest, automation has been discussed in practice among fact-checking organizations and journalists [26, 34]. While professional fact-checking remains principally manual, some organizations are working on preliminary

prototypes [23, 8, 72] to automate various fact-checking steps, with signs that they can be potentially useful as complementary assistive solutions with human supervision [75, 32].

Fact-Checking Attacks. In addition to recent advances, previous work studied adversarial attacks on models by changing the formulation of claims [78, 30, 6]. This primarily aimed at diagnostically revealing the dataset’s and models’ biases without considering malicious intents, i.e., the evidence databases were assumed to contain only factual information. To the best of our knowledge, Du et al. [20] is the only work that studied automated evidence manipulation attacks by synthesizing AI-generated articles given the claim [90]. However, their approach lacked a comprehensive analysis and formulation of the threat model and possible attack vectors.

Our Work. We take analogies from journalism, where manipulated media constitutes a major challenge [19, 1]. We assume an adversary that disrupts the automatic fact-checking process by *automatically manipulating evidence repositories* to obscure or introduce misleading evidence. We propose a broad taxonomy (Figure 2) to derive our systematic exploration of evidence manipulation attacks. The taxonomy spans different dimensions: the attacker’s **targets** (evidence camouflaging or planting as in Figure 1), the **constraints** (the control they have over modifying the repository and the original context), and the **capabilities** (the models available to launch the attack). We also evaluate the attacks with respect to the attacker’s **knowledge** (the attacker’s dataset and the white- or black-box access to the evidence retrieval and verification models). We highlight that these attacks can negatively affect humans in the loop [48, 83] (e.g., models potentially assisting fact-checkers or end-users) – models should allow the interpretability of the reached verdict via, e.g., inspecting the salient evidence [54, 79, 7]. However, by camouflaging evidence, attackers could perturb or deprioritize the originally-relevant evidence. Thus, it might not be retrieved or be irrelevant/inconclusive if it is (sometimes even to humans). In contrast, by planting targeted factually-wrong evidence, humans in the loop might be deceived by these campaigns (i.e., spear disinformation [93]). Overall, this might cause a false sense of security, especially for end-users, given the lack of a verdict or enforcing the manipulated one.

Why Should We Study Fact-Checking Attacks? Even under human supervision, attacks that compromise the integrity of models have dangerous implications, ranging from Denial of Service (DoS) to automatically manipulating critical sources needed for human verification. Besides, these tools might be used more widely in the future [8], given the rapid progress of NLP. In addition, automated fact-checking has also been considered a promising sustainable solution to detect machine-generated text [90]. Given this potential, it is crucial to proactively understand the vulnerabilities and limitations of fact-checking models and design adversary-aware ones, now and before large-scale deployment.

Why Should We Study AI-Generated Attacks? Large

Language Models (LLMs) [12, 13] can generate highly credible and plausible content that humans often struggle to detect [39, 2, 15]. While human-generated content remains what mainly fuels current disinformation campaigns [22, 18, 73, 62], the wide accessibility of LLMs might enable and facilitate the creation of disinformation and automatic manipulation at scale, calling for an early evaluation of such threats.

Contributions. In summary, we make the following contributions: 1) We propose a systematic taxonomy to conduct the first comprehensive investigation of automated evidence manipulation attacks. 2) We propose extensive and highly successful attacks that vary in their **targets**, stealthiness, context-preserving **constraints**, and the adversary’s **capabilities** and **knowledge**. 3) We discuss models’ limitations, future defense directions, and the need to model possible malicious manipulations in the design of fact-verification models.

2 Preliminaries and Related Work

This section briefly introduces the automatic fact-checking frameworks and the technical methods we used to construct the attacks. We report previous real-world examples of evidence manipulation that motivate and derive our work. Finally, we discuss our contributions in comparison with related work.

FEVER Dataset and Framework. The FEVER dataset [79] consists of over 185k claims manually written based on Wikipedia. Each claim is annotated as one of three labels: ‘Supported’ (SUP - 80k train, 6k dev. sets), ‘Refuted’ (REF - 29k train, 6k dev. sets), or ‘Not Enough Info’ (NEI - 35k train, 6k dev. sets). The REF and NEI claims were constructed by instructing annotators to generate mutations of correct claims (e.g., negation, entity substitution). SUP and REF claims were labelled with the golden evidence needed for verification. There have been other specific, yet smaller, datasets (e.g., scientific [84] and COVID-19 claims [61]). However, we use FEVER due to its popularity and large size. We use the training set (or subsets from it) to train the attack models and perform the attacks on the dev. set.

The task involves the open-domain verification of claims, where the golden evidence is not pre-identified at test time. Specifically, the task consists of three steps: 1) document retrieval (obtaining relevant Wikipedia pages given their titles and the claim), 2) evidence retrieval (selecting evidence sentences from the retrieved pages), and 3) verifying the claim given the retrieved sentences. Thorne et al. [79] proposed a simple baseline that retrieves pages and evidence sentences based on TF-IDF vectors followed by an entailment model [50]. Many other improvements have been achieved by employing state-of-the-art transformers [37, 42] in both the retrieval and verification tasks [92, 43, 49]. We test the attacks on the **KGAT** [43] as one of the most prominent models and due to its easy-to-use public implementation. It uses a BERT-based evidence retrieval that was trained contrastively on golden evidence vs. other random sentences. Then, it is

used to rank sentences according to the claim. The verification model is based on a graph neural network with BERT or RoBERTa backbones for representations. The number of evidence sentences used in the verification step is capped to the top 5 retrieval results. We also test on CorefBERT [89] that initializes the KGAT verification model with a BERT model fine-tuned to better handle contextual coreferential relations.

NLP Adversarial Attacks. Previous work generated adversarial attacks by word-level substitutions based on semantic constraints via word embeddings search [5] or contextualized replacements [41]. More recent work used imperceptible changes [11] to manipulate the output of NLP classifiers. We apply these attacks to perturb the evidence to achieve the evidence camouflaging **target**; they distort the salient snippets within the evidence rather than semantically shifting the polarity with respect to the claim.

AI-Generated and Re-written Evidence. We utilize conditional language generation to achieve targeted disinformation given claims, meeting the evidence planting **target**. We also use methods related to the task of text re-writing (e.g., style transfer [67], sentiment-changing [10], paraphrasing [45], and factual modification [77, 65]). Specifically, we conduct claim-guided evidence re-writing to 1) remove claim-salient snippets by paraphrasing or conditional generation for the camouflaging **target**, or 2) align the evidence with the wrong claim for the planting **target**.

Evidence Manipulation: Examples. Being an open source, Wikipedia is susceptible to manipulative edits [59, 20]. Some of these are designed to cause vandalism and be humorous [82], and thus, are easy to be detected. However, some could last for as long as several years [86]. It was even subject to pervasive organized disinformation campaigns that lasted for almost a decade to promote political or ideological orientations (e.g., far-right groups) [17]. Other incidents included deleting incriminating information [74], deleting political scandals [85, 80], and editing a description of a medical procedure from ‘controversial’ to ‘well documented and studied’ [73], **closely matching** our attacks’ **targets**: evidence camouflaging and evidence planting.

While we use a Wikipedia-based dataset, the concept of seeding erroneous evidence can be applied to other mediums, social platforms, and websites, sometimes with even less constraint and moderation than Wikipedia. Case studies [38, 21] demonstrate events where participants compiled *evidence collages* of verified and unverified information (making it harder to verify) and used them to affect the public, journalists, and authorities. Thus, we take analogies from these incidents and investigate whether evidence manipulation can be automated by AI technologies to attack fact-verification models.

Related Work. Du et al. [20] studied a similar task to ours. However, via the lens of our taxonomy, they only studied one type of planting attacks. In contrast, we extend the **targets** to evidence **camouflaging**, proposing *stealthier* (sometimes completely factual) attacks that hide the facts instead of intro-

ducing evidently false content. Via camouflaging, we highly succeed in *attacking correct claims*, which was not covered in their work. We further extend the **planting** attacks and propose an evidence-rewriting attack that is more *context-preserving* (varying the **constraint** dimension) yet more successful. Even within the same constraints, we address multiple limitations reported in their work. We generate evidence that is better coordinated with the claims and more similar to the golden evidence distribution. As a result, we produce both *more successful and more plausible attacks* while still having a limited-**knowledge** adversary. Our planting attacks *show more success in SUP to REF inversion*, which was not possible at all previously, and reveal limitations of fact-verification models when faced with contradicting evidence.

3 Threat Model

We assume an adversary \mathcal{A} that targets a fact-checking model \mathcal{M} via evidence manipulation to serve a political agenda or achieve personal gain. \mathcal{M} might be employed (by defender \mathcal{D}) to automatically flag disinformation or assist fact-checkers or end-users by outputting warnings and pointing to related evidence. \mathcal{M} consists of retrieval and verification models ($\mathcal{R}_{\mathcal{D}}$ and $\mathcal{V}_{\mathcal{D}}$, respectively). Similarly, the adversary has retrieval and verification models ($\mathcal{R}_{\mathcal{A}}$ and $\mathcal{V}_{\mathcal{A}}$, respectively) that mirror \mathcal{M} . \mathcal{D} has a labelled fact-verification dataset $\mathcal{S}_{\mathcal{D}}$. \mathcal{A} has a dataset $\mathcal{S}_{\mathcal{A}}$, where $\mathcal{S}_{\mathcal{A}} \subseteq \mathcal{S}_{\mathcal{D}}$. In the following, we outline the taxonomy of the attacks, as depicted in Figure 2.

1) Adversary’s Targets. Rather than generically assuming that \mathcal{A} aims to fool \mathcal{M} , we take inspiration from previously observed manual evidence manipulation attempts to further categorize the attacks’ logical targets into *camouflaging* and *planting*. This is also motivated by the potential deceptive implications of these targets on humans.

In **camouflaging**, \mathcal{A} intends to hide the sentences needed to verify the claim (e.g., [74, 85, 80]). Simply removing them might be suspicious and not always applicable (e.g., removing image captions). Thus, we investigate *more subtle* attacks that work as a ‘smarter delete’ by changing the evidence such that it is less relevant to the claim (because it is either perturbed or does not contain the needed information anymore). These attacks can be applied to both REF and SUP claims. As a result, the claims would mostly become unverifiable, and the model would change its prediction to NEI. In **planting**, \mathcal{A} intends to actively change the narrative to change \mathcal{M} ’s prediction (a less subtle adversary, e.g., [73, 17]). This can be done by i) partial re-writing of the initially relevant evidence or ii) inserting fully newly generated sentences to, e.g., have more flexibility or pre-emptively fill the data void [25]. The first can be used to, e.g., change the prediction from REF to SUP, while the second also allows changing from NEI to SUP.

2) Adversary’s Constraints. We set two constraints for \mathcal{A} : how much the attacks need to preserve the context, and how the evidence repository can be modified.

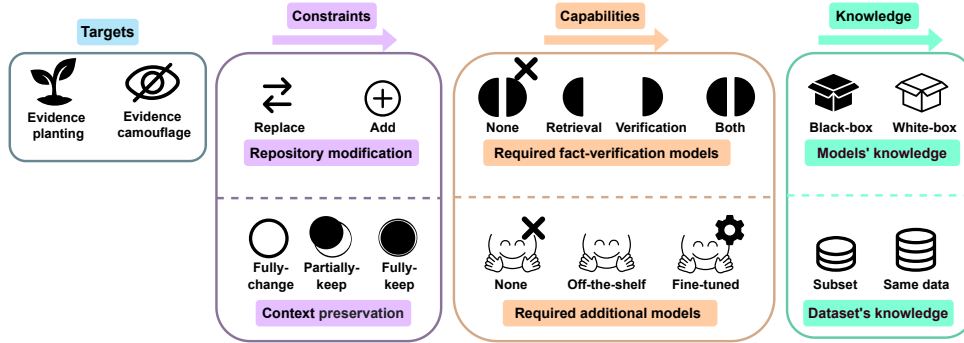


Figure 2: Taxonomy of the threat model’s dimensions. We categorize and evaluate the attacks in terms of the adversary’s **targets**, **constraints** (preserving context and modifying the evidence repository), **capabilities** (which fact-verification and other external models are needed to compute the attack), and **knowledge** (access to the downstream fact-verification models and dataset). Arrows indicate an increasing direction of the dimension.

Many works in adversarial NLP assumed that adversarial sentences should preserve the entailment/label in order to be used as a diagnostic tool for the models’ robustness [6, 5, 78]. However, since we study disinformation and information manipulation, we do not exclusively assume that the needed facts still exist. Instead, the manipulations should be stealthy by being sensical and grammatical. Besides, they might need to completely or partially preserve the **context**¹ to avoid detection in the case of, e.g., a highly moderated page or website, or pass in disinformation within partially factual content to increase the perceived credibility [21]. In our attacks, sentence editing can preserve the context more than generating entirely new sentences, and imperceptible attacks and paraphrases fully preserve the context by not adding new information.

We also analyze the attacks with respect to the repository **modification** method needed for the attack to succeed. In the camouflaging attacks, we empirically found that \mathcal{A} needs to ‘replace’ the original evidence with the manipulated one. However, for planting attacks, \mathcal{A} can have an ‘add’ control only. We found that even when the planted evidence exists along with the old one², \mathcal{M} can still be swayed to agree with a wrong claim. This is especially relevant in setups beyond Wikipedia, where \mathcal{A} might be constrained by not having a ‘replace’ access to a specific source (e.g., a credible newspaper or a governmental source that is hard to infiltrate). Instead, they might resort to spamming the Internet and other repositories and sources with the intended narratives.

Finally, as we work on a Wikipedia-based dataset, we have a single evidence repository. However, in practice, the constraints can also include how many sources/repositories the adversary can access to poison or modify.

3) Adversary’s Capabilities. Next, we analyze the attacks in terms of the models \mathcal{A} needs to obtain/train in order to compute the attack. Specifically, we outline if \mathcal{A} needs to

¹By ‘context’, we mean how much information within the evidence sentence is replaced by new, possibly incorrect, information.

²An example of that in the case of Wikipedia would be to create a new page or append the evidence to another page.

| Target | Constraints | | Capabilities | | Attack | Labels |
|--------|--------------|---------|--------------|--------|---|--------|
| | Modification | Context | FV models | Others | | |
| | | | | | Lexical Variation (based on [5]) | R+S |
| | | | | | Contextualized replace (based on [41]) | R+S |
| | | | | | Imperceptible (based on [11]) | R+S |
| | | | | | Imperceptible _{Ret} (based on [11]) | R+S |
| | | | | | Omitting paraphrase | R+S |
| | | | | | Omitting generate | R+S |
| | | | | | Claim-aligned rewriting +stance filtering | R |
| | | | | | Claim-aligned rewriting _{Ret} +retrieval filtering | R |
| | | | | | Supporting generation +stance filtering | NEI+R |
| | | | | | Claim-conditioned article generation (introduced in [20]) | NEI+R |

Table 1: The investigated permutations of the taxonomy’s dimensions and the attack methods that satisfy them. The ‘Labels’ column indicates which labels this attack can target, based on the attack’s properties or our empirical findings.

have fact-verification models ($\mathcal{R}_{\mathcal{A}}$ or $\mathcal{V}_{\mathcal{A}}$) in addition to other external off-the-shelf or fine-tuned models (e.g., a language generation model). For example, relevant-evidence editing attacks must have $\mathcal{R}_{\mathcal{A}}$ that ranks and returns the potentially relevant sentences, attacks targeting the entailment step need to have $\mathcal{V}_{\mathcal{A}}$, generating sentences from scratch might not need a retrieval but requires a language generator (either off-the-shelf or fine-tuned), etc.

4) Adversary’s Knowledge. As an orthogonal dimension, we evaluate the attacks in varying degrees of \mathcal{A} ’s knowledge, particularly the access and knowledge about $\mathcal{R}_{\mathcal{D}}$, $\mathcal{V}_{\mathcal{D}}$, and $\mathcal{S}_{\mathcal{D}}$. For the retrieval, we study a white-box scenario (i.e., $\mathcal{R}_{\mathcal{A}} = \mathcal{R}_{\mathcal{D}}$) and black-box scenarios where the architecture

is either the same or different. To minimize the attacks’ assumptions, we *never* use the white-box verification model to construct the attack (i.e., $\mathcal{V}_{\mathcal{A}} \neq \mathcal{V}_{\mathcal{D}}$), and we do not assume any knowledge about its exact framework. For all our attacks, we set $\mathcal{V}_{\mathcal{A}}$ as a model trained on pairs of claims and single evidence sentences, while $\mathcal{V}_{\mathcal{D}}$ is based on a graph neural network to capture the relationship among the evidence. Also, the backbone models can differ (e.g., BERT vs. RoBERTa). In practice, these white- and black-box scenarios can depend on whether a classifier is released by a developing company or only available as an API or a web interface [3].

Finally, we evaluate a setup where $\mathcal{S}_{\mathcal{A}} \subset \mathcal{S}_{\mathcal{D}}$. For Wikipedia, having a same-distribution dataset subset is a reasonable assumption, as the main limitation here would be to write and annotate the claims (i.e., the size of the dataset), assuming the dataset cannot be obtained in other ways. Beyond Wikipedia, the taxonomy can potentially extend to scenarios where \mathcal{D} ’s dataset is proprietary or from a different distribution.

4 Attacks on Fact-Verification Models

In this section, we describe the details of the investigated attacks, shown as a summary in Table 1. Starting from permutations of the proposed taxonomy, we explore possible technical methods that satisfy them. As discussed in section 3, we found that certain attack targets might need specific assumptions on the constraints and capabilities. Thus, exhaustive permutations are not feasible. Given the attack method, we indicate to which ground-truth labels it can be applied. Some attacks have inherent and logical properties of the labels they can target, e.g., camouflage is possible for REF or SUP labels since NEI labels do not have relevant evidence to begin with. Moreover, ‘claim-aligned rewriting’ is ideally for REF. However, for others, we indicate our empirical findings of what combinations of labels were possible (e.g., planting attacks were hardly successful on SUP).

In addition, Figure 3 depicts the attacks’ general flow. As discussed in section 3, attacks might or might not need a retrieval step depending on whether they edit existing evidence³ or generate a new one. After the attack sentences are computed, the evidence repository is modified according to the constraints. The attacks are then tested on the downstream model \mathcal{M} by first retrieving from all the manipulated evidence repository and then performing the verification step. In the following, we first discuss camouflaging, then planting attacks. To visualize the attacks with examples, see Figure 1 and Table 10 in Appendix B.

4.1 Camouflaging Attacks \rightleftharpoons R+S

Camouflaging attacks assume a ‘replace’ evidence manipulation constraint and can be applied to SUP and REF examples.

³We never assume that relevancy annotations (i.e., golden evidence labels) are required to run the attack at test time.

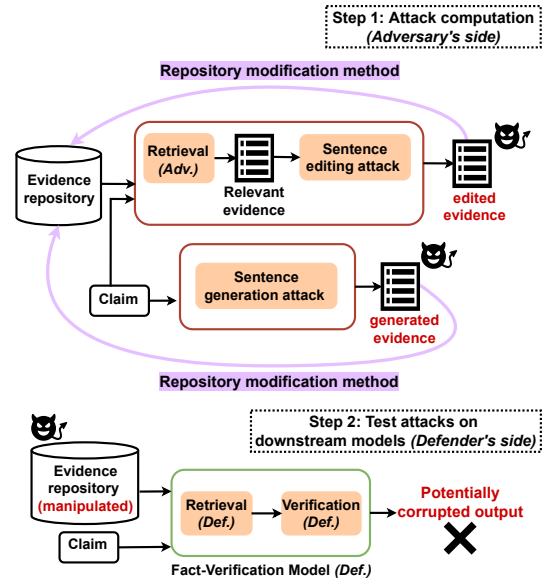


Figure 3: Attacks’ general pipeline. Some attacks might first need to retrieve the relevant evidence. Others can be constructed given the claims only. Next, the attack is tested on the downstream FEVER model \mathcal{M} (Step 2).

4.1.1 Lexical Variation

This attack is based on introducing lexical changes to attack the verification model $\mathcal{V}_{\mathcal{A}}$. Alzantot et al. [5] proposed to generate natural language adversarial examples via black-box access to a classification model. They use a population-based optimization algorithm that generates candidate sentences by finding the N nearest neighbors of a word based on GloVe embeddings [52]. Other techniques were employed to filter out unfitting words (e.g., a distance threshold and ensuring that nearest neighbors are synonyms [47]). The algorithm returns candidates that maximize the required target label.

This method was used previously to generate claim-based attacks on FEVER [30]. We here apply it to perturb the evidence while keeping the claims fixed. As a proxy to $\mathcal{V}_{\mathcal{D}}$, $\mathcal{V}_{\mathcal{A}}$ is a RoBERTa_{BASE} model trained on pairs of claims and golden evidence. For NEI claims, the evidence is selected from the retrieval results returned by $\mathcal{R}_{\mathcal{A}}$. We then apply the black-box attack against $\mathcal{V}_{\mathcal{A}}$. For each claim, we attempt to perturb the top sentences returned by $\mathcal{R}_{\mathcal{A}}$, where the target classification for SUP claims is REF and vice versa. Although this is a targeted attack, we show in our experiments that the perturbed sentences are generally less likely to be retrieved, achieving the camouflaging target.

4.1.2 Contextualized Replace

The previous lexical variation attack is limited in considering the context of the sentence since it uses GloVe embeddings with fixed nearest neighbors. To solve that, Li et al. [41] intro-

duced the BERT-attack to get more fluent and higher-quality perturbations. It is also a black-box attack against a classifier model (e.g., BERT). First, salient words in the sentence s are extracted by ranking the classification probability drop of the correct label o_y when masking a word w_i to form a masked sequence s_{w_i} : $I_{w_i} = o_y(s) - o_y(s_{w_i})$.

Then another pre-trained BERT masked language model (hence without fine-tuning: 🧠) is used to generate candidates for the ranked salient words. This has the advantage of being more context-aware and dynamic, without using heuristics such as a POS checker. The perturbations are restricted by a budget ϵ on the words to replace and a probability threshold on the masked language model’s candidates. The algorithm then returns the candidates maximizing a wrong prediction. Here, $\mathcal{V}_{\mathcal{A}}$ is a BERT_{BASE} model fine-tuned on sentence pairs.

4.1.3 Imperceptible 🕒 🕒 🕒

We examine a stealthy attack where the changes performed are invisible or imperceptible. Boucher et al. [11] used encoding-specific perturbations to produce indistinguishable sentences that nevertheless fool NLP classifiers. This might enable malicious actors to hide documents or avoid content moderation [11], a highly similar scenario to our camouflaging target.

This attack mainly breaks the tokenization step by replacing characters with their homoglyphs and inserting invisible characters, directionality, or deletion control characters. As these characters are outside the models’ dictionaries, the tokens would be mapped to *UNK* or incorrect sub-words. The attack is also performed via black-box access to a model and a differential evolution optimization algorithm [70] to minimize the logits of the correct prediction o_y : $x_{\mathcal{A}} = \arg \min_x o_y(x)$, bounded by a perturbation budget ϵ on the total number of changes. We use the previously mentioned BERT classifier as $\mathcal{V}_{\mathcal{A}}$. In our experiments, as expected, we observed that it often changes (and consequently hides) the tokens that are sensitive to the claim (e.g., entities, main verbs), affecting $\mathcal{V}_{\mathcal{A}}$ and indirectly later the retrieval step by $\mathcal{R}_{\mathcal{D}}$ as well.

4.1.4 Imperceptible_{Ret} 🕒 🕒 🕒

To further limit \mathcal{A} ’s capabilities, we then design a version of the imperceptible attacks that only needs a retrieval model. Ideally, if the main entities mentioned in the claim (c) are hidden in the evidence, $\mathcal{R}_{\mathcal{A}}$ (and then $\mathcal{R}_{\mathcal{D}}$) will have low scores for these sentences, i.e., the evidence would be hidden. Thus, instead of minimizing the correct label probability, we here minimize the ranking score of the evidence with respect to the claim: $x_{\mathcal{A}} = \arg \min_x \mathcal{R}_{\mathcal{A}}(x, c)$.

4.1.5 Omitting Paraphrase 🕒 🕒 🕒

As ‘imperceptible’ attacks produce indistinguishable sentences, they keep the sentences’ correctness. However, they

only hide the sentences from models while still being available to online readers. On the other hand, the ‘contextualized replace’ attack could replace the relevant snippets but might introduce syntactic errors and incorrect information, violating the full preservation of the context constraint.

To meet both goals, we propose a sentence re-writing attack based on paraphrasing or abstractive summarization. As there are usually many different ways to write a summary of a sentence, \mathcal{A} here aims to pick the sentence that omits the claim-salient snippets from the evidence. Specifically, we use an off-the-shelf paraphrasing model, based on the PEGASUS abstractive summarization model [91], to generate paraphrases for the top-retrieved evidence. This step is claim-agnostic. Next, we use $\mathcal{R}_{\mathcal{A}}$ as an adversarial filter to select the paraphrase that minimizes the retrieval ranking with respect to the claim, c : $x_{\mathcal{A}} = \arg \min_x \mathcal{R}_{\mathcal{A}}(x, c)$. The reasoning here is that paraphrases that leave out the important parts should be ranked lower by $\mathcal{R}_{\mathcal{A}}$.

This attack is highly stealthy; the re-writings are fluent as they are not based on word-level perturbations. In addition, it does not introduce false or even unrelated evidence, meeting the complete preservation of the context constraint. It also does not require \mathcal{A} to either have a verification model or fine-tune the additionally used paraphrasing model.

4.1.6 Omitting Generate 🕒 🕒 🕒

In some sentences, it might be difficult to find evidence paraphrases that omit the claim-relevant parts. Thus, we investigate another omitting variant that assumes that \mathcal{A} is not constrained by keeping the context. As we exclude deleting evidence as an attack (as discussed in section 3), we study a more subtle approximation: Given the evidence, we generate alternative evidence that should leave out the relevant parts. We fine-tune a GPT-2 model [55] to generate supporting evidence given claims (details later in section 4.2.3). Next, we use the old evidence as a generation prompt. As the model is fine-tuned to generate supporting evidence, the generated sentence should have some overlap in topics and context with the old evidence, ideally making it a plausible alternative. To exclude sentences that copied the relevant parts from the old evidence, we again pick the sentence with the lowest retrieval score by $\mathcal{R}_{\mathcal{A}}$. We show the workflow of these two omitting attacks in Figure 4.

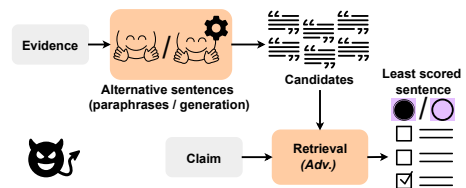


Figure 4: Omitting paraphrase and generate attacks.

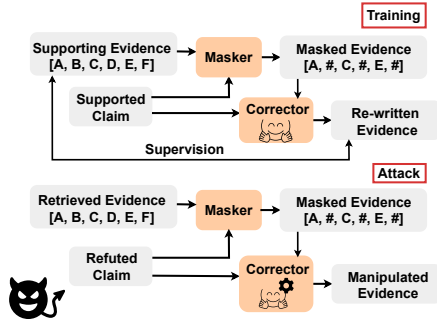


Figure 5: We design a distantly-supervised claim-aligned evidence re-writing attack inspired by the factual error correction of claims approach in [77].

4.2 Planting Attacks

Next, we discuss planting attacks that attempt to produce evidence with a supporting factual stance to the claim. All planting attacks assume that either a ‘replace’ or ‘add’ modification method can be applied.

4.2.1 Claim-aligned Re-writing **R**

To create evidence agreeing with a wrong claim, one can re-write the relevant, likely contradicting, evidence. This can partially keep the original context. Thus, compared to previous work [20], it can be stealthier than generating entirely new evidence. To perform re-writings, we ideally need training data in the format of <claims, refuting evidence, supporting re-writes>, which is unavailable. We thus use a distant supervision method. Thorne et al. [77] proposed a two-stage framework to factually correct claims such that they are better supported by the retrieved evidence. We here employ their approach while reversing the task; we edit the evidence to agree with the claim. This can be a harder generation task since the evidence sentences are usually longer.

The framework, shown in Figure 5, consists of a masker (Mask) and a corrector (Corr). First, Mask replaces claim-salient parts in the evidence (e.g., supporting or contradicting) with placeholders, yielding masked evidence s' : $s' = \text{Mask}(s)$. Second, the corrector network is trained to fill in the blanks while conditioning on the claim c : $\tilde{s} = \text{Corr}(c, s')$. As distant supervision, Corr is trained on pairs of SUP claims and their masked golden supporting evidence, and it is instructed to reconstruct the evidence: $\tilde{s} = s$. The goal here is to produce evidence that agrees with the claim. We use a masking method based on masking the top important tokens according to a BERT $\mathcal{V}_{\mathcal{A}}$ (similar to the ‘contextualized replace’ attack); we empirically found that it outperforms the LIME masker [58] used in [77]. The corrector network is a T5 encoder-decoder model [56]. Then, to run the attack at test time, the framework is applied to REF claims and the top retrieved evidence sentences by $\mathcal{R}_{\mathcal{A}}$ to convert them to supporting ones.

+Stance Filtering. To further evaluate the attack’s success

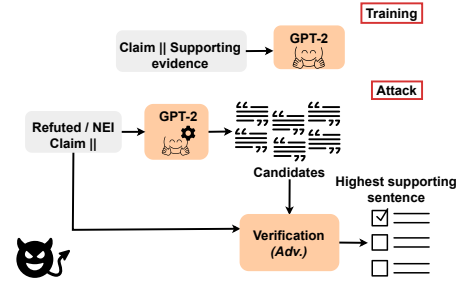


Figure 6: ‘Supporting generation’ attack.

rate, we study a variant that samples different re-writes candidates using top- k sampling [33] from the trained corrector $\{\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_n\}$ and then picks the sample that maximizes the SUP class probability o_{supp} of $\mathcal{V}_{\mathcal{A}}$: $\tilde{s}_{\mathcal{A}} = \arg \max_{\tilde{s}} o_{\text{supp}}(\tilde{s})$.

4.2.2 Claim-aligned Re-writing_{Ret} **R**

We implement a variant of the previous attack that leverages $\mathcal{R}_{\mathcal{A}}$ (instead of $\mathcal{V}_{\mathcal{A}}$) in the masking step for both the training and the attack computation. Similarly, we mask each word w_i in the evidence s and compute $\mathcal{R}_{\mathcal{A}}$ ’s score for the masked sentence s'_{w_i} w.r.t. the claim c :

$$I_{w_i} = \mathcal{R}_{\mathcal{A}}(s'_{w_i}, c)$$

Then, we rank words in ascending order of these scores and mask the top k ; the most important words should ideally cause the lowest retrieval scores when masked. The corrector model is trained the same way as in the previous attack but with the new masking output.

+Retrieval Filtering. To improve the attack, we sample different re-writes candidates from the corrector. As this attack does not assume the availability of $\mathcal{V}_{\mathcal{A}}$, the attack sentences are picked using $\mathcal{R}_{\mathcal{A}}$: $\tilde{s}_{\mathcal{A}} = \arg \max_{\tilde{s}} \mathcal{R}_{\mathcal{A}}(\tilde{s}, c)$. The sentences highly relevant to the claim are also likely to be agreeing with it since the masking should have removed the contradicting snippets and the corrector should yield supporting sentences.

4.2.3 Supporting Generation **NEI+R**

The ‘claim-aligned re-writing’ attack starts from relevant evidence; thus, it partially preserves the context. However, \mathcal{A} might seek to distribute diverse supporting sentences instead of re-writing a single one (e.g., for spamming). Additionally, in some cases, it could be hard to reverse the stance from partial re-writes, e.g., for NEI claims that do not have highly relevant evidence, making the masking required for re-writing less defined. Therefore, we study an attack based on generating new supporting evidence given the claim.

As shown in Figure 6, we first fine-tune GPT-2 to generate supporting evidence given a claim. As we do not have training pairs of <wrong claims, supporting evidence>, we use pairs of <correct claims, supporting evidence>, similar

to the previous distant supervision approach. The training sequence is: $\langle \text{claim} \rangle \parallel \langle \text{evidence} \rangle$. To run the attack at test time, we prompt the fine-tuned GPT-2 with the REF or NEI claims, followed by \parallel . We fine-tune GPT-2 instead of using it off-the-shelf for two reasons: 1) to adapt to the FEVER writing style, and 2) the evidence should entail the claim, not be a continuation of it.

+Stance Filtering. Nevertheless, text-generation models can have limited coordination between the input and output [68] (one of the reported limitations in [20]). To tackle that limitation, we sample from the fine-tuned model and take the samples maximizing the SUP probability of a BERT $\mathcal{V}_{\mathcal{A}}$ (excluding exact copies of claims), similar to the previous stance-filtering re-writing attack.

4.2.4 Claim-conditioned Article Generation NEI+R

We fit the ‘AdvAdd’ method [20] within our taxonomy. The adversary here has limited **knowledge** and **capabilities**. The attack uses the claim to conditionally generate articles using the Grover model [90] (no extra fine-tuning or filtering w.r.t. the claim), and it assumes that the article would be used to create a new Wikipedia page. We exclude the ‘AdvMod-paraphrase’ [20] because it yields unrealistic attacks (short direct reiteration of claims). We also exclude the ‘AdvMod-KeyReplace’ [20] because it is not intended to fool humans (it produces sentences that do not logically support the claim but are only superficially similar to it). It is important to note that ‘AdvMod’ attacks differ substantially from our camouflaging attacks since they do not edit the relevant evidence. Instead, they edit an article by appending new sentences that are variants of the claim itself, *not* the original evidence.

5 Evaluation

We first show the attacks’ performance. We then evaluate the attacks under different **constraints** and **knowledge** settings and post-hoc claim paraphrasing. Next, we show qualitative examples. Finally, we discuss a use-case of planting attacks against the SUP label. We show in the main paper the results on KGAT (BERT_{BASE}). In Appendix A, we outline more attacks’ implementation details. In Appendix B, we report the results on CorefBERT_{BASE}, KGAT (RoBERTa_{LARGE}), and CorefRoBERTa_{LARGE}. Code and data will be available at: <https://github.com/S-Abdelnabi/Fact-Saboteurs>.

5.1 Attacks’ Performance

We show the attacks’ performance on the KGAT (BERT_{BASE}) model in Table 2. We compute the model’s accuracy before and after the attack (lower \rightarrow more successful attack). We also measure the percentage of perturbed sentences that were retrieved by $\mathcal{R}_{\mathcal{D}}$ (‘recall’) and the ratio of predictions that

changed to NEI (‘ \rightarrow NEI’). These metrics measure how well the attacks align with the **targets**; e.g., recall is hypothesized to be higher and ‘ \rightarrow NEI’ lower for planting attacks. All planting attacks are reported using the more constrained ‘add’ modification assumption, i.e., the original evidence still exists. All attacks edit/add at most 5 sentences; this is to compute attacks’ lower bounds but the attacker can, in principle, perform more changes. We summarize our findings as follows:

1) Consistent with the **targets**, attack sentences are less likely to be recalled in the camouflaging attacks. Also, predictions mainly changed to NEI instead of the opposite polarity (i.e., the relevant evidence becomes hidden or irrelevant). The













| Attack | SUP | REF | NEI | Attack Recall | \rightarrow NEI |
|---|-------------|-------------|-------------|---------------|-------------------|
| - (baseline) | 89.0 | 71.2 | 72.4 | - | - |
| Camouflaging  | | | | | |
| Lexical variation  | 68.9 | 65.4 | - | 42.1 | 73.6 |
| Contextualized replace  | 50.7 | 59.7 | - | 30.3 | 69.3 |
| Imperceptible ($\epsilon = 5$) | | | | | |
| Homoglyph | 39.6 | 50.3 | - | 55.2 | 83.6 |
| Reorder | 37.8 | 49.5 | - | 55.1 | 81.8 |
| Delete  | 38.9 | 49.7 | - | 60.5 | 79.4 |
| Imperceptible _{Ret} | | | | | |
| Homoglyph ($\epsilon = 5$) | 62.3 | 60.5 | - | 31.5 | 88.9 |
| Homoglyph ($\epsilon = 12$)  | 25.9 | 42.6 | - | 16.5 | 90.1 |
| Omitting paraphrase  | 51.0 | 54.3 | - | 54.4 | 83.8 |
| Omitting generate  | 29.9 | 46.8 | - | 30.9 | 87.9 |
| Planting  | | | | | |
| Claim-aligned re-writes | - | 51.2 | - | 95.2 | 4.4 |
| +stance filtering  | - | 38.4 | - | 94.4 | 1.8 |
| Claim-aligned re-writes _{Ret} | - | 53.8 | - | 86.4 | 4.9 |
| +retrieval filtering  | - | 43.7 | - | 99.1 | 1.8 |
| Supporting generation | - | 61.2 | 60.5 | 70.1 | 11.4 |
| +stance filtering  | - | 42.0 | 32.2 | 85.7 | 3.8 |
| Claim-conditioned article generation [20]  | - | 42.4 | 15.5 | | |

Table 2: Accuracy before and after attacks (%), recall of perturbed evidence by $\mathcal{R}_{\mathcal{D}}$ (%), and ‘ \rightarrow NEI’ (%) (ratio of predictions that changed to NEI). The ‘Claim-conditioned article generation’ results are from [20] (‘AdvAdd-full’).

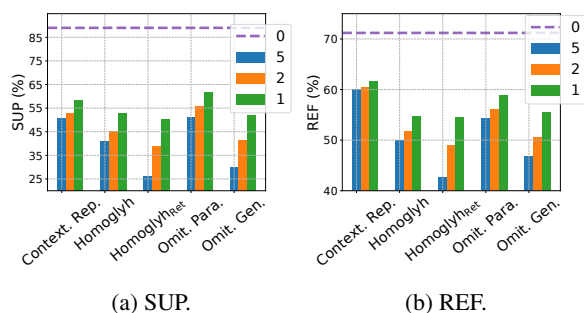


Figure 7: Camouflaging attacks when limiting the maximum changed evidence to 5, 2, or 1, vs. the ‘no attack’ baseline.

opposites are true for planting attacks.

2) For camouflaging, ‘imperceptible’ attacks are highly successful while they keep the sentences visually unchanged. The ‘omitting generate’ attack is also closely effective while, in contrast, it actually removes the information.

3) For planting attacks, candidate sampling and filtering increase the attacks’ success rate. In addition, the re-writing is as successful as generation, *outperforming the baseline* [20] for REF claims while being more context-preserving. It is also more frequently retrieved, possibly because the starting evidence is already relevant.

4) The ‘claim-conditioned article generation’ [20] is the strongest attack for NEI. However, its results are computed by adding 10 paragraphs to the repository, while the rest of our planting attacks are computed by adding 2 sentences only. Also, as reported in [20], the success rate might be overestimated as the Grover model tends to copy the claims exactly for $\sim 20\%$ of the cases. In contrast, our ‘supporting generation’ attack can produce *more plausible sentences* (more discussion and results are in section 5.3 and Appendix B).

5) For attacks with both retrieval and verification variants (‘imperceptible’ at $\epsilon = 5$ and claim-aligned re-writes), the verification one is stronger in affecting accuracy, possibly because the verification model is more precise in finding the important tokens. In contrast, the retrieval model might assign high importance to overlapping yet non-content words (e.g., ‘is’). However, increasing the top-words pool (e.g., the perturbation budget for ‘imperceptible’ attacks) can still highly increase the retrieval variant success.

Summary #1 (Targets): For both the planting and camouflaging targets, the model’s performance degrades significantly under many attacks and across all labels.

5.2 Constraints

Moreover, we investigate and discuss different **constraints**. The first is the context; Table 2 shows that attacks work well even under the restrictive context-preserving constraint. The ‘imperceptible’ attacks do not introduce any changes to the

evidence, yet, they are the most effective camouflaging attack. The ‘omitting paraphrase’ also works relatively well (compared to other perturbation attacks such as the ‘contextualized replace’) while it is fluent, stealthy, factual, and does not introduce irrelevant information.

Next, we study a setup where the adversary might be limited in the number of evidence sentences to edit/add. Figure 7 shows the camouflaging attacks when the maximum allowed edits range from 5 to 1. In each setting, the top n relevant sentences (ranked by \mathcal{R}_A) are edited. Even with 1 edited sentence, attacks can still be successful. For example, the ‘imperceptible’ attack can drop the total accuracy to 53.7%, vs. 45.0% when editing at most 5 sentences. While this can be explained by the scarcity of golden evidence per claim in FEVER, it indicates that the adversary can use the retrieval model to selectively corrupt the most important evidence without needing golden relevancy annotations.

Figure 8 shows a similar experiment for planting attacks. Here, the adversary is limited in the number of evidence sentences to *add* to the repository – *without removing* the existing golden evidence. While adding more sentences increases the attacks’ success rate, a large drop can still be achieved by adding only one (e.g., the REF accuracy dropped to 44% via evidence re-writes, and the NEI dropped to 19.5% via article generation [20]). This suggests that models are sensitive to even the slightest presence of supporting evidence to claims.

Finally, as shown in Table 3, we observed that camouflaging attacks work *only* under a ‘replace’ repository modification method⁴. In contrast, the gap in performance of the ‘claim-aligned re-writing’ attack under the ‘add’ and ‘replace’ methods is minimal, suggesting that the adversary can be nearly as successful *without* removing the existing evidence.

Summary #2 (Constraints): Attacks are still highly successful under the full-context preservation constraint and when fewer sentences are changed/added.

5.3 Knowledge

Previous experiments are performed assuming the adversary has the white-box retrieval model, $\mathcal{R}_A = \mathcal{R}_D$, and the same dataset, $\mathcal{S}_A = \mathcal{S}_D$, when training the models needed for the attack. In this section, we relax these assumptions and study different **knowledge** variations.

To evaluate a black-box setting of \mathcal{R}_D ⁵, we train the BERT retrieval model but with different random initialization. We evaluate another restricted setup where the architecture of \mathcal{R}_A and \mathcal{R}_D is different. We use the retrieval output of the

⁴This is based on our empirical evaluation of current attacks and models, rather than being an inherent property of the attack. E.g., future camouflaging attacks might be successful with partial ‘replace’ over one source or by adding evidence that gets retrieved/prioritized over the relevant evidence.

⁵A reminder: the verification model \mathcal{V}_A is never a white-box nor the same architecture as \mathcal{V}_D .

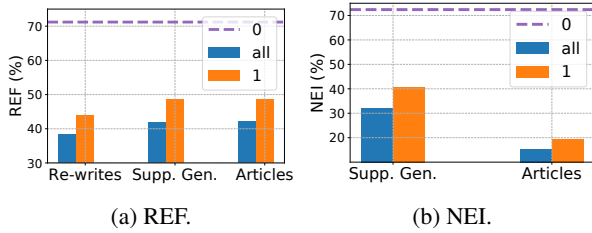


Figure 8: Planting attacks when the maximum added evidence is ‘all generated’ (2 sentences for re-writes and supporting generation and 10 paragraphs for article generation [20]) or 1 vs. the ‘no attack’ baseline. Article generation results are from [20] (‘AdvAdd-full’ and ‘AdvAdd-min’).

Enhanced Sequential Inference Model (ESIM) [14] used in previous FEVER work [49] (LSTMs with alignment model). We compare these two setups in Table 4. These attacks use \mathcal{R}_A to retrieve the relevant sentences that the attack edits (e.g., ‘imperceptible’ or ‘contextualized replace’), or to also construct the attack sentences themselves (e.g., ‘omitting paraphrase’). The white-box and black-box BERT cases have nearly the same performance. Even when using ESIM (a less powerful model), the attacks have a high success rate (e.g., for the ‘imperceptible’ attack, the accuracy dropped to 47% vs. 45% in the white-box case).

Additionally, for black-box scenarios, the adversary needs to train proxy fact-verification models (\mathcal{R}_A and \mathcal{V}_A). Also, some attacks need to fine-tune additional models for language generation (e.g., T5 or GPT-2). Thus, we show the attacks’ performance vs. the size of the dataset available to the adversary in Figure 9. The attacks are nearly as successful when having

| Attack | Method | SUP (%) | REF (%) |
|------------------------|---------|---------|---------|
| - | - | 89.0 | 71.2 |
| Imperceptible | Replace | 39.7 | 50.3 |
| | Add | 88.3 | 70.6 |
| Contextualized replace | Replace | 50.7 | 59.8 |
| | Add | 88.8 | 70.3 |
| Omitting paraphrase | Replace | 51.0 | 54.3 |
| | Add | 88.8 | 71.0 |
| Claim-aligned re-write | Replace | - | 49.2 |
| | Add | - | 51.2 |

Table 3: ‘Add’ vs. ‘Replace’ repository modification methods for a sample of camouflaging and planting attacks.

| Attack | \mathcal{R}_A | Knowledge | SUP (%) | REF (%) |
|------------------------|-----------------|-----------|---------|---------|
| Imperceptible | BERT | WB | 39.6 | 50.3 |
| | | BB | 40.6 | 49.9 |
| | ESIM | - | 43.1 | 50.9 |
| Contextualized replace | BERT | WB | 50.7 | 60.1 |
| | | BB | 50.8 | 59.8 |
| | ESIM | - | 53.1 | 60.9 |
| Omitting paraphrase | BERT | WB | 55.5 | 56.1 |
| | | BB | 54.5 | 55.8 |

Table 4: Attacks when changing the adversary’s retrieval model, \mathcal{R}_A .

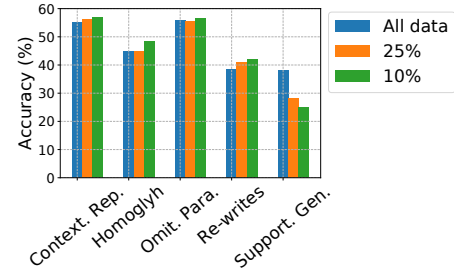


Figure 9: Attacks with different assumptions about the adversary’s dataset size; subsets are chosen randomly.

only 10% of the data (the maximum absolute difference is ~ 3.5 percentage points).

Interestingly, the attack success *increases* for the ‘supporting generation’ attack when decreasing the training data (accuracy decreased to 25.1% when fine-tuning with 10% of the data, outperforming the 28.9% by the ‘article generation’ baseline [20]). We found that models fine-tuned with more data tend to generate more diverse sentences, better matching their training data. In contrast, models fine-tuned with a small subset can have simpler sentences that more directly support the claim. On the other hand, off-the-shelf models (e.g., Grover) can often, trivially and unrealistically, copy the claims exactly [20]. For further analysis, we show histograms of claim-evidence sentence embeddings’ distances in Figure 11; not only is the 10% ‘supporting generation’ more successful than the baseline [20], but it can also achieve a *better trade-off between the attack’s success and its plausibility*.

Summary #3 (Knowledge): Attacks do not need white-box access to the victim model and can be (even more) successful with only 10% of the data.

5.4 Robustness to Post-Hoc Claim Edits

So far, the claims used to construct the attacks are also the ones used in the final evaluation of the victim model. However, adversaries may not have full control over the propagation and digestion of claims and thus over the phrasings used in verification. Therefore, attacks need to not overfit (for both the retrieval and verification steps) the claim-phrasings used in construction. To test this, we created paraphrases of claims and tested them against the *already-computed* attack sentences by repeating \mathcal{D} ’s retrieval and verification given the new claims (step 2 in Figure 3). To create paraphrases, we use the PEGASUS model used in the ‘omitting paraphrase’ attack. To ensure that paraphrases are semantically equivalent, we draw different samples and take the one with the highest retrieval score to the original claim that also contains all of its named entities [69] (e.g., to exclude sentences that might replace a person’s name with a pronoun). We discard examples where only exact matches were found or not all named entities exist. Next, we test the paraphrases on the downstream model

\mathcal{M} with no attacks. We use the examples that retained the same prediction for further analysis (70% of the data, after all exclusions). These measures are to ensure that the drop in performance can be attributed to the attacks, not because the new claims are semantically different. This is also important since previous work [78] has shown that models are sensitive to claim phrasing patterns. New claims might include syntactic and lexical changes or double negation (Table 9).

Table 5 shows that the attacks’ performance on the original and paraphrased claims are comparable. The attacks also consistently achieve the corresponding targets (indicated by the ‘→ NEI’ ratio) instead of performing random changes. This experiment also suggests that potential defenses based on claim paraphrasing might not be effective.

Summary #4: Attacks work well even after post-hoc modifications and paraphrasing of the claims.

5.5 Qualitative Analysis

Examples of the attacks are in Table 10 (Appendix B). We summarize the main qualitative observations as follows:

- 1) As expected, the ‘lexical variation’ had lower quality than the ‘contextualized replace’. However, the latter still had syntactic mistakes, such as breaking the sentence with commas or dots to remove the important parts.
- 2) The ‘imperceptible’ attacks (both the verification and retrieval variants) change the relatively salient words. The retrieval variant usually changes words overlapping with the claim (e.g., the main subject, but even less crucial words such as prepositions). In contrast, the verification variant might focus on the entailment (even non-overlapping words). This explains why the retrieval variants affect the retrieval of perturbed sentences more (Table 2). It also implies that attackers might have an incentive to use them if they want to hide the sentences from users as well.
- 3) The ‘omitting paraphrase’ attack has very high quality and is also factual. However, it fails if all samples contain

| Attack | Claims | SUP | REF | NEI | → NEI |
|---|--------|------|------|------|-------|
| - | o/p | 92.2 | 71.2 | 75.4 | - |
| Imperceptible | o | 41.1 | 51.6 | - | 84.9 |
| | p | 44.8 | 46.6 | - | 87.1 |
| Imperceptible _{Ret} | o | 27.0 | 44.5 | - | 93.3 |
| | p | 27.6 | 38.3 | - | 93.5 |
| Omitting paraphrase | o | 54.9 | 56.7 | - | 88.8 |
| | p | 61.6 | 53.7 | - | 89.1 |
| Claim-aligned re-writing | o | - | 40.5 | - | 1.4 |
| | p | - | 39.1 | - | 2.1 |
| Claim-aligned re-writing _{Ret} | o | - | 45.0 | - | 1.5 |
| | p | - | 42.8 | - | 1.8 |
| Supporting generation | o | - | 44.1 | 33.8 | 1.7 |
| | p | - | 41.3 | 32.9 | 2.2 |

Table 5: Attacks optimized with the original claims (o) and tested afterwards on paraphrased claims (p).

the claim-relevant part. Increasing the candidate pool size or training omitting models might increase the attack’s success.

4) The ‘omitting generate’ attack can drastically decrease the performance. However, it might lead to limited coherency between the original evidence and the new one, which might affect the overall context.

5) The ‘claim-aligned re-writing’ attack can work even if there is no exact word-level or a short span overlap with the claim. It can also partially keep the context, depending on the original evidence’s length (we mask the top 13 tokens).

6) As discussed in section 5.3, fine-tuning GPT-2 (in the ‘supporting generation’ attack) can produce more elaborate evidence compared to using off-the-shelf models like Grover [20]. However, as similarly observed in [20], the ‘supporting generation’ may incorrectly respond to claims with negation [35] and end up producing refuting evidence.

5.6 Planting Attacks on Correct Claims

As reported in [20], generating refuting evidence to correct claims has many challenges. One of them is automatically creating meaningful counterclaims. However, adversaries can circumvent that by manually writing counterclaims, then automatically generating the evidence [20].

To test that, we manually crafted counterclaims for 150 SUP claims. Our employed strategies were to use negations,

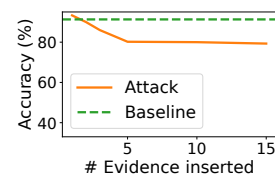


Figure 10

| Evidence | SUP | → NEI |
|-----------|------|-------|
| Baseline | 91.3 | - |
| No golden | 42.0 | 86.4 |
| + Planted | 52.6 | 38.4 |

Table 6

Figure 10: Planting attacks with ‘add’ modification against SUP examples subset. Table 6: Performance (%) with original evidence, removing golden evidence, and adding the generated evidence (without the golden).

| |
|--|
| Claim: Fox 2000 Pictures released the film Soul Food. |
| Counterclaim: Columbia Pictures released the film Soul Food. |
| Original: <u>Soul Food</u> is a 1997 American comedy drama film produced by Kenneth ‘Babyface’ Edmonds, Tracey Edmonds and Robert Teitel and released by Fox 2000 Pictures. |
| Planted 🗑️: <u>Columbia Pictures released Soul Food</u> on December 12, 2012, as the second film in the Jim Henson Company film Picture Show. |
| Planted 🗑️: <u>Columbia Pictures released Soul Food</u> on December 4, 2009, as a pre-quel to the 2009 film The Divergent Series. |
| Planted 🗑️: <u>Columbia Pictures released Soul Food</u> on November 30, 2004 as the second North American release on VHS, but later discontinued production. |
| Original prediction: SUP (0.96) |
| After-attack prediction: SUP (0.86) |

Table 7: Examples of attacks against correct claims. The planted counter-evidence is added to the original. These sentences were among the top-5 retrieval output.

oppositions, and replacing with a similar entity for both mutually exclusive and possibly coexistent events, whenever it would fit (Table 11). We then used the counterclaims to generate supporting evidence (i.e., should ideally counter the original claim) via the fine-tuned GPT-2 model. Next, we add the planted evidence to the existing one and re-test against the original claim.

Figure 10 shows the attack’s results. Contrary to [20], where the accuracy always *increased* after the attack, we show that it is possible to decrease it by adding more sentences (capped at 5 after retrieval). Further, Table 6 shows the accuracy when removing the golden evidence and then adding the generated one. The ‘→ NEI’ ratio decreased after the addition, showing that the generated sentences can have, to some extent, the required polarity. Nonetheless, the attack has limited success, partially because counterclaims with negations could end up with evidence agreeing with the original claims, in addition to counterclaims with non-contradicting replacements [20] (Table 12). However, in many cases, *even when the generated evidence logically refutes the original claim, the model retained its predictions* (see Table 7 and Table 13), revealing a critical limitation we discuss next.

Summary #5: We achieve more success in SUP to REF inversion, revealing other potential limitations.

6 Discussion

We here discuss the limitations and implications of our work, models’ limitations, and the potential directions that we deem promising to robustify fact-verification models.

6.1 Limitations.

Human-in-the-loop. We envision that fact-checking models might be more commonly used in the future as assistive solutions to fact-checkers [8] or ultimately to end-users [83] to, e.g., output warnings. In both cases, we believe our attacks might affect humans by misleading them or denying the service. An important follow-up that we leave for future work is to evaluate the attacks by measuring such effects. This can be methodologically complex as it involves studying how to: perform these manipulations realistically and ethically, choose topics, measure the attacks’ success via measuring users’ perception [9], and control for users’ knowledge [53] and experience (e.g., fact-checkers vs. users).

Beyond FEVER. FEVER allowed large-scale experiments and training. While we opted for a more comprehensive evaluation of the threat model, Du et al. show success on smaller datasets [20]. However, it remains unknown and ought to be evaluated how our attacks perform on other datasets with possibly different topics and characteristics.

Wikipedia as a (Relatively) Credible Source. While Wikipedia can be publicly edited, it is subject to administra-

tion to remove factually wrong or biased content [87, 57]. This gives it a relative consensus of credibility compared to other online sources and makes it highly read [59], even among fact-checkers [88]. Adversaries can exploit this wide trust to pass in disinformation or wipe traces of facts. While the Wikipedia community tirelessly resists disinformation [74], this is not free of flaws (e.g., the Croatian Wikipedia incident [17]). We hypothesize that some of our attacks (e.g., the context-preserving ones) can be stealthy even under administration. However, it can be complex to measure their potency and detection resistance without actual edits.

Beyond Single-Source Datasets. We work on Wikipedia to conform to current large-scale benchmarks; sizable datasets that match real-world fact-checking are still lacking [24]. In practice, fact-checkers usually rely on many sources [88]. Our attacks can, in principle, be applied to other sources [38]; however, some might not be publicly available or easy to tamper with. While our work lays the foundation for attacks in this domain, Wikipedia manipulations may affect only one of these sources, reducing the practical effect of these attacks on the whole manual fact-verification process. On the other hand, bridging this discrepancy between practitioners and automated fact-checking frameworks regarding the considered sources is one of our main takeaways that we discuss next.

6.2 Implications

Ethical Considerations. We emphasize that most of the studied attacks are based on already publicly available models, and some do not need any extra fine-tuning. Moreover, we work on a dataset containing claims that are generally not designed to be sensitive in nature, limiting any potential abuse.

"Only Finding Waldo": Models’ Limitations. Planting attacks can considerably succeed even when as low as one evidence sentence is inserted and with the presence of the original evidence. They also succeed on instances where the model is originally highly confident. This could be partially attributed to the sparsity of golden evidence for many claims in FEVER. However, our observations on generating refuting evidence to correct claims might indicate another underlying problem. In many cases, even when the generated evidence *logically refutes* the claim and the retrieved refuting evidence *outnumbers* the supporting one, the model did not flip its prediction to REF (Table 7). At first, this can be considered a sign of robustness. However, it is possibly the exact reason why it is easy to flip the prediction of wrong claims to SUP; models might be looking for *any agreement* with the evidence without considering counter stances. This is a plausible explanation, given that models were not trained with an evidence contradiction setup. However, it hints at a potential limitation in models’ inference that should be investigated since *the fact-checking process in practice inherently entails weighing different stances*.

Beyond Fact-Preserving Attacks. While we employ at-

tacks that target current AI vulnerabilities (e.g., imperceptible perturbations), we *choose and argue for a broader scope of our work* that goes beyond adversarial examples in the sense of imperceptible perturbations and semantic equivalence. Instead, we broadly study how AI can be leveraged to create targeted disinformation and deceptive evidence manipulations at scale, impacting models and potentially humans as well. In such a human-centric task, simulating human-created manipulations becomes the holy grail of the attacks. These semantically driven attacks can also be more pervasive across models (see Appendix B), potentially motivating their adoption by adversaries. *Even under such semantic changes, we argue that models do not show the intended behavior.* Fact-checkers do not base their verdict on a single piece of evidence, nor do they blindly trust the evidence’s plausibility [66]. Thus, future work should bridge this discrepancy and design ‘adversary-aware’ defenses that better align with these practices and exploit the persisting attacks’ limitations. We further explain these ideas in what follows.

6.3 How to Robustify Fact-Checking Models?

Diversifying Evidence Sources. Current camouflaging attacks need to replace the original evidence with the manipulated one; otherwise, they generally fail. Thus, in practice, fact-checking models should rely on diverse and independent sources, while also considering cross-platform coordination [19], to reduce the likelihood of being manipulated by a single adversarial campaign. Other evidence metadata, such as its source, should be included in the model’s design [54] to capture features such as the source’s credibility, biases, or polarity. This resembles the ‘two-source’ and ‘source triangulation’ rules of verification in journalism [66, 1].

Detecting Perturbations. In addition, some camouflaging attacks can leave artifacts or perturbations enabling their identification (e.g., NLP adversarial attacks). While it might not be possible to easily recover the original evidence, human-in-the-loop systems might issue warnings about potential manipulations. The effectiveness of such warnings should also be studied [36]. On the other hand, imperceptible perturbation attacks might allow recovering the evidence upon detection via leveraging, e.g., an OCR [11], or utilizing recent language models that render text as images [60].

Circular Verification against Planting Attacks. As discussed, models should represent opposing stances among the evidence. A possible inspectable solution would be to cluster the evidence with respect to its stance [63]. Moving beyond that, models should ideally contrast these opposing evidence given factors such as their source, plausibility, commonsense reasoning [35], and inter and intra-consistency.

Language models may have limited factuality even in response to factual claims [40]. *These limitations of attacks are, in fact, defense opportunities for detection based on high-level semantics.* While they all may agree with the claim, different

generated samples might be inconsistent or contradicting in other details (see Table 7). Similarly, a single sample might contain possibly incorrect information, beyond what supports the claim. This can fuel detection defenses of what can be called ‘circular fact-checking’: Given the evidence, extract and verify follow-up claims and reach a plausibility decision via aggregation. This, again, echos the circular nature of information gathering and verification in investigative journalism [66], lateral reading as a fact-checkers’ practice [88], in addition to the veracity verification of manipulated media guidelines [1].

A Need for Practical Datasets. As discussed in section 6.1, FEVER may be limited in matching the practical challenges of fact-verification. Therefore, there is a need to develop other datasets or augment FEVER with synthetic evidence to allow the development of models that adhere to the best practices of fact-verification. Claims should have multiple confirming or denying evidence pieces, and the evidence repositories could be partially contaminated to simulate potential adversaries. Our attacks could promisingly contribute to constructing such datasets, similar to the line of work that constructs synthetic data to help detect real fake news [31, 44].

Other Attacks. The results in section 5.3 show that there could be a trade-off between the generated evidence complexity and the attacks’ success rate. Our approach of fine-tuning and sampling can have higher success while better imitating the dataset’s distribution. Other possible approaches would be to enforce the entailment between the claim and the generated evidence by training a Sequence-to-Sequence model with a verification model [10]. Moreover, future work might study the effects of different prompts when generating evidence; e.g., is it possible to affect the stance of the evidence with biased variations (e.g., subtle linguistic cues [51]) of the claim? Finally, our work is not meant to be an exhaustive evaluation of all possible attacks, as such an evaluation might be intractable and can only grow with the improvements in language generation and understanding [13].

7 Conclusion

We propose a taxonomy to comprehensively study evidence and information manipulation attacks against fact-verification models. Inspired by real-life incidents of Wikipedia edits, we set the attacks’ semantic targets to evidence camouflaging and planting. We then design technical methods that adversaries could utilize to achieve those targets, given the taxonomy’s dimensions. Compared to previous work, we propose an extensive range of stealthier, more context-preserving, and more plausible attacks, all while simultaneously achieving higher or similar success rates and extending the attacks to all labels. We show that adversaries can decrease the performance of models even under restrictive threat models. We highlight the limitations of models’ inference and discuss possible defenses by drawing insights from fact-verification in journalism.

Acknowledgements

This work was partially funded by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

- [1] *A course by Reuters News Agency: Identifying and Tackling Manipulated Media*. [\[Link\]](#).
- [2] *Abstracts written by ChatGPT fool scientists*. [\[Link\]](#).
- [3] *AI-Text Classifier*. [\[Link\]](#).
- [4] Hunt Allcott and Matthew Gentzkow. “Social media and fake news in the 2016 election”. In: *Journal of economic perspectives* 31.2 (2017), pp. 211–36.
- [5] Moustafa Alzantot et al. “Generating Natural Language Adversarial Examples”. In: *EMNLP*. 2018.
- [6] Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. “Generating Label Cohesive and Well-Formed Adversarial Claims”. In: *EMNLP*. 2020.
- [7] Pepa Atanasova et al. “Generating Fact Checking Explanations”. In: *ACL*. 2020.
- [8] *Automated Fact-checking*. [\[Link\]](#).
- [9] Mahmoudreza Babaei et al. “Analyzing biases in perception of truth in news stories and their implications for fact checking”. In: *IEEE Transactions on Computational Social Systems* 9.3 (2021), pp. 839–850.
- [10] Eugene Bagdasaryan and Vitaly Shmatikov. “Spinning Language Models: Risks of Propaganda-as-a-Service and Countermeasures”. In: *S&P*. 2022.
- [11] Nicholas Boucher et al. “Bad Characters: Imperceptible NLP Attacks”. In: *S&P*. 2022.
- [12] Tom Brown et al. “Language models are few-shot learners”. In: *NeurIPS*. 2020.
- [13] *ChatGPT*. [\[Link\]](#).
- [14] Qian Chen et al. “Enhanced LSTM for Natural Language Inference”. In: *ACL*. 2017.
- [15] Elizabeth Clark et al. “All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text”. In: *ACL | IJCNLP*. 2021.
- [16] *Coronavirus: The human cost of virus misinformation*. [\[Link\]](#).
- [17] *Croatian Wikipedia Disinformation Assessment-2021*. [\[Link\]](#).
- [18] *Disinformation for Hire, a Shadow Industry, Is Quietly Booming*. [\[Link\]](#).
- [19] Joan Donovan and Brian Friedberg. *Source hacking: Media manipulation in practice*. Data & Society Research Institute, 2019.
- [20] Yibing Du, Antoine Bosselut, and Christopher D. Manning. “Synthetic Disinformation Attacks on Automated Fact Verification Systems”. In: *AAAI*. 2022.
- [21] *Evidence Collages*. [\[Link\]](#).
- [22] Hany Farid. “Creating, Using, Misusing, and Detecting Deep Fakes”. In: *Journal of Online Trust and Safety* 1.4 (2022).
- [23] *Full Fact AI*. [\[Link\]](#).
- [24] Max Glockner, Yufang Hou, and Iryna Gurevych. “Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation”. In: *EMNLP*. 2022.
- [25] Michael Golebiewski and Danah Boyd. *Data voids: Where missing data can easily be exploited*. Data & Society Research Institute, 2019.
- [26] Lucas Graves. “Understanding the promise and limits of automated fact-checking”. In: *Reuters Institute for the Study of Journalism* (2018).
- [27] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. “A survey on automated fact-checking”. In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 178–206.
- [28] Naemul Hassan et al. “The quest to automate fact-checking”. In: *computation+ journalism symposium*. 2015.
- [29] Naemul Hassan et al. “Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster”. In: *KDD*. 2017.
- [30] Christopher Hidey et al. “DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking”. In: *ACL*. 2020.
- [31] Kung-Hsiang Huang et al. “Faking Fake News for Real Fake News Detection: Propaganda-loaded Training Data Generation”. In: *arXiv* (2022).
- [32] *In Argentina, fact-checkers latest hire is a bot*. [\[Link\]](#).
- [33] Daphne Ippolito et al. “Automatic Detection of Generated Text is Easiest when Humans are Fooled”. In: *ACL*. 2020.
- [34] *Is the future of fact-checking automated?* [\[Link\]](#).
- [35] Liwei Jiang et al. “‘I’m Not Mad’: Commonsense Implications of Negation and Contradiction”. In: *NAACL-HLT*. 2021.
- [36] Ben Kaiser et al. “Adapting security warnings to counter online disinformation”. In: *USENIX Security*. 2021.
- [37] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL-HLT*. 2019.
- [38] Peaks M Krafft and Joan Donovan. “Disinformation by design: The use of evidence collages and platform filtering in a media manipulation campaign”. In: *Political Communication* 37.2 (2020), pp. 194–214.
- [39] Sarah Kreps, R Miles McCain, and Miles Brundage. “All the news that’s fit to fabricate: AI-generated text as a tool of media misinformation”. In: *Journal of Experimental Political Science* 9.1 (2022), pp. 104–117.
- [40] Nayeon Lee et al. “Factuality Enhanced Language Models for Open-Ended Text Generation”. In: *NeurIPS*. 2022.
- [41] Linyang Li et al. “BERT-ATTACK: Adversarial Attack Against BERT Using BERT”. In: *EMNLP*. 2020.
- [42] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv* (2019).
- [43] Zhenghao Liu et al. “Fine-grained Fact Verification with Kernel Graph Attention Network”. In: *ACL*. 2020.
- [44] Grace Luo, Trevor Darrell, and Anna Rohrbach. “NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media”. In: *EMNLP*. 2021.
- [45] Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. “Paraphrasing revisited with neural machine translation”. In: *EACL*. 2017.
- [46] *Meta’s Third-Party Fact-Checking Program*. [\[Link\]](#).
- [47] Nikola Mrkšić et al. “Counter-fitting Word Vectors to Linguistic Constraints”. In: *NAACL-HLT*. 2016.
- [48] Thanh Tam Nguyen et al. “Factcatch: Incremental pay-as-you-go fact checking with minimal user effort”. In: *ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020.

- [49] Yixin Nie, Haonan Chen, and Mohit Bansal. “Combining fact extraction and verification with neural semantic matching networks”. In: *AAAI*. 2019.
- [50] Ankur Parikh et al. “A Decomposable Attention Model for Natural Language Inference”. In: *EMNLP*. 2016.
- [51] Roma Patel and Ellie Pavlick. “Was it “stated” or was it “claimed”? : How linguistic bias affects generative language models”. In: *EMNLP*. 2021.
- [52] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *EMNLP*. 2014.
- [53] Gordon Pennycook, Tyrone D Cannon, and David G Rand. “Prior exposure increases perceived accuracy of fake news.” In: *Journal of experimental psychology: general* 147.12 (2018), p. 1865.
- [54] Kashyap Popat et al. “DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning”. In: *EMNLP*. 2018.
- [55] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* (2019).
- [56] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer.” In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67.
- [57] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. “Linguistic models for analyzing and detecting biased language”. In: *ACL*. 2013.
- [58] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *NAACL*. 2016.
- [59] Roy Rosenzweig. “Can history be open source? Wikipedia and the future of the past”. In: *The journal of American history* 93.1 (2006), pp. 117–146.
- [60] Phillip Rust et al. “Language Modelling with Pixels”. In: *arXiv* (2022).
- [61] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. “COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic”. In: *ACL | IJCNLP*. 2021.
- [62] Mohammad Hammas Saeed et al. “TROLLMAGNIFIER: Detecting state-sponsored troll accounts on reddit”. In: *S&P*. 2022.
- [63] Tal Schuster et al. “Stretching Sentence-pair NLI Models to Reason over Long Documents and Clusters”. In: *Findings of EMNLP*. 2022.
- [64] Tal Schuster et al. “Towards Debiasing Fact Verification Models”. In: *EMNLP-IJCNLP*. 2019.
- [65] Darsh Shah, Tal Schuster, and Regina Barzilay. “Automatic fact-guided sentence modification”. In: *AAAI*. 2020.
- [66] Ivor Shapiro et al. “Verification as a strategic ritual: How journalists retrospectively describe processes for ensuring accuracy”. In: *Journalism Practice* 7.6 (2013), pp. 657–673.
- [67] Rakshith Shetty, Bernt Schiele, and Mario Fritz. “A4NT: Author Attribute Anonymity by Adversarial Training of Neural Machine Translation”. In: *USENIX Security*. 2018.
- [68] Kai Shu et al. “Fact-enhanced synthetic news generation”. In: *AAAI*. 2021.
- [69] *SpaCy*. [\[Link\]](#).
- [70] Rainer Storn and Kenneth Price. “Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces”. In: *Journal of global optimization* 11.4 (1997), pp. 341–359.
- [71] *Tackling online disinformation*. [\[Link\]](#).
- [72] *TECH & CHECK*. [\[Link\]](#).
- [73] *The Covert World Of People Trying To Edit Wikipedia—For Pay*. [\[Link\]](#).
- [74] *There’s a lot Wikipedia can teach us about fighting disinformation*. [\[Link\]](#).
- [75] *This Washington Post fact check was chosen by a bot*. [\[Link\]](#).
- [76] James Thorne and Andreas Vlachos. “Automated Fact Checking: Task Formulations, Methods and Future Directions”. In: *COLING*. 2018.
- [77] James Thorne and Andreas Vlachos. “Evidence-based Factual Error Correction”. In: *ACL | IJCNLP*. 2021.
- [78] James Thorne et al. “Evaluating adversarial attacks against multiple fact verification systems.” In: *EMNLP/IJCNLP*. 2019.
- [79] James Thorne et al. “FEVER: a Large-scale Dataset for Fact Extraction and VERification”. In: *NAACL-HLT*. 2018.
- [80] *Twitter account highlights history of Ottawa staffers making Wiki edits*. [\[Link\]](#).
- [81] *Twitter finally turns to the experts on fact-checking*. [\[Link\]](#).
- [82] *Vandalism on Wikipedia*. [\[Link\]](#).
- [83] Nguyen Vo and Kyumin Lee. “Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News”. In: *EMNLP*. 2020.
- [84] David Wadden et al. “Fact or Fiction: Verifying Scientific Claims”. In: *EMNLP*. 2020.
- [85] *Wikipedia edits from inside Parliament removing scandals from MPs’ pages, investigation finds*. [\[Link\]](#).
- [86] *Wikipedia:List of hoaxes on Wikipedia*. [\[Link\]](#).
- [87] *Wikipedia:Wikipedia is not a forum*. [\[Link\]](#).
- [88] Sam Wineburg and Sarah McGrew. “Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information”. In: *Teachers College Record* 121.11 (2019), pp. 1–40.
- [89] Deming Ye et al. “Coreferential Reasoning Learning for Language Representation”. In: *EMNLP*. 2020.
- [90] Rowan Zellers et al. “Defending against neural fake news”. In: *NeurIPS*. 2019.
- [91] Jingqing Zhang et al. “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization”. In: *ICML*. 2020.
- [92] Jie Zhou et al. “GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification”. In: *ACL*. 2019.
- [93] Mary Ellen Zurko. “Disinformation and Reflections From Usable Security”. In: *IEEE Security & Privacy* 20.3 (2022), pp. 4–7.

A Implementation Details

To train the attack model \mathcal{V}_a , we fine-tune BERT_{BASE} or RoBERTa_{BASE} models for 4 epochs on pairs of claims and golden evidence (for SUP and REF claims). For NEL, we pick the top 3 retrieved sentences for each claim (these should be more challenging than taking random sentences).

To run the ‘lexical variation’ attack, we follow the authors’ `code` and distances’ hyperparameters but change the target model to RoBERTa. Words to replace are randomly sampled with probabilities proportional to the number of neighbors each word has in the embedding space [5]. We adapt the ‘contextualized replace’ `code` to the entailment task. We perturb at most 15% of tokens in the sentence and set a probability threshold of 1.0e-5 on the BERT MLM candidates. We allow sub-words substitutes. We use an embedding distance of 0.4

between the counter-fitted vectors [47]. For the ‘imperceptible’ attacks, we use the untargeted versions of the attack with a maximum of only 3 iterations for the genetic algorithm (vs. 10 in the original paper). Increasing the iterations’ number may lead to even higher success rates; however, these attacks are expensive to run on the whole dataset (> 90,000 claim-evidence pairs). For attacks against $\mathcal{V}_{\mathcal{A}}$, we run the attacks only on the pairs (among the top-5) where $\mathcal{V}_{\mathcal{A}}$ ’s predictions are initially correct. Since the attacks do not assume golden relevancy annotations, the labels of all retrieved sentences are set to the original claims’ label (i.e., SUP or REF). We run the ‘imperceptible_{Ret}’ on all the top-5 retrieved evidence to minimize the retrieval score.

For ‘omitting paraphrase’, we use the PEGASUS model fine-tuned for paraphrasing. For each sentence, we generate 20 candidates using beam search (then select the lowest retrieval candidate). The GPT-2 model used in ‘omitting generate’ and later in the ‘supporting generation’ is trained on pairs of claims and supporting evidence for 20 epochs with a batch size of 4 and a learning rate of 0.00003. For both attacks, we use top-*k* sampling. For ‘omitting generate’, we also generate 20 candidates (then select the lowest retrieval candidate). For ‘supporting generation’, we select the top 2 sentences from 160 samples (increasing the samples’ number helped to have better attacks).

For the ‘claim-aligned re-writing’ attack, we adapt [77]’s code to re-write evidence instead of claims. We use the BERT-score masker explained in the main paper. During training, we mask the top 16 tokens. We train the T5 model for 12 epochs with a batch size of 4 and a learning rate of 0.0001. To run the attack, we mask the top 13 tokens. Depending on the masking, the T5 model re-writes single masked words or a whole span. Since this attack ideally assumes that the starting evidence is relevant, we run it only on the top 2 relevant evidence sentences. In the sampling and filtering variants, we generate 60 candidates using top-*k* sampling. Finally, due to time and computation resources’ constraints and the scale of the experimental evaluation, it is difficult to perform an exhaustive hyperparameter search. Further tuning of the hyperparameters can lead to higher success rates; our results are a lower bound.

B Other Results and Examples

In Table 8, we show the attacks’ performance (without any further adaptation) on CorefBERT_{BASE}, KGAT (RoBERTa_{LARGE}), and CorefRoBERTa_{LARGE}. Most of the attacks are still effective across models. As ‘imperceptible’ attacks depend on the model’s vocabulary, their performance can slightly degrade when transferred from BERT to RoBERTa. Increasing the perturbation budget can yield similar performance. Attacks that are based on semantically removing or adding information needed for verification are consistent across models.


Table 9 shows examples of the automatically-created claim paraphrases (section 5.4). Table 10 shows qualitative examples (section 5.5). Tables 11, 12, and 13 show more examples of planting attacks against the SUP label (section 5.6). Finally, Figure 11 shows histograms of sentence embeddings’ distances between claims and evidence, for both golden and generated evidence (section 5.3). Our attack can lead to better matching of the golden evidence distribution compared to the baseline [20].

| Attack | SUP (%) | | | REF (%) | | | NEI (%) | | |
|--|---------|------|------|---------|------|------|---------|------|------|
| | #1 | #2 | #3 | #1 | #2 | #3 | #1 | #2 | #3 |
| - (baseline) | 87.5 | 91.5 | 92.2 | 72.8 | 74.7 | 77.5 | 72.8 | 68.8 | 70.0 |
| Camouflaging 🌀 🚫 | | | | | | | | | |
| Lexical variation 🔍 🗑️ 📄 | 67.7 | 73.6 | 74.5 | 66.6 | 69.9 | 71.6 | - | - | - |
| Contextualized replace 🔍 🗑️ 📄 | 50.0 | 55.8 | 55.6 | 60.8 | 63.9 | 64.0 | - | - | - |
| Imperceptible | | | | | | | | | |
| Homoglyph (ε = 5) | 39.9 | 60.6 | 65.1 | 52.5 | 60.1 | 60.7 | - | - | - |
| Homoglyph (ε = 12) | 33.6 | 49.7 | 52.0 | 47.9 | 54.7 | 52.4 | - | - | - |
| Reorder (ε = 5) | 37.4 | 47.7 | 49.9 | 52.3 | 54.8 | 51.4 | - | - | - |
| Reorder (ε = 12) | 32.4 | 36.8 | 34.9 | 48.0 | 49.3 | 42.7 | - | - | - |
| Delete (ε = 5) 🔍 🗑️ 📄 | 39.2 | 60.8 | 66.1 | 52.5 | 60.7 | 59.8 | - | - | - |
| Imperceptible _{Ret} | | | | | | | | | |
| Homoglyph (ε = 12) 🔍 🗑️ 📄 | 26.6 | 36.4 | 37.0 | 44.8 | 50.3 | 45.6 | - | - | - |
| Omitting paraphrase 🔍 🗑️ 📄 | 50.7 | 56.8 | 55.5 | 55.8 | 60.7 | 58.4 | - | - | - |
| Omitting generate 🔍 🗑️ 📄 | 30.2 | 33.8 | 31.6 | 48.9 | 51.9 | 47.4 | - | - | - |
| Planting 🗑️ 📄 | | | | | | | | | |
| Claim-aligned re-writes 🔍 🗑️ 📄 | - | - | - | 36.9 | 44.9 | 42.1 | - | - | - |
| Claim-aligned re-writes _{Ret} 🔍 🗑️ 📄 | - | - | - | 43.1 | 48.8 | 47.6 | - | - | - |
| Supporting generation 🔍 🗑️ 📄 | - | - | - | 39.7 | 43.2 | 45.2 | 34.5 | 25.6 | 30.0 |

Table 8: Attacks on CorefBERT_{BASE} (#1), KGAT (RoBERTa_{LARGE}) (#2), and CorefRoBERTa_{LARGE} (#3).

| Original Claim | Paraphrase |
|--|--|
| Tilda Swinton is a vegan. | There is a person named Tilda Swinton who is a vegan. |
| Murda Beatz’s real name is Marshall Mathers. | Marshall Mathers is Murda Beatz’s real name. |
| Hourglass is performed by a Russian singer-songwriter. | Hourglass is a song by a Russian singer-songwriter. |
| Fox 2000 Pictures released the film Soul Food. | The film Soul Food was released by Fox 2000 Pictures. |
| Charles Manson has been proven innocent of all crimes. | Charles Manson has not been proven guilty of any crimes. |

Table 9: Automatically created claim paraphrases.

Lexical Variation 

Claim: Ann Richards was professionally involved in politics (**Label:** SUP).

Original: Richards was the second **female** governor of Texas, and was frequently noted **in** the media **for** her outspoken feminism and her **one** liners.

Edited: Richards was the second **daughters** governors **du** Texas, and became frequently noted **for** the media **in** her outspoken feminism and her **eden** liners.

Contextualized Replace 

Claim: James VI and I was a major advocate of a single parliament for Scotland and England (**Label:** SUP).


Original: He was a **major advocate** of a single **parliament** for **England** and **Scotland**.

Edited: He was a **broad activist** of a single **legislature** for **Britain** and **Ireland**.


Claim: Ernest Medina participated in the My Lai Massacre (**Label:** SUP).


Original: He was the commanding officer of Company C, ... , the unit responsible for the My **Lai** Massacre ...


Edited: He was the commanding officer of company C, ..., the unit responsible for the My **,** Massacre ...

Imperceptible/Imperceptible_{ret} 

Claim: Nicholas Brody is a character on Homeland (**Label:** SUP).

Edited : Nicholas 'Nick' Brody, played by actor Damian Lewis, is a fictional **character** on the **American television** series **Homeland** on Showtime.

Edited : **Nicholas 'Nick' Brody**, played by actor Damian Lewis, is a fictional character on the American television series **Homeland on** Showtime.

Omitting Paraphrase 

Claim: Murda Beatz's real name is Marshall Mathers. (**Label:** REF).


Original: Shane Lee Lindstrom (born February 11, 1994) , professionally known as Murda Beatz, is a Canadian hip hop record producer from Fort Erie, Ontario.

Edited: Murda Beatz is a hip hop record producer from Fort Erie, Ontario.

Claim: Fox 2000 Pictures released the film Soul Food. (**Label:** SUP).

Original: Soul Food is a 1997 American comedy drama film produced by Kenneth 'Babyface' Edmonds, Tracey Edmonds and Robert Teitel and released by Fox 2000 Pictures.


Edited: The 1997 American comedy drama film Soul Food was produced by Kenneth 'Babyface' Edmonds, and was released by Fox 2000 Pictures.

Omitting Generate 

Claim: Damon Albarn's debut album was released in 2011 (**Label:** REF).

Original: Raised in Leytonstone , East London and around Colchester , Essex , Albarn attended the Stanway School , where he met Graham Coxon and eventually formed Blur , whose debut album Leisure was released in 1991 to mixed reviews.

Edited: Born in Leytonstone, east London, his first exposure to music came in 1991 at the age of seven, when he was discovered by Dr. Paul Barbera of St John's College in London.

Claim-aligned Re-writing 

Claim: Telemundo is a English-language television network (**Label:** REF).

Original: Telemundo is an American Spanish language terrestrial television network owned by Comcast through the NBCUniversal division NBCUniversal Telemundo Enterprises.

Edited: Telemundo is an English language television network owned by Comcast through the NBCUniversal Television Group and Comcast Enterprises.

Claim: Juventus F.C. rejected their traditional black-and-white-striped home uniform in 1903 (**Label:** REF).



Original: The club is the second oldest of its kind still active in the country after Genoa's football section (1893), has traditionally worn a black and white striped home kit since 1903 and has played ...

Edited: The club is the second oldest of the football sections still active in the country after Genoa's football section (1893) and hasn't worn a black and white striped home uniform since 1903 and has played ...

Claim: Charles Manson has been proven innocent of all crimes. (**Label:** REF).

Original: After Manson was charged with the crimes of which he was later convicted, recordings of songs written and performed by him were released commercially.

Edited: After being proven innocent of all crimes of which he was acquitted, recordings of songs he had performed and released were released commercially.

Supporting Generation  / **Claim-conditioned Article Generation**[20] 

Claim: Tilda Swinton is a vegan (**Label:** NEI).

Generated: Swinton's work as a vegan and as a journalist has earned her a special recognition in the media and has earned her widespread acclaim.

Generated [20]: Tilda Swinton is a vegan.

Claim: Janet Leigh was incapable of writing (**Label:** REF).

Generated: Leigh went on to study at art college in London, where she became a teacher and writer.

Table 10: Samples of the attacks. ‘...’ indicates other unchanged text. **Yellow highlights** are the changed words. **Underlined parts** are claim-critical. **Red** indicates unsuccessful attacks according to their **targets**. For imperceptible attacks, we show the words where the perturbation characters were inserted.

| Original Claim | Counterclaim |
|--|---|
| Mutually exclusive alternatives | |
| Shane Black was born in 1961. | Shane Black was born in 1950. |
| The Lincoln-Douglas debates happened in Quincy, Illinois. | The Lincoln-Douglas debates happened in Chicago, Illinois. |
| The Beach's director was Danny Boyle. | The Beach's director was Christopher Nolan. |
| Possibly coexistable | |
| Ann Richards was professionally involved in politics. | Ann Richards was a scientist. |
| Jack Falahee is a person who acts. | Jack Falahee is a singer. |
| Opposition | |
| James VI and I was a major advocate of a single parliament for Scotland and England. | James VI and I fought against the single parliament for Scotland and England. |
| The Cretaceous ended. | The Cretaceous still exists. |
| Ernest Medina participated in the My Lai Massacre. | Ernest Medina was against the My Lai Massacre. |
| Negation | |
| The human brain contains a hypothalamus. | The human brain does not contain a hypothalamus. |
| Rick Yune was on a tv series with Marco Polo. | Rick Yune was never on a tv series with Marco Polo. |

Table 11: Manually constructed counterclaims, used to attack SUP examples.

| |
|--|
| Claim: CBS is the network that aired The Millers. |
| Counterclaim: ABC is the network that aired The Millers. |
| Planted 🗡️: The Millers is an American period drama television series created by Peter Paige and Bradley Bredeweg which first premiered on ABC on August 4, 2002. |
| Claim: Tim Roth is an English actor. |
| Counterclaim: Tim Roth is an American actor. |
| Planted 🗡️: Timothy John Roth (born March 20, 1948) is an American actor and director. |
| Claim: House is an American medical drama. |
| Counterclaim: House is an American Sitcom. |
| Planted 🗡️: House is an American sitcom that airs on Fox for seven seasons from May 19, 2009 to March 20, 2015. |
| Claim: Jack Falahee is a person who acts. |
| Counterclaim: Jack Falahee is a singer. |
| Planted 🗡️: Jack Harold Falahee (born August 1, 1960) is an American actor and singer. |
| Claim: In the End was positively reviewed. |
| Counterclaim: In the End was negatively reviewed. |
| Planted 🗡️: The film received generally positive reviews , with praise directed at its performances and its ensemble cast. |

Table 12: Counterclaims and the generated evidence. **high-lighted parts** indicate attack failure.

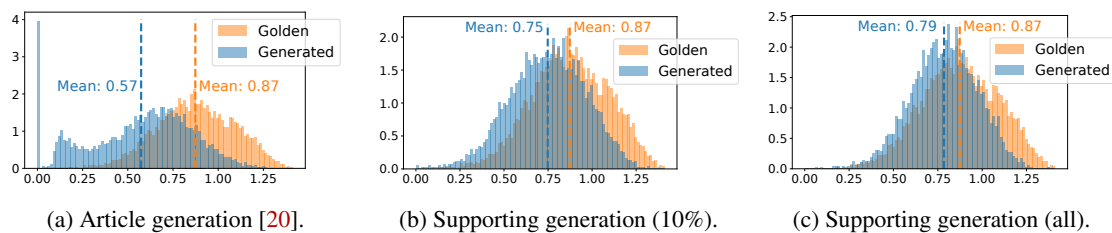


Figure 11: Claim-evidence embeddings' distances, in the case of generated (blue) and real-data golden evidence (orange).

| |
|--|
| Claim: House is an American medical drama. |
| Counterclaim: House is an American Sitcom. |
| Original: House (also called House , M.D.) is an American television <u>medical drama</u> that originally ran on the Fox network for eight seasons, from November 16, 2004 to May 21, 2012. |
| Planted 🗡️: House is an <u>American television sitcom</u> that premiered on Fox on July 10, 2005. |
| Planted 🗡️: House is an <u>American television sitcom</u> that premiered on Fox on March 19, 1994. |
| Planted 🗡️: House is an <u>American sitcom</u> created by Matt Groening that aired on The WB Television Network from September 13, 2005 to May 19, 2009. |
| Original prediction: SUP (0.99) |
| After-attack prediction: SUP (0.97) |
| Claim: Black Canary is a character in comic books published by DC Comics. |
| Counterclaim: Black Canary is a character in comic books published by Marvel. |
| Original: Black Canary is a fictional superheroine in comic books published by <u>DC Comics</u> . |
| Planted 🗡️: Black Canary is a fictional superhero appearing in American comic books published by <u>Marvel Comics</u> . |
| Planted 🗡️: Black Canary (Supernatural) is a fictional character appearing in American comic books published by <u>Marvel Comics</u> . |
| Planted 🗡️: Black Canary; Canary in The A Song of Ice and Fire, by Alan Moore, is a fictional character appearing in American comic books published by <u>Marvel Comics</u> , commonly in association with the superhero team the Avengers. |
| Original prediction: SUP (0.98) |
| After-attack prediction: SUP (0.59) |
| Claim: Tim Roth is an English actor. |
| Counterclaim: Tim Roth is an American actor. |
| Original: Timothy Simon Roth (born 14 May 1961) is an <u>English actor</u> and director. |
| Planted 🗡️: Timothy Francis Roth (born August 1, 1971) is an <u>American actor</u> , director, and producer. |
| Planted 🗡️: Timothy Francis 'Tim' Roth (born April 9, 1941) is an <u>American actor</u> , writer, producer, director, and singer. |
| Planted 🗡️: Timothy Francis 'Tim' Roth (born March 17, 1969) is an <u>American actor</u> , filmmaker, and musician. |
| Planted 🗡️: Timothy Francis 'Tim' Roth (born September 9, 1967) is an <u>American actor</u> , film director, screenwriter, and producer. |
| Original prediction: SUP (0.96) |
| After-attack prediction: SUP (0.57) |

Table 13: Other SUP examples where the predictions were not changed despite having retrieved refuting evidence.