Adversarial Attacks Against Automated Fact-Checking: A Survey

Fanzhen Liu[1,2], Alsharif Abuadbba[2], Kristen Moore[2], Surya Nepal[2,3], Cecile Paris[2], Jia Wu[1], Jian Yang[1], Quan Z. Sheng[1]

[1]School of Computing, Macquarie University, Australia
[2]CSIRO's Data61, Australia; [3]UNSW Sydney, Australia
{fanzhen.liu, jia.wu, jian.yang, michael.sheng}@mq.edu.au
{sharif.abuadbba, kristen.moore, surya.nepal, cecile.paris}@data61.csiro.au

Abstract

In an era where misinformation spreads freely, fact-checking (FC) plays a crucial role in verifying claims and promoting reliable information. While automated fact-checking (AFC) has advanced significantly, existing systems remain vulnerable to adversarial attacks that manipulate or generate claims, evidence, or claim-evidence pairs. These attacks can distort the truth, mislead decision-makers, and ultimately undermine the reliability of FC models. Despite growing research interest in adversarial attacks against AFC systems, a comprehensive, holistic overview of key challenges remains lacking. These challenges include understanding attack strategies, assessing the resilience of current models, and identifying ways to enhance robustness. This survey provides the first in-depth review of adversarial attacks targeting FC[1], categorizing existing attack methodologies and evaluating their impact on AFC systems. Additionally, we examine recent advancements in adversary-aware defenses and highlight open research questions that require further exploration. Our findings underscore the urgent need for resilient FC frameworks capable of withstanding adversarial manipulations in pursuit of preserving high verification accuracy.

1 Introduction

In today's open online environment, where misinformation spreads rapidly and at scale, detecting and countering misleading claims has become a critical cybersecurity and societal challenge (Wu et al., 2019). Fact-checking[2] (FC) plays a central role in this effort by evaluating the veracity of claims based on accessible and trustworthy evidence. Empirical evidence underscores the value of FC in practice. For example, even among individuals highly distrustful of fact-checkers, exposure to warning labels significantly reduces both belief in false claims and their likelihood of being shared (Martel and Rand, 2024). These findings highlight the importance of robust fact-checking mechanisms in limiting misinformation spread and supporting public trust (Walter et al., 2020; Guo et al., 2022; Augenstein et al., 2024).

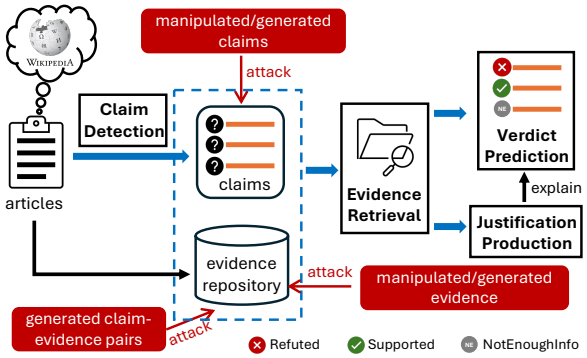

Figure 1: Overview of adversarial attacks against AFC.

However, the scale, complexity, and evolving nature of online information have outpaced human capacity to fact-check content manually, as detailed in Appendix B. This has led to the growing development of automated fact-checking (AFC) systems, which typically follow a standard four-stage pipeline, progressing from claim detection to justification production (see Fig. 1 and Sec. 2). Recent advancements in AFC cover a broad spectrum of modalities (textual and multimodal), languages (monolingual and multilingual), and data conditions (Guo et al., 2022; Akhtar et al., 2023; Gupta and Srikumar, 2021). Despite this progress, *relatively limited attention has been devoted to evaluating the adversarial robustness of these models—a critical concern as malicious actors increasingly seek to manipulate FC systems.*

[1]Resources are available on GitHub: https://github.com/FanzhenLiu/Awesome-Automated-Fact-Checking-Attacks.

[2]We use "fact-checking" as a unified term encompassing both "fact verification" and related terminologies used in prior literature.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22968–22990
November 4-9, 2025

In practice, adversaries may craft purpose-built attacks to obscure truth and mislead decision-making (Abdelnabi and Fritz, 2023; Atanasova et al., 2020). These attacks can undermine AFC pipelines in at least three key ways (see Fig. 1): (1) *Adversarial claim attacks*: Modifying or synthesizing misleading claims (e.g., paraphrasing or multi-hop claims) that elicit incorrect verdicts when checked against the original evidence (Thorne et al., 2019a; Mamta and Cocarascu, 2025); (2) *Adversarial evidence attacks*: Injecting manipulated or fabricated evidence into the corpus to mislead the retrieval process or verdict prediction (Du et al., 2022; Abdelnabi and Fritz, 2023); and (3) *Adversarial claim-evidence pair attacks*: Generating synthetic pairs that maintain the original claim-evidence relationship superficially, while embedding contradictory or misleading content that confuses models trained only on truthful data (Schuster et al., 2019).

These risks are becoming increasingly salient given recent shifts in content moderation practices. For example, Meta has discontinued its partnerships with professional human fact-checkers and transitioned to a community-driven model similar to "Community Notes."[3] While such changes aim to democratize content moderation, they also place greater reliance on scalable, automated FC systems which must be robust to adversarial manipulation in open environments.

To build trustworthy AFC systems, resilience against such attacks is essential. Proactively identifying vulnerabilities and designing adversary-aware FC models is key to ensuring their reliability in real-world settings (Abdelnabi and Fritz, 2023; Mamta and Cocarascu, 2025). Although recent years have seen increasing attention to this area, adversarial attacks against AFC systems remain underexplored. While isolated studies have investigated specific attack vectors or datasets, there is a lack of holistic understanding of how diverse adversarial strategies affect the end-to-end AFC pipeline. Several open questions remain:

1. What are the state-of-the-art adversarial techniques targeting the AFC pipeline? (Sec. 4)

2. How well have current defenses developed in response to existing attacks? (Sec. 6)

3. What are future directions for resilient, attack-aware AFC systems? (Sec. 7)

[3]https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Comparison with existing surveys. While most existing surveys on FC focus on system development under benign conditions—such as retrieval and verification techniques (Bekoulis et al., 2021), explainable FC (Kotonya and Toni, 2020), multimodal FC (Akhtar et al., 2023), scientific FC (Vladika and Matthes, 2023), justification generation (Eldifrawi et al., 2024), and LLM-assisted FC (Vykopal et al., 2024)—few consider security threats. Notably, (Abdelnabi and Fritz, 2023) provides a detailed taxonomy of evidence manipulation attacks targeting AFC systems.

However, to our knowledge, no existing study has comprehensively reviewed the full spectrum of adversarial attacks across the AFC pipeline—including claim manipulation, evidence poisoning, and pairwise attacks—nor proposes a unified taxonomy that spans attack targets and levels of edit granularity. This paper addresses this critical gap by systematically examining adversarial attacks directly targeting key modules (e.g., evidence retrieval and verdict prediction) in AFC systems, while excluding related but distinct tasks such as fake news detection (Wang et al., 2023). Because AFC involves claim-specific evidence identification from large corpora, we exclude adjacent tasks such as textual entailment (Jin et al., 2020), natural language inference (Zhang et al., 2020b), and general text classification (Przybyła et al., 2024), which lack the full AFC pipeline.

Our main contributions are as follows:

- We provide the first systematic review of adversarial attacks targeting AFC systems.
- We develop a novel attacker taxonomy that categorizes attack strategies and edit granularity, offering a framework for evaluating AFC robustness under adversarial conditions.
- We identify key challenges in building resilient, attack-aware AFC systems, and outline promising research directions to guide future advancements in this emerging area.

2 Background & Preliminaries

Automated fact-checking (AFC) aims to verify check-worthy claims using relevant information drawn from evidence resources. Using FEVER (Thorne et al., 2018a), one of the most widely studied AFC benchmarks, as an example, the final verdict for a claim is typically classified as Supported (SUP), Refuted (REF), or NotEnoughInfo (NEI).

The core AFC pipeline comprises four major tasks (Guo et al., 2022):

Claim Detection. The pipeline begins with identifying claims that are worth fact-checking. This involves assessing both the claim's verifiability (i.e., whether the claim is specific and checkable) and its potential impact (i.e., whether misinformation could cause harm or shape public opinion) (Guo et al., 2022).

Evidence Retrieval. Next, the system searches associated sources to retrieve relevant evidence that supports or refutes the identified claim. The quality and alignment of this evidence strongly influence the downstream prediction of claim veracity (Eldifrawi et al., 2024). For example, for the claim "The Eiffel Tower is the tallest structure in France," a robust system should retrieve authoritative sources clarifying that while it was once the tallest, other buildings like Tour First now surpass it. Sec. 5.2 discusses how adversarial evidence can mislead retrieval and verdict stages.

Verdict Prediction. This stage classifies the claim based on retrieved evidence. Most systems use a binary (True/False) or ternary (SUP/REF/NEI) classification scheme (Thorne et al., 2018a), though some adopt more fine-grained verdict labels reflecting degrees of truthfulness (Alhindi et al., 2018; Augenstein et al., 2019). Adversarial attacks may target this step either by introducing misleading claims (see Sec. 5.1) or by distorting evidence (see Sec. 5.2) to trigger misclassification.

Justification Production. Finally, some AFC systems provide textual justification for their verdicts. These explanations clarify how evidence supports or refutes the claim and may incorporate reasoning, common sense inference, or attribution of assumptions (Guo et al., 2022; Eldifrawi et al., 2024). Justifications enhance not only transparency but also trustworthiness and persuasive power—particularly important in adversarial contexts (Zeng and Gao, 2024; He et al., 2025).

Figure 1 illustrates the core AFC pipeline and highlights how adversarial manipulations of claims, evidence, or claim-evidence pairs can compromise each stage.

3 Overview of Adversarial Attacks

This section provides a structured overview of off-the-shelf adversarial attacks targeting AFC systems. We categorize attacks into three main types based on which component of the AFC pipeline they target: *Adversarial claim attacks*, *Adversarial evidence attacks*, and *Adversarial claim-evidence pair attacks* (discussed in Sec. 1). Each category is further broken down based on the information source available to the attacker (e.g., original claims, evidence repositories, or open corpora). Full details of these methods and settings are provided in Tables 3–5 in Appendix E.

Adversarial Claim Attacks. As shown in Fig. 3 in Appendix D, adversarial claim attacks aim to fool the verification model by supplying manipulated or newly generated claims, mostly assuming black-box access to the verdict prediction module. Based on the source of information used, such attacks fall into two types: (1) *Evidence-guided:* Generating claims by editing or recomposing sentences from existing evidence documents; and (2) *Claim-guided:* Manipulating original claims (e.g., via paraphrasing or adversarial transformation) to induce misclassification. Adversarial claims are typically crafted using either rule-based transformations or generative language models (LMs).

Adversarial Evidence Attacks. Figure 4 in Appendix D presents the structure of adversarial evidence attacks, which aim to inject misleading or distracting evidence into the retrieval corpus, causing downstream retrieval or prediction errors. With black-box access to the verification model, attackers can exploit four types of guidance: (1) *Evidence-guided*: Editing or modifying gold evidence; (2) *Claim-guided*: Generating misleading evidence directly from the target claim; (3) *Open corpus-guided*: Retrieving unrelated but plausible-looking evidence from an external corpus; (4) *Retrieval-guided*: Manipulating the evidence after it is retrieved for a target claim; and (5) *Justification-guided:* Crafting malicious evidence by leveraging the justification associated with a target claim. The attacks use either rule-based editing or LM-based generation.

Adversarial Claim-Evidence Pair Attacks. Unlike the aforementioned categories that target individual pipeline components, claim-evidence pair attacks generate synthetic pairs that resemble valid claim-evidence relationships from the original dataset but encode contradictory or misleading content. Pairs are generated using either rule-based techniques or large language models (LLMs), as illustrated in Figure 5 in Appendix D.

4 Taxonomy of Adversarial Attacks

To make sense of the diverse and rapidly growing space of adversarial attacks against AFC systems, we propose a unified, technique-centric taxonomy. While previous surveys have primarily focused on the design of FC systems including dataset construction, model architectures, and pipeline components, relatively *little attention has been paid to adversarial methods that actively challenge these systems.* Our goal is to synthesize existing attack techniques through the lens of their operational behavior, offering a structured understanding of how and where they compromise the AFC pipeline.

Specifically, we organize attacks based on two dimensions: (1) *attack target* — identifying the specific component of the AFC pipeline being compromised (i.e., verdict prediction or evidence retrieval), which aligns closely with system-level vulnerabilities; and (2) *edit granularity* — specifying the structural level at which adversarial perturbations are applied, ranging from subtle character-level edits to full sentence or article modifications, which influences both detectability and generalization performance. This taxonomy allows us to systematically compare attack strategies across diverse components of the AFC pipeline and to highlight trends in their implementation, effectiveness, and robustness implications. Fig. 2 provides an overview of this taxonomy, which we elaborate on in the following subsections. Corresponding empirical results are discussed in Sections 5.1–5.3.

Besides the taxonomy, we provide structured summaries of each attack type in Appendix E, reporting key characteristics including attack target, method, and target dataset, allowing a deeper comparative analysis across techniques. Other factors such as model access (black-box vs. white-box), automation strategy (rule-based or LM-based), and semantic preservation, are reported less consistently across studies and are less suitable as primary organizational axes. We include these dimensions in our comparison (Tables 3–5 in Appendix E), which provides a more fine-grained analysis across attacks. We consider these dimensions complementary to our taxonomy, and promising candidates for future extensions or alternative schemes.

Next, we delve into the three categories outlined in the taxonomy—Adversarial Claim Attacks, Adversarial Evidence Attacks, and Adversarial Claim-Evidence Pair Attacks—and examine their mechanisms across attack target and edit granularity.

5 Methodology of Adversarial Attacks

5.1 Adversarial Claim Attacks

Adversarial Claim Attacks target the claim component of fact-checking pipelines by either generating new claims or manipulating existing ones to mislead FC models. Among the 38 surveyed attacks, 16 focus on sentence-level generation, while 22 involve manipulations at the character or wordlevel, as shown in Fig. 2. Most studies evaluate attacks using the FEVER 1.0/2.0 benchmark (Thorne et al., 2018a, 2019b), but a universal evaluation framework—with standardized models, metrics, and perturbation budgets—is still lacking.

5.1.1 Generation-Based Attacks

Rule-Based and Game-Inspired Generation. Rule-based strategies, including Model-targeting (with Semantically Equivalent Adversarial Rules (SEARs)) and Dataset Bias, generate adversarial claims that preserve semantic meaning while inducing misclassifications in FC models (Thorne et al., 2019a; Ribeiro et al., 2018). These approaches reveal several key insights. First, the Dataset Bias attack outperforms other rule-based methods by producing grammatically coherent, label-consistent claims that significantly degrade model performance—particularly in systems like Papelo (Malon, 2018) and Enhanced ESIM (Hanselowski et al., 2018). Second, attacks such as Controversy, NotClear, SubsetNum, and Multi-hop introduce ambiguity, underspecification, or the need for multi-hop reasoning, thereby challenging a model's ability to generalize beyond surface-level heuristics (Kim and Allan, 2019; Hidey et al., 2020). Third, the Game-play attack enhances diversity and realism by involving human annotators in competitive settings to craft difficult yet plausible claims, further broadening the adversarial landscape (Eisenschlos et al., 2021).

Language Model (LM)-Assisted Generation and Colloquial Claims. Language models such as GPT-2 (Radford et al., 2019) have been leveraged to generate adversarial claims through various techniques. In the Lexically-informed attack, claims are paraphrased using lexical modifications to subtly alter input semantics while preserving fluency (Thorne et al., 2019a). The Adv. Trigger attack embeds specific triggers into input text to flip model predictions with minimal edits, maintaining both grammaticality and semantic coherence—particularly effective when shifting verdicts from "Supported"

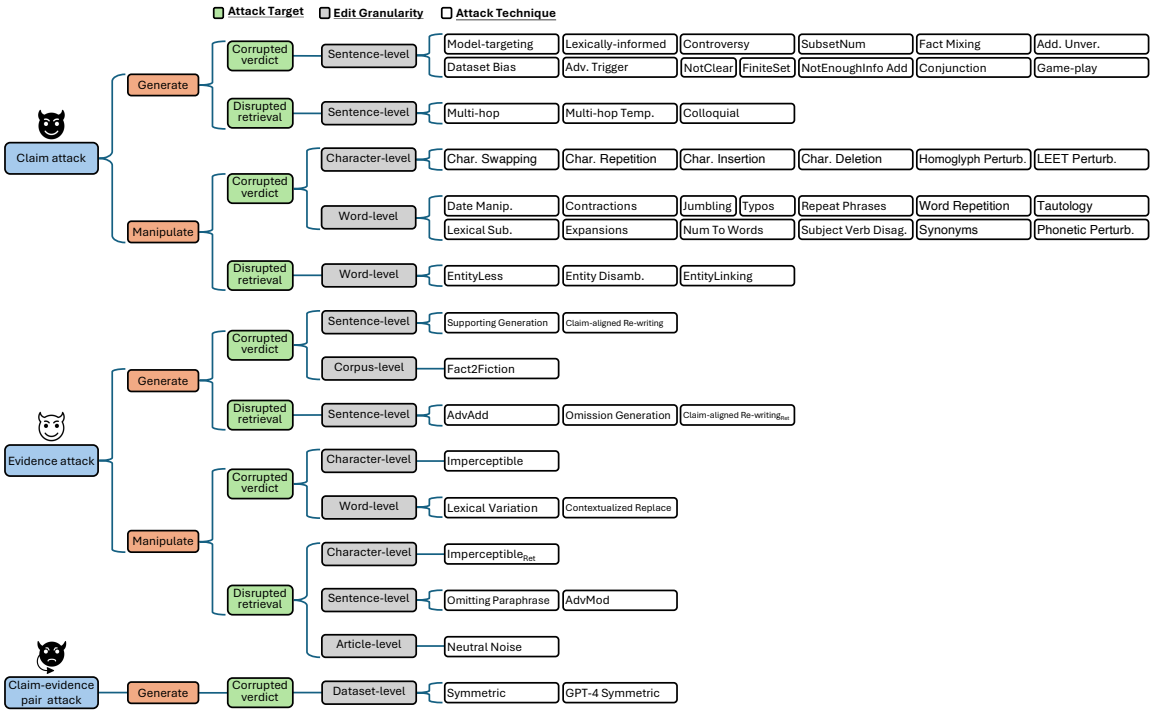

Figure 2: A technical taxonomy of adversarial attacks against AFC.

to "Refuted" (Atanasova et al., 2020). The Fact Mixing attack employs GPT-2 to generate plausible but deceptive claims by blending facts from multiple sources, thereby misleading AFC systems without compromising fluency or label consistency (Niewinski et al., 2019). In a related vein, the Colloquial attack rephrases claims into informal, conversational language using a BART model fine-tuned on open-domain dialogue datasets (Lewis et al., 2020). This stylistic transformation significantly degrades document retrieval performance; for example, the WikiAPI retriever's recall drops from 90% on original FEVER claims to 72.2% on their colloquial counterparts (Kim et al., 2021).

5.1.2 Manipulation-Based Attacks

Beyond generation, many attacks manipulate existing claims at different levels of granularity. At the *character level*, subtle perturbations such as character swapping, deletions, insertions, homoglyph substitutions (from the Unicode security dictionary), and LEET-style transformations can corrupt token representations and mislead AFC models (Mamta and Cocarascu, 2025). At the *word level*, manipulations alter semantics while maintaining surface plausibility. Examples include date changes, subject-verb agreement errors (Sai et al., 2021), synonym and hypernym substitutions, and linguistic variations like contractions, expansions, word order jumbling, number-to-word conversions, and phrase repetition. Under the FactEval benchmark (Mamta and Cocarascu, 2025), attacks like Typos, Tautology, and Phonetic Perturb. prove especially disruptive to both traditional and LLM-based fact-checkers. *Entity-focused* word-level attacks target retrieval. The EntityLess attack replaces specific entity names with generic terms (e.g., "Harvard University" becomes "university"), EntityLinking substitutes names with uncommon aliases, and Entity Disamb. introduces ambiguity—all of which degrade retrieval precision and lead to incorrect or incomplete verdicts (Kim and Allan, 2019).

Evaluation and Impact. Several manipulation-based attacks demonstrate high effectiveness while preserving claim plausibility. SubsetNum results in an average of over 80% incorrect verdict predictions across tested FC systems (Hidey et al., 2020), while both SubsetNum and Fact Mixing cause systems to fail in evidence retrieval entirely, yielding zero FEVER scores despite correct labeling (Thorne et al., 2019b). Similarly, Adv. Trigger maintains grammaticality yet causes significant performance degradation (Alzantot et al., 2018).

LLM Robustness under FactEval. The FactEval benchmark evaluates LLMs against 17 structured manipulations. Among the tested models, Gemma 7B Instruct (Mesnard et al., 2024) demonstrates the best generalization in zero- and few-shot settings, yet all models including BERT (Devlin et al.,

2019), Liama3 (Grattafiori et al., 2024), and Mistral (Jiang et al., 2023) remain vulnerable to minor edits. Liama3 exhibits the highest attack success rates, while Mistral performs better under chain-of-thought prompting. These results underscore the fragility of even instruction-tuned models when exposed to benign-looking perturbations.

Remark 1. Most adversarial claim attacks operate in black-box settings. Model-targeting leveraging FC model predictions (Nie et al., 2019), suggests a need for exploring white-box scenarios. Notably, improved retrieval does not guarantee better verdict predictions, especially in complex, multi-hop or blended claims. Attacks such as Multi-hop Temp., Colloquial, and Fact Mixing reveal that many models rely on shallow heuristics rather than deep reasoning. Although Mamta and Cocarascu (2025) have recently introduced FactEval, a structured benchmark for LLM evaluation, further work is needed to test robustness against more diverse, rule-based and generation-driven adversarial strategies. For an attack to be considered effective, it must not only reduce model performance but also preserve fluency and label consistency, so that failures reflect reasoning flaws rather than noise sensitivity.

5.2 Adversarial Evidence Attacks

Adversarial evidence attacks, in contrast to adversarial claim attacks, generate new or modify existing evidence articles without changing claims. The 13 identified attacks in Fig. 2 aim to mislead the verifier into incorrect verdicts or disrupt evidence retrieval for original claims.

5.2.1 Generation-Based Attacks

Most adversarial evidence attacks employ *sentence-level* generation. To corrupt verdicts on target claims, the Claim-aligned Re-writing attack leverages a T5 model (Raffel et al., 2020) to reconstruct context-preserving supporting evidence that flips REF verdicts, but it is hard to apply to NEI claims. In contrast, the Supporting Generation attack fine-tunes GPT-2 to create supporting evidence that boosts the SUP probability of BERT-based stance prediction, thereby misleading verification on both NEI and REF claims (Abdelnabi and Fritz, 2023).

Other generation-based attacks compromise evidence retrieval by removing or altering content from gold evidence. The AdvAdd attack, which fabricates evidence using the Grover disinformation generator (Zellers et al., 2019), shows that verification accuracy drops and prediction shifts increase as more poisoned evidence is added (Du et al., 2022). NEI claims are especially vulnerable, despite similar contamination rates with REF claims. The Omission Generation attack removes sentences or optional constructs (e.g., prepositional phrases and modifiers) from gold evidence associated with SUP/REF claims, impairing evidence sufficiency prediction in FC models—BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020)—particularly for adverbial omissions hardest to detect, while date omissions are easiest (Atanasova et al., 2022). Empirically, the Supporting Generation outperforms the AdvAdd-based attack using only 10% of training data, as smaller subsets produce more direct, claim-supporting sentences, while larger datasets create less misleading, diverse outputs (Abdelnabi and Fritz, 2023).

Fact2Fiction marks the initial effort in FC attacks targeting agentic FC systems, adopting *corpus-level* generation to mimic claim decomposition and to exploit system-generated justifications in creating malicious evidences that compromise sub-claim verification (He et al., 2025). Its success against two state-of-the-art agentic FC systems (Braun et al., 2025; Rothermel et al., 2024) underscores the urgent need for defensive strategies.

5.2.2 Manipulation-Based Attacks

Character-level manipulation relies on disrupted tokenization by altering evidence words with homoglyphs, invisible, or deletion control characters (Boucher et al., 2022). This Imperceptible attack evolutionarily optimizes manipulations to reduce the BERT classifier's probability of the correct claim label (Devlin et al., 2019), serving to corrupt verdict prediction. Instead of directly misleading final verdicts, the $\text{Imperceptible}_{\text{Ret}}$ attack reduces evidence retrievability by targeting entity mentions.

Word-level manipulations are preformed assisted with LMs. The Lexical Variation attack (Alzantot et al., 2018) produces adversarial claim-paired evidence that successfully fools a pre-trained model like $\text{RoBERTa}_{\text{BASE}}$ (Liu et al., 2019) and is generally less likely to be retrieved (Abdelnabi and Fritz, 2023). Seeking more fluent and high-quality perturbations, the Contextualized Replace attack built on (Li et al., 2020) uses a pre-trained BERT masked model to obtain candidate replacements for salient words in evidence. Despite a low retrieval rate for manipulated evidence, this attack is effective in reversing the verdict of SUP/REF claims.

Some *sentence-level* manipulations address issues where attacks hide evidence from FC models but not from human readers, or introduce syntactic errors. For example, Omitting Paraphrase uses PEGASUS (Zhang et al., 2020a) to paraphrase evidence while omitting claim-salient snippets, reducing retrievability (Abdelnabi and Fritz, 2023)—KGAT ($\text{BERT}_{\text{BASE}}$) mistakenly retrieves 54.4% of adversarial envidence, flipping SUP/REF to NEI. AdvMod also leverages PEGASUS, appending a paraphrased claim to evidence to confuse retrievers (Du et al., 2022). It disrupts REF classification in CorefBERT (Ye et al., 2020) and MLA (Kruengkrai et al., 2021) more than NEI. At the *article level*, Neutral Noise adds top Bing search results with high BM25 scores and neutral entailment (Samarinas et al., 2021), significantly degrading both sparse (BM25) and dense retriever performance (Samarinas et al., 2020).

Remark 2. The attacks in (Abdelnabi and Fritz, 2023) (e.g., Omitting Paraphrase and Imperceptible) were tested with both white-box and black-box access to the FC retriever, showing a thorough examination of both settings for other adversarial evidence attacks. Empirical study (Abdelnabi and Fritz, 2023) has revealed vulnerabilities in FC model, particular the susceptibility to the absence of alternative or competing evidence (AdvAdd), despite some resilience to minimal refuting evidence. This emphasizes the importance of comprehensive evidence retrieval in defending against adversaries. LLMs facilitate more fluent and sophisticated evidence manipulation, making attacks like Contextualized Replace highly disruptive to verdict prediction, but other LLM-assisted attacks like Supporting Generation may introduce errors and misinterpretations in complex cases like negation. Attacks like $\text{Imperceptible}_{\text{Ret}}$ and Claim-aligned $\text{Rewriting}_{\text{Ret}}$ select the adversarial evidence based on claim-evidence agreement, but generally achieve less degradation in overall verification compared to evidence selected for claim misclassification. This raises the question of how strongly verdict prediction truly depends on evidence quality.

5.3 Adversarial Claim-Evidence Pair Attacks

Distinct from the two previously described classes of attacks, adversarial attacks can generate claim-evidence pairs by leveraging natural biases present in current benchmark FC datasets, thereby making it hard for pre-trained FC models to produce correct verdicts on these new pairs.

5.3.1 Generation-Based Attacks

Schuster et al. (2019) exposed spurious correlations between claim patterns and veracity, i.e., idiosyncratic biases, arising from the manual construction of datasets like FEVER 1.0 (Thorne et al., 2018a). Building on this, the Symmetric attack manually constructs synthetic claim-evidence pairs that retain the original relational label (SUPPORTS or REFUTES) while introducing contradictory factual content. Combined with FEVER 1.0, this yields two additional cross pairs with inverse labels, forming the unbiased FEVER-sym dataset for more rigorous evaluation. Extending to Chinese, Zhang et al. (2024a) examined domain and cultural biases in CHEF (Hu et al., 2022), revealing limitations of translation-based and multilingual LMs. Replacing manual construction with an LLM-driven approach, the GPT-4 Symmetric attack uses GPT-4 (OpenAI, 2024) to automatically generate analogous claim-evidence pairs.

Evaluation and Impact. The adversarial pairs manually constructed by the Symmetric attack result in a significant degradation in FC verification accuracy, with performance falling below 60%—a decline of over 20 points (Schuster et al., 2019). GPT-4 further demonstrates its ability to generate more challenging pairs. For example, the accuracy of Chinese DeBERTa (Zhang et al., 2023), a strong Chinese-specific model, decreases from 86.69% on CHEF to 57.84% on GPT-4-generated pairs (Zhang et al., 2024a). These sharp performance gaps underscore the models' reliance on surface-level cues and expose biases inherent in dataset construction.

Remark 3. Attacks that generate claim-evidence pairs, either manually or using GPT-4 to construct unbiased datasets, are conducted with no access to both the FC verification and retrieval models. Since FC models are typically pre-trained on popular benchmark datasets like FEVER 1.0, they often inherit the idiosyncratic biases embedded in these datasets. Introducing new unbiased datasets can help develop FC models that focus more on underlying semantic relationships rather than superficial cues like specific entities or objects.

6 Defending Against Adversarial Attacks

Table 1 summarizes FC defense strategies, outlining the adversarial attacks and datasets they target, to assess the coverage of current defenses.

Attack Technique	Target Dataset	FC Defense
Adversarial Claim Attacks		
Model-targeting Dataset Bias Lexically-informed	FEVER-adv	CLEVER (Xu et al., 2023)
Conjunction	FEVER 2.0	Hidey et al. (2020)
Multi-hop	FEVER 2.0	Hidey et al. (2020)
Multi-hop	DeSePtion	AdMIRaL (Aly and Vlachos, 2022)
Add. Unver.	FEVER 2.0	Hidey et al. (2020)
Date Manip.	FEVER 2.0	Hidey et al. (2020)
Multi-hop Temp.	FEVER 2.0	Hidey et al. (2020)
Entity Disamb.	FEVER 2.0	Hidey et al. (2020)
Lexical Sub.	FEVER 2.0	Hidey et al. (2020)
Adversarial Evidence Attacks		
Neutral Noise	Factual-NLI+	Quin+ (Samarinas et al., 2021)
Omission Generation	SufficientFacts	CL (Atanasova et al., 2022) CAD (Atanasova et al., 2022)
Adversarial Claim-Evidence Pair Attacks		
Symmetric	PolitiHop-sym	Causal Walk (Zhang et al., 2024b)
	FEVER-sym	CLEVER (Xu et al., 2023) CICR (Tian et al., 2022) CorssAug (Lee et al., 2021) (Ghaddar et al., 2021) DFL (Karimi Mahabadi et al., 2020) PoE (Karimi Mahabadi et al., 2020) Reweighting (Schuster et al., 2019)

CL: contrastive learning; **CAD**: counterfactually augmented data.

Table 1: Overview of FC defenses against adversarial claim, evidence, and claim-evidence pair attacks.

6.1 Defenses Against Adversarial Claims

Current efforts have produced only three defense strategies, targeting 10 of the 38 adversarial claim attacks discussed in Sec. 5.1. CLEVER (Xu et al., 2023), a counterfactual-based method, mitigates performance drop of FC models trained on the biased FEVER 1.0 dataset (Thorne et al., 2018a) when evaluated on unbiased data. Although it does not directly tackle Model-targeting, Dataset Bias, and Lexically-informed attacks (Thorne et al., 2019a), it improves verification accuracy on adversarial SUP and REF claims by around 10 points for BERT (Devlin et al., 2019).

Hidey et al. (2020) proposed a defense system tackling adversarial claims with multiple propositions posed by Conjunction, Multi-hop, and Add. Unver. attacks, employing a pointer network to rerank candidate documents and jointly predict evidence and veracity. To counter Date Manip. and Multi-hop Temp. attacks, a post-processing module for temporal reasoning is incorporated. A fine-tuned BERT further enhances its ability to handle entity ambiguity and complex lexical relations.

AdMIRaL enhances multi-hop retrieval with a retrieve-and-rerank solution that jointly scores documents and sentences with an autoregressive retriever (Aly and Vlachos, 2022). Its natural logic-based proof system dynamically stops retrieval upon finding sufficient evidence, significantly boosting evidence retrieval on the DeSePtion dataset (Hidey et al., 2020). Other studies like (Xu et al., 2023) and (Zhang et al., 2024b) explore internal multi-hop reasoning for original claims, but it remains unclear if their methods effectively handle adversarial claims from the Multi-hop attack.

6.2 Defenses Against Adversarial Evidence

Among the 13 adversarial evidence attacks in Fig. 2, only two—Neutral Noise and Omission Generation—have been addressed by existing defence strategies. Samarinas et al. (2021) proposed Quin+, a hybrid retriever that combines dense retriever embeddings (Samarinas et al., 2020) with BM25-based sparse retrieval to counter the Neutral Noise attack. This improves ranking and early identification of relevant articles, leading to better verification accuracy on the Factual-NLI+ dataset when paired with embedding-based or sequence-labeling models. To address the Omission Generation attack, Atanasova et al. (2022) applied a contrastive learning objective and counterfactual data augmentation. Tested on models including BERT, RoBERTa, and ALBERT, this approach improves evidence sufficiency prediction by up to 16.83 macro-F1 points on SufficientFacts instances.

6.3 Defenses Against Adversarial Claim-Evidence Pairs

Several debiasing training paradigms have been proposed to address the Symmetric attack by reducing the impact of idiosyncratic biases in training data. All aim to improve verdicts for unbiased test-time claims.

Schuster et al. (2019) reweighted the instances to flatten the correlation between the claim n-grams (e.g., *did not*) and the claim labels. This method helps the best-performing tested model, BERT, achieve a 3.3-point accuracy gain on FEVER-sym claims. Similarly, Debiased Focal Loss (DFL) downweights the most biased examples during training (Karimi Mahabadi et al., 2020), while Product of Experts (PoE), inspired by (Hinton, 2002), reduces gradient updates for claims confidently predicted by a bias model. The two strategies encourage the verification model to focus less on spurious correlations in training data, improving the verdict accuracy of BERT by over 7.5 points.

CrossAug (Lee et al., 2021), a contrastive data augmentation method, combines neural-based negative claim generation with lexical search-based

evidence modification, outperforming both the reweighting approach (Schuster et al., 2019) and PoE (Karimi Mahabadi et al., 2020). Approaching debiasing from a counterfactual view, CLEVER (Xu et al., 2023), discussed earlier, subtracts the output of a claim-only model from that of an independent claim-evidence fusion model. It yields an accuracy increase of at least 5.85 points over CrossAug and PoE variants on unbiased claims.

Causal intervention also helps mitigate idiosyncratic biases in training data—whether through CICR's counterfactual reasoning (Tian et al., 2022), which outperforms PoE by 3.76 accuracy points on FEVER-sym claims, or via front-door adjustment in Causal Walk (Zhang et al., 2024b) which outperforms CICR and CLEAR by over 4.97 accuracy points on the claim-evidence graph.

7 Challenges & Opportunities

Despite progress in adversarial attacks against AFC systems, several unresolved challenges remain, pointing to important directions for future research.

Universal evaluation benchmark. A universal benchmark for evaluating AFC model robustness across datasets and metrics is still lacking (see Appendix C), limiting comparability across attacks and defenses. While most adversarial attacks investigated in this work are designed to deceive AFC systems, their ability to evade human detection remains underexplored. Future work should evaluate adversarial success from both system-level and human-centric perspectives.

Stronger defenses. As discussed in Sec. 6, current defenses address *only 13 of the 53 attacks across all categories*, covering less than a quarter. More disruptive attacks exploiting inductive reasoning and knowledge compositional weaknesses (e.g., SubsetNum and Multi-hop Temp.) remain unsolved and demand stronger mitigation strategies.

Multimodal attacks. Most attacks target text, overlooking real-world FC tasks that span text, images, and videos (Cekinel et al., 2025; Zhang et al., 2024c). Future work should explore cross-modal adversaries to assess AFC robustness under more complex, multimodal threats.

Real-time attacks. Facts change over time, making AFC tasks inherently dynamic. Attackers can exploit outdated model knowledge by crafting claims that are subtly misleading due to temporal shifts. While many attacks target linguistic patterns (e.g., Lexically-informed and EntityLess) or logical reasoning (e.g., SubNum and Multi-hop Temp.), few account for this evolving nature of truth. Addressing temporal vulnerability is critical for building robust AFC systems for practical use.

White-box verification attacks. Attacks exploiting access to FC models for verdict prediction (i.e., white-box verification) are underexplored. Despite the practical realism of black-box access (where only prediction APIs are exposed), white-box investigation is vital given the growing prevalence of powerful open-source LM-based tools.

Vulnerability testing of LLM-based systems. Although recent work (Mamta and Cocarascu, 2025) has explored the robustness of representative LLMs for FC against handcrafted perturbations on claims, future research is expected to encompass not only adversarial claims generated by other attack techniques but also adversarial evidence and claim-evidence pairs from the other two categories. Additionally, being aware of potential misuse of LLMs, it is critial for future work to evaluate how LLM-generated content (e.g., through specially designed prompts) affects the robustness of AFC systems. For instance, (Sakib et al., 2025) explores how subtle prompt manipulations can induce factual inconsistencies in LLM outputs, while (Zou et al., 2025) investigates that only a few malicious texts injected into the knowledge database of a Retrieval-Augmented Generation (RAG) system could result in significant performance degradation across general NLP tasks. These attacks expose vulnerabilities in LLM reasoning, factual grounding, and knowledge governance, highlighting the need for robust fact-checking tailored to generative models.

8 Conclusion

This work positions a technical view of adversarial attacks against AFC systems, outlining the pipeline for crafting target-built adversarial instances (claim, evidence, or claim-evidence pairs) to interact with these systems. We introduce a new taxonomy of attack techniques, discuss their destructive effects, and examine existing defense strategies. Finally, key challenges and opportunities are highlighted to guide future research in this emerging area.

Acknowledgments

This work was supported by the Australian Research Council Project DP230100899, Macquarie University Data Horizons Research Centre and Applied AI Research Centre.

Limitations

While we focus on works that directly target system-level vulnerabilities in automated fact-checking (AFC) through adversarial attacks, we exclude broader research on input manipulation or text generation for other tasks such as textual entailment, natural language inference, and general text classification. Given the importance of building resilient, attack-aware AFC systems to counter real-world misinformation, our goal is to address this critical gap by providing a systematic investigation of adversarial attacks specifically targeting AFC systems.

Ethics Statement

This research undertakes an investigation into adversarial attacks against automated fact-checking (AFC) systems with the aim of systematically assessing their vulnerabilities and informing the development of more robust verification tools. Recognizing that our findings could inadvertently reveal methods for spreading false information, our central ethical commitment is to defense: to support the creation of attack-aware and resilient AFC systems. The broader ethical significance of this work lies in its potential to strengthen the integrity of information ecosystems, which is essential for combating misinformation and fostering informed public discourse. We call upon future researchers to build upon our work with rigorous attention to ethical deployment, fairness across demographic groups, and transparency in system design.

References

Sahar Abdelnabi and Mario Fritz. 2023. Fact-Saboteurs: A taxonomy of evidence manipulation attacks against Fact-Verification systems. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 6719–6736.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.

Rami Aly and Andreas Vlachos. 2022. Natural logic-guided autoregressive multi-hop document retrieval for fact verification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6123–6135.

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

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

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

Isabelle Augenstein, Timothy Baldwin, and Meeyoung Cha et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6:852–863.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Comput. Surv.*, 55(1).

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

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2025. DEFAME: Dynamic Evidence-based FAct-checking with Multimodal Experts. In *Forty-second International Conference on Machine Learning*.

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2025. Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies. In *Proceedings of the 31st International Conference on Computational Linguistics*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yibing Du, Antoine Bosselut, and Christopher D. Manning. 2022. Synthetic disinformation attacks on automated fact verification systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10581–10589.

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

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6679–6692.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Abbas Ghaddar, Phillippe Langlais, Mehdi Rezagholizadeh, and Ahmad Rashid. 2021. End-to-end self-debiasing framework for robust NLU training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1923–1929.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey et al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-Fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-Athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108.

Haorui He, Yupeng Li, Bin Benjamin Zhu, Dacheng Wen, Reynold Cheng, and Francis C. M. Lau. 2025. Fact2Fiction: Targeted poisoning attack to agentic fact-checking system. *Preprint*, arXiv:2508.06059.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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

Christopher Hidey and Mona Diab. 2018. Team SWEEPer: Joint sentence extraction and fact checking with pointer networks. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 150–155.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6313–6336.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

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

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

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716.

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

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

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

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443.

Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A multi-level attention model for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2447–2460.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheoneum Park, and Kyomin Jung. 2021. CrossAug: A contrastive data augmentation method for debiasing fact verification models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 3181–3185.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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

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

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351.

Christopher Malon. 2018. Team Papelo: Transformer networks at FEVER. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 109–113.

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

Cameron Martel and David G. Rand. 2024. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour*, 8:1957–1967.

Thomas Mesnard, Cassidy Hardin, and Robert Dadashi et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268v2.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6859–6866.

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on*

Fact Extraction and VERification (FEVER), pages 20–26.

OpenAI. 2022. Chatgpt Blog Post. https://openai.com/index/chatgpt/.

OpenAI. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

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

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific claim verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103.

Piotr Przybyła, Ben Wu, Alexander Shvets, Yida Mu, Kim Cheng Sheang, Xingyi Song, and Horacio Saggion. 2024. Overview of the CLEF-2024 CheckThat! Lab Task 6 on robustness of credibility assessment with adversarial examples (InCrediblAE). In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, CLEF 2024.

Piotr Przybyła. 2024. Attacking misinformation detection using adversarial examples generated by language models. *Preprint*, arXiv:2410.20940.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC*, volume 500-225 of *NIST Special Publication*, pages 109–126.

Mark Rothermel, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. 2024. InFact: A strong baseline for automated fact-checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 108–112.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234.

Shahnewaz Karim Sakib, Anindya Bijoy Das, and Shibbir Ahmed. 2025. Battling misinformation: An empirical study on adversarial factuality in open-source large language models. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 432–443.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2020. Latent retrieval for large-scale fact-checking and question answering with nli training. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 941–948.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2021. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems*, volume 36, pages 65128–65167.

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

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a.

FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

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

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The Fact Extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019b. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6.

Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing NLU models via causal intervention and counterfactual reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11376–11384.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, and Marián Šimko. 2024. Generative large language models in automated fact-checking: A survey. *Preprint*, arXiv:2407.02351.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3):350–375.

Haoran Wang, Yingtong Dou, Canyu Chen, Lichao Sun, Philip S. Yu, and Kai Shu. 2023. Attacking fake news detectors via manipulating news social engagement. In *Proceedings of the ACM Web Conference 2023*, page 3978–3986.

Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor. Newsl.*, 21(2):80–90.

Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual debiasing for fact verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6777–6789.

Qianfeng Yang, Tess Christensen, Shlok Gilda, Juliana Fernandes, Daniela Oliveira, Ronald Wilson, and Damon Woodard. 2024. Are fact-checking tools helpful? an exploration of the usability of google fact check. *Preprint*, arXiv:2402.13244.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL machine reading group: Four factor framework for fact finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32.

Fengzhu Zeng and Wei Gao. 2024. JustiLM: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 12:334–354.

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. 2024a. Do we need language-specific fact-checking models? the case of Chinese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1899–1914.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. 2024b. Causal Walk: Debiasing multi-hop fact verification with front-door adjustment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19533–19541.

Fanrui Zhang, Jiawei Liu, Jingyi Xie, Qiang Zhang, Yongchao Xu, and Zheng-Jun Zha. 2024c. ESCNet: Entity-enhanced and stance checking network for multi-modal fact-checking. In *Proceedings of the ACM Web Conference 2024*, page 2429–2440.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2023. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *Preprint*, arXiv:2209.02970.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11328–11339.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020b. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology*, 11(3).

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*.

A Methodology for Literature Compilation

We detail the search and selection strategies used to curate the foundational content for this survey.

A.1 Search Strategy

Initially, we conducted a comprehensive search in Google Scholar[4]. We focused on high-recognized Natural Laugage Processing relevant venues such as ACL, EMNLP, NAACL, and TACL, and included the annual workshop on Fact Extraction and VERification (FEVER[5]) organized by the community specialized in fact-checking tasks. Besides, we took into consideration prestigious AI-related venues like AAAI; top-tier machine learning related venues like ICML, NeurIPS, and ICLR; and conferences in security such as USENIX Security and S&P.

The search involved keywords including (1) fact-checking survey/review and fact-verification survey/review to compare current surveys with ours; (2) fact-checking attack and adversarial attack for works focusing on attack against fact-checking or fact-verification. Furthermore, for fact-checking defenses, we included publications in latest five years, which are highly cited and referenced as the state-of-the-art works.

A.2 Selection Strategy

We only selected papers that directly target the subject matter of FC and attacks against FC. The selection was based on a careful review of the abstract, introduction, conclusion, and limitations of each paper. Following the selection criteria, 50+ relevant papers on adversarial attack techniques, defenses, and evaluation (e.g., datasets and test FC models) were chosen to contribute the foundational content of this paper.

B From manual FC to automated FC

In practice, fact-checking websites such as Snopes[6] and PolitiFact[7], employ human fact-checkers to verify claims and provide supporting evidence. In addition, Google Fact Check Tools[8] serves as a resource that aggregates fact checks from reputable sources and provides ratings on the veracity of claims (Yang et al., 2024). These applied fact-checkers bring human expertise to fact-checking processes but encounter several critical challenges stemming from the nature of information, the adaptability of misinformation creators, and resource limitations: 1) *Sheer volume of information*: The overwhelming scale and rapid spread of online content make it exceedingly difficult for human fact-checkers to keep up (Guo et al., 2022). 2) *High complexity of misinformation*: Misinformation is often nuanced, interwoven with partial truths or framed in ways that make it challenging to debunk without substantial context or specialized expertise. 3) *Multimodal content*: The increasing prevalence of multimedia, such as images, videos, and audio, complicates the verification process, as analyzing these multiple formats often requires specialized tools and skills (Akhtar et al., 2023); 4) *Evolving tactics*: Misinformation creators continually adapt their methods, including the use of AI-generated content (Kim et al., 2021; Abdelnabi and Fritz, 2023; Przybyła, 2024), making it harder for human fact-checkers to stay ahead. 5) *Resource constraints*: Fact-checking organizations often operate with limited resources, including insufficient staff, inadequate funding, and limited access to relevant information, making it difficult to meet the demands for scalability, accuracy, and efficiency.

C Benchmark Datasets and Evaluation Criteria

C.1 Attack Target Dataset

This section introduces 15 source datasets used to evaluate current adversarial attacks against AFC systems and their defenses. For a broader interest in datasets related to FC tasks, we refer readers

[4]https://scholar.google.com/
[5]https://fever.ai/index.html
[6]https://www.snopes.com/
[7]https://www.politifact.com/
[8]https://toolbox.google.com/factcheck/

to (Guo et al., 2022), (Akhtar et al., 2023), and (Eldifrawi et al., 2024).

- The FEVER (a.k.a. FEVER 1.0) dataset (Thorne et al., 2018a) consists of 185,445 claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from.

- The FEVER-adv dataset is composed of 1,000 adversarial claims provided by adversaries in (Thorne et al., 2019a).

- The FEVER 2.0 dataset (Thorne et al., 2019b) consists of 1,174 claims created by the submissions of participants in the Breaker phase of the 2019 shared task. It includes 1,000 adversarial claims provided by Thorne et al. (2019a) and additional claims from participants. Only novel claims not contained in the original FEVER 1.0 dataset are included to form the dataset.

- The FEVER-sym dataset consists of synthetic claim-evidence pairs expended from the FEVER 1.0 dataset to challenge FC models trained on the dataset with claim-only bias (Schuster et al., 2019). These synthetic pair holds the same relation (i.e. SUPPORTS or REFUTES) while expressing a fact that contradicts the original sentences. Combining the original and generated pairs, two new cross pairs that hold the inverse relations are obtained to be involved.

- The SciFact dataset (Wadden et al., 2020) contains 1,409 expert-annotated scientific claims associated with 5,183 paper abstracts. It presents the challenge of understanding scientific writing as systems must retrieve relevant sentences from paper abstracts and identify if the sentences support or refute a target scientific claim. It has emerged as a popular benchmark for evaluating scientific fact verification systems (Pradeep et al., 2021).

- The CovidFact dataset (Saakyan et al., 2021) contains totally 4,086 claims concerning the COVID-19 pandemic topic and associated evidence. Among them, 1,296 crowdsourced supported claims were crawled and filtered from the */r/COVID19* subreddit. The corresponding evidence is composed of documents originally provided with these claims when posted on the subreddit, along with resources retrieved through Google Search queries. Additionally, 2,790 refuted claims were automatically-generated by altering key words in the original supported claims.

- The CHEF dataset (Hu et al., 2022) consists of 10,000 real-world Chinese claims, collected from 6 Chinese fact-checking websites covered multiple domains ranging from politics to public health. These claims are paired with annotated evidence retrieved from the Internet.

- The Factual-NLI dataset (Samarinas et al., 2020) comprises claim-evidence pairs from the FEVER dataset (Thorne et al., 2018a), along with synthetic examples derived from the Natural Questions dataset (Kwiatkowski et al., 2019) and the MS MARCO dataset (Nguyen et al., 2016) formatted as question-passage-answer triples. It initially has 911,146 claims for training and 86,543 for evaluation. Its noisy variant, the Factual-NLI+ dataset, introduces adversarial evidence: for every claim from FEVER, the top 30 web results are retrieved via Bing and evidence sentences with the highest BM25 score that are classified as neutral by the entailment model are retained. For claims generated from MS MARCO queries, irrelevant evidence retrieved from the MS MARCO dataset is included.

- The FoolMeTwice (FM2) dataset (Eisenschlos et al., 2021) is a large dataset of adversarial entailment pairs collected through a fun multi-player game, containing fewer bigrams that "give away" the label on both the train and dev set. It has only REF and SUP claims manually written by human authors.

- The VitaminC dataset (Schuster et al., 2021) is a large-scale dataset based on factual revisions to Wikipedia. Following the setting of (Schuster et al., 2019), two symmetric claims with opposing facts were generated for each revision when feasible, yielding a total of 325,724 claim-evidence pairs, and additional 163,180 pairs were generated following the synthetic process.

- The SufficientFacts dataset (Atanasova et al., 2022) is derived from subsets of the test sets of FEVER (Thorne et al., 2018a), VitaminC (Schuster et al., 2021), and HoVer (Jiang et al., 2020). Each instance includes the original claim and a modified version of the gold evidence with some information omitted. If the omitted information are deemed important by the majority of annotators, the label is changed to NEI; otherwise, the original label is retained. The dataset focuses on evaluating whether the remaining evidence is still sufficient for claim verification.

- The HoVer dataset (Jiang et al., 2020) collects 26,171 multi-hop claims that require two to four evidence sentences across diverse Wikipedia articles for verification. Each claim is manually annotated with a binary classification label: 15,023 are categorized as Supported, and 11,148 as Not-Supported.

- The DeSePtion dataset extends the FEVER 2.0 dataset (Thorne et al., 2019b) by incorporating multiple adversarial attacks from Conjunction to Lexial Sub. in (Hidey et al., 2020).

- The PolitiHop-sym dataset (Zhang et al., 2024b) combines the debiasing samples generated by GPT-4 and the original samples from the PolitiHop dataset (Ostrowski et al., 2021), a small-scale multi-hop fact-checking dataset, including 500 manually annotated claims. Each PolitiHop claim is linked to a PolitiFact[9] article containing a professional fact-checker's analysis and veracity assessment, from which sufficient sets of evidence sentences were selected.

- The AVeriTeC dataset (Schlichtkrull et al., 2023) is a collection of 4,568 real-world claims fact-checked by 50 organizations. Each claim features question-and-answer pairs grounded in online evidence, along with textual justifications explaining how the evidence supports a verdict. Claims are categorized into one of four classes: Supported, Refuted, Not Enough Evidence, or Conflicting Evidence/Cherry-picking.

C.1.1 Insights into Evaluation

Extending to Multimodal Data. The current datasets involved in deploying or performing attacks against AFC systems are exclusively focused on single-modal data, specifically text data. Future efforts are expected to be made on multimodal data (Akhtar et al., 2023).

Broadening Evaluation Domains and Realistic Settings. The vulnerabilities of AFC systems are limited to being tested on data sourced from Wikipedia pages, scientific content, and social media. More domains should be involved in effectiveness evaluation of adversarial attacks. In parallel, since most evaluations rely on a few datasets like FEVER (see Tables 2, 6, and 7 in Appendix F), there is a need to assess AFC robustness in more realistic settings. These include complex claims requiring decomposition (Hu et al., 2025) and real-world claims paired with high-quality, web-retrieved evidence that ensures sufficiency and avoids temporal leakage (Schlichtkrull et al., 2023), where adversarial vulnerabilities may manifest differently.

C.2 Evaluation Metrics

C.2.1 Evaluation on Claim Verification

- **FEVER Score** (Thorne et al., 2018b): accuracy for claim verification, conditioned on retrieving at least one complete supporting evidence set for SUP and REF claims;

- **Correctness Rate** (Thorne et al., 2019a): percentage of grammatically correct and coherent claims with correct labels and evidence;

- **Potency** (Thorne et al., 2019a): average reduction (from 1) in FEVER score across FC systems;

- **Adjusted Potency** (Thorne et al., 2019a): weighted average reduction in FEVER score across FC systems, adjusted by correctness rate;

- **Accuracy** (Thorne et al., 2018b): percentage of claims with correctly predicted veracity labels;

- **Resilience** (Thorne et al., 2019a): average FEVER score scaled by correctness rate, where the weights correspond to the correctness rate for each adversary;

- **Macro-F1** (Atanasova et al., 2020): unweighted average of the F1 scores computed for each claim class;

[9] https://www.politifact.com/

- **'→ NEI' (%)** (Abdelnabi and Fritz, 2023): ratio of verdict predictions that change to NEI;
- **Attack Success Rate** (He et al., 2025): proportion of target claims where the attack successfully induces the intended verdict inversion;
- **System Fail Rate** (He et al., 2025): proportion of target claims where the attack causes the fact-checking system to produce an incorrect verdict.

C.2.2 Evaluation on Evidence Retrieval

Document Recall (Kim et al., 2021): percentage of ground-truth documents (e.g., Wikipedia page in FEVER datasets) correctly retrieved by the fact-checking system;

Evidence Recall (Thorne et al., 2019a): percentage of ground-truth evidence sentences correctly retrieved by the fact-checking system;

Adversarial Evidence Recall (Abdelnabi and Fritz, 2023): percentage of adversarial evidence sentences retrieved by the fact-checking system.

Successful Injection Rate (He et al., 2025): proportion of malicious evidences among all retrieved evidences.

C.2.3 Evaluation of Adversarial Claim Quality

- **Perplexity** (Atanasova et al., 2020): perplexity of claims with prepended triggers measures how unpredictable a language model finds the modified claims compared to natural text;
- **Semantic Similarity** (Atanasova et al., 2020): averaged semantic similarity between each claim and associated triggers;
- **Claim quality** (Atanasova et al., 2020): overall quality of generated claims assessed by two independent annotators on a 1-5 scale.

D Figures for Adversarial Attack Overview

We provide Figs. 3 to 5 as complementary for our structured overview of current adversarial attacks against AFC systems in Sec. 3

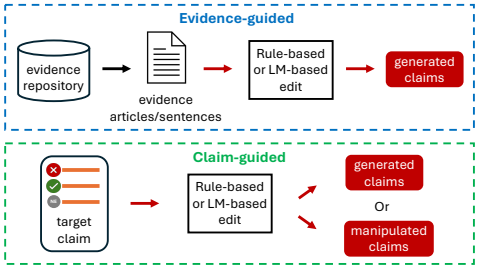

Figure 3: Overview of adversarial claim attacks: evidence-guided and claim-guided.

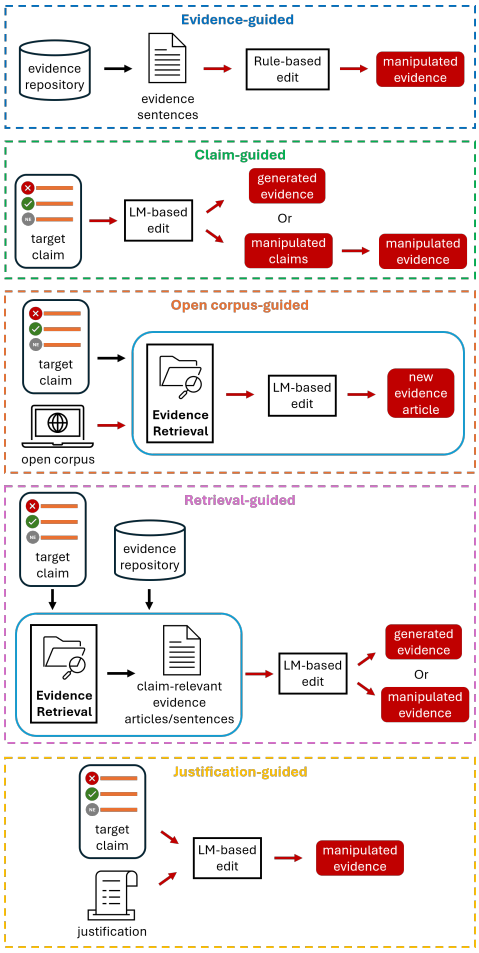

Figure 4: Overview of adversarial evidence attacks: evidence-guided, claim-guided, open corpus-guided, retrieval-guided, and justification-guided.

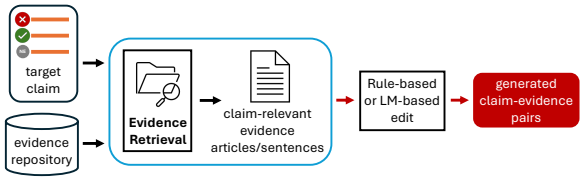

Figure 5: Overview of adversarial claim-evidence pair attacks.

E Characteristics of Adversarial Attacks

To support a more comprehensive comparison of adversarial attacks from both security and technique perspectives, Tables 3 to 5 present key characteristics: direct attack target, method, source dataset (i.e., which dataset a certain attack were performed on), and access setting (black-box vs. white-box) for FC retrieval and verification models. The attack target notation (e.g., "SUP $\rightarrow$ REF") indicates a deliberate verdict shift from the gold label to a target label, while "Generic" signifies no specific target verdict.

F Summary Tables of Adversarial Attacks' Performance

While ensuring a fair comparison across existing studies is challenging due to different settings, we provide summary tables of adversarial attacks, which offer an approximate view of performance differences. For clarity, we focus on evaluations conducted on FEVER 1.0 for both claim attacks (Table 6) and evidence attacks (Table 7). Since the category of evidence–claim attacks comprises only two cases, we report both in Table 2.

Attack Technique	Target Dataset	Test FC System	Acc		Macro-F1	
			Org	Gen	Org	Gen
Symmetric	FEVER 1.0	NSMN	81.8	58.7	–	–
		ESIM	80.8	55.9	–	–
		GPT-3.5-Turbo	86.2	58.3	–	–
GPT-4 Symmetric	CHEF	GPT-4 + DeBERTa	77.34	43.68	75.26	44.57
		mDeBERTa	82.45	53.27	80.68	51.09
		GPT-3.5-Turbo	80.00	53.73	55.25	36.78
		BERT-base	76.35	38.56	75.36	37.62
		Attention-based	78.96	39.98	78.12	39.62
		KGAT	79.55	39.61	76.97	38.67
		Chinese DeBERTa	86.69	57.84	84.98	54.31
		GPT-4-Turbo	85.60	65.20	60.70	47.12

Acc: Accuracy; **Org**: Original dataset pairs; **Gen**: Generated claim-evidence pairs; **–**: This metric was not assessed.
Test FC systems: NSMN (Nie et al., 2019); ESIM (Gardner et al., 2018); GPT-3.5-Turbo (OpenAI, 2022); GPT-4, GPT-4-Turbo (OpenAI, 2024); DeBERTa (He et al., 2021); mDeBERTa (He et al., 2023); BERT (Devlin et al., 2019); Attention-based (Gupta and Srikumar, 2021); KGAT (Liu et al., 2020); Chinese DeBERTa (Zhang et al., 2023).

Table 2: Evaluation of adversarial claim-evidence pair attacks' performance against test FC systems on FEVER 1.0 or CHEF.

Attack Technique	Target Dataset	Attack Settings					Attack Target[1]
		Attack Method	BB-Ret	WB-Ret	BB-Ver	WB-Ver	
Rule-based Approaches							
Model-targeting (Thorne et al., 2019a)	FEVER 1.0	Generate meaning-preserving transformed claims guided by model predictions using Semantically Equivalent Adversarial Rules (SEARs) and (Nie et al., 2019)	✓		✓	✓[2]	SUP → REF REF → SUP
Dataset Bias (Thorne et al., 2019a)	FEVER 1.0	Introduce out-of-distribution patterns into adversarial claims via transformations: entailment preserving rewrites, simple negations, and complex negations	✓		✓		SUP → REF REF → SUP
EntityLess (Kim and Allan, 2019)	FEVER 1.0	Retain only generic entity terms in manipulated claims	✓		✓		Disrupted retrieval
EntityLinking (Kim and Allan, 2019)	FEVER 1.0	Replace entity names in claims with uncommon alternatives	✓		✓		Disrupted retrieval
Controversy (Kim and Allan, 2019)	FEVER 1.0	Generate REF claims associated with conflicting evidence sentence pairs	✓		✓		REF → SUP
NotClear (Kim and Allan, 2019)	FEVER 1.0	Generate NEI claims based on evidence sentences with the phrase "not clear"	✓		✓		NEI → SUP/REF
FiniteSet (Kim and Allan, 2019)	FEVER 1.0	Generate claims guided by temporal reasoning for time-dependent events	✓		✓		SUP → NEI REF → NEI
SubsetNum (Kim and Allan, 2019)	FEVER 1.0	Generate SUP claims guided by subset reasoning over entity relationships	✓		✓		SUP → NEI
NotEnoughInfo Add (Kim and Allan, 2019)	FEVER 1.0	Add NEI claims to balance adversarial classes	✓		✓		NEI → REF/SUP
Conjunction (Hidey et al., 2020)	FEVER 1.0	Generate claims with clauses from single Wikipedia pages	✓		✓		REF → SUP/NEI NEI → SUP/REF
Multi-hop (Hidey et al., 2020)	FEVER 1.0	Generate claims with clauses across diverse Wikipedia pages	✓		✓		Disrupted retrieval
Add. Unver. (Hidey et al., 2020)	FEVER 1.0	Generate NEI claims with additional unverifiable propositions	✓		✓		NEI → REF/SUP
Date Manip. (Hidey et al., 2020)	FEVER 1.0	Manipulate claims with date manipulation heuristics: arithmetic, range, and verbalization	✓		✓		Generic
Multi-hop Temp. (Hidey et al., 2020)	FEVER 1.0	Generate claims with temporal relations linking entities across Wikipedia pages	✓		✓		Disrupted retrieval
Entity Disamb. (Hidey et al., 2020)	FEVER 1.0	Generate claims containing ambiguous entities	✓		✓		Disrupted retrieval
Lexical Sub. (Hidey et al., 2020)	FEVER 1.0	Manipulate claims by applying genetic algorithm to replace synonyms, hypernyms, or hyponyms	✓		✓		Generic
Game-play (Eisenschlos et al., 2021)	FM2	Handcraft plausible but hard claims through a multi-player game	✓		✓		SUP → REF/NEI REF → SUP/NEI
Contractions (Mamta and Cocarascu, 2025)	FEVER 1.0	Replace words in claims with contractions (e.g., *do not* → *don't*)	✓		✓		Generic
Expansions (Mamta and Cocarascu, 2025)	FEVER 1.0	Replace a contraction in claims with its full form (e.g., *don't* → *do not*)	✓		✓		Generic
Jumbling (Mamta and Cocarascu, 2025)	FEVER 1.0	Perturb a claim by randomly changing the order of its words	✓		✓		Generic
Num to Words (Mamta and Cocarascu, 2025)	FEVER 1.0	Convert all numerical digits in claims to their word equivalents (e.g., 2 → *two*)	✓		✓		Generic
Repeat Phrases (Mamta and Cocarascu, 2025)	FEVER 1.0	Append the first quarter of the claim to the end of the original claim	✓		✓		Generic
Subject Verb Disag. (Mamta and Cocarascu, 2025)	FEVER 1.0	Introduce subject-verb agreement errors (i.e., singular vs plural) into claims	✓		✓		Generic
Typos (Mamta and Cocarascu, 2025)	FEVER 1.0	Manipulate claims by swapping adjacent characters	✓		✓		Generic
Word Repetition (Mamta and Cocarascu, 2025)	FEVER 1.0	Duplicate a random selected word in the claim immediately after itself	✓		✓		Generic
Synonyms (Mamta and Cocarascu, 2025)	FEVER 1.0	Manipulate claims by replacing adjectives with synonyms from WordNet	✓		✓		Generic
Tautology (Mamta and Cocarascu, 2025)	FEVER 1.0	Append *and true is true* three times at the end of the claim	✓		✓		Generic
Phonetic Perturb. (Mamta and Cocarascu, 2025)	FEVER 1.0	Apply phonetic perturbations using a human-written dictionary with a word-level perturbation budget	✓		✓		Generic
Character Swapping (Mamta and Cocarascu, 2025)	FEVER 1.0	Randomly swap adjacent characters within a word in the claim	✓		✓		Generic
Character Repetition (Mamta and Cocarascu, 2025)	FEVER 1.0	Randomly duplicate a non-initial/final character within a word in the claim	✓		✓		Generic
Character Insertion (Mamta and Cocarascu, 2025)	FEVER 1.0	Randomly select a non-initial/final character from a word and insert it after the selected character	✓		✓		Generic
Character Deletion (Mamta and Cocarascu, 2025)	FEVER 1.0	Randomly delete a non-initial/final character within a word in the claim	✓		✓		Generic
Homoglyph Perturb. (Mamta and Cocarascu, 2025)	FEVER 1.0	Replace characters with homoglyphs from the Unicode Security dictionary	✓		✓		Generic
LEET Perturb. (Mamta and Cocarascu, 2025)	FEVER 1.0	Replace letters with visually similar symbols using a predefined dictionary	✓		✓		Generic
LM-based Approaches							
Lexically-informed (Thorne et al., 2019a)	FEVER 1.0	Generate claims by replacing nouns and adjectives with WordNet synset lemmas, followed by a neural back-translation paraphrasing model	✓		✓		SUP → REF REF → SUP
Fact Mixing (Niewinski et al., 2019)	FEVER 1.0	Generate claims by controlling GPT-2 for text generation with novel target vocabulary, mixing facts across articles	✓		✓		Generic
Adv. Trigger (Atanasova et al., 2020)	FEVER 1.0	Generate claims using GPT-2 fine-tuned with triggers optimized by extending HotFlip with RoBERTa fine-tuned based on Semantic Textual Similarity	✓		✓		Generic
Colloquial (Kim et al., 2021)	FEVER 1.0	Generate colloquial claims by BART-large fine-tuned on Wizard of Wikipedia	✓		✓		Disrupted retrieval

BB-Ret: Black-box attack without access to the FC retrieval model's architecture; **WB-Ver**: White-box attack with access to the FC retrieval model's architecture; **BB-Ver**: Black-box attack without access to the FC verification model's architecture; **WB-Ver**: White-box attack with access to the FC verification model's architecture; "✓" signifies that the corresponding access setting has been explored for the attack; a blank indicates it has not. **NEI**: NotEnoughInfo; **SUP**: Supported; **REF**: Refuted.

[1] Regarding the target attack, the notation "A → B" denotes a corrupted verdict in which the gold label A is intentionally altered to the target label B, and "Generic" means no specified verdict shift for corrupted verdict.

[2] We indicate white-box verification access for NSMN, one of the test FC models, since the Model-targeting attack leverages its internal predictions to generate adversarial claims.

Table 3: A summary of adversarial claim attack techniques, categorized into rule-based and LM-based approaches.

Attack Technique	Target Dataset	Attack Settings					Attack Target[1]
		Attack Method	BB-Ret	WB-Ret	BB-Ver	WB-Ver	
AdvAdd[2] (Du et al., 2022)	FEVER 1.0 SciFact CovidFact	Add adversarial evidence generated by Grover	✓		✓		Disrupted retrieval
AdvMod (Du et al., 2022)	FEVER 1.0	Create adversarial evidence by PEGASUS for real-time verification	✓		✓		Disrupted retrieval
Imperceptible (Boucher et al., 2022)	FEVER 1.0	Manipulate evidence characters by replacing with homoglyphs, reordering, or deleting them; Minimize correct claim label probability of BERT classifier	✓	✓	✓		SUP → REF REF → SUP
$\text{Imperceptible}_{\text{Ret}}$ (Boucher et al., 2022)	FEVER 1.0	Manipulate evidence characters; Minimize claim-specific evidence ranking score of retrieval model	✓		✓		Disrupted retrieval
Lexical Variation (Alzantot et al., 2018)	FEVER 1.0	Create adversarial claim-paired evidence by RoBERTaBASE	✓		✓		SUP → REF REF → SUP
Contextualized Replace (Li et al., 2020)	FEVER 1.0	Pre-train BERT to replace masked salient words for maximum classification drop	✓	✓	✓		SUP → REF REF → SUP
Omitting Paraphrase (Abdelnabi and Fritz, 2023)	FEVER 1.0	Paraphrase evidence by PEGASUS; Minimize claim-specific retrieval ranking	✓	✓	✓		Disrupted retrieval
Omitting Generate (Abdelnabi and Fritz, 2023)	FEVER 1.0	Fine-tune GPT-2 to create alternative evidence omitting key parts	✓		✓		Disrupted retrieval
Claim-aligned Re-writing (Abdelnabi and Fritz, 2023)	FEVER 1.0	Mask top important tokens by BERT verification model; T5 reconstructs masked supporting evidence; Select Top-K maximizing SUP probability of BERT	✓		✓		REF → SUP
Claim-aligned $\text{Re-writing}_{\text{ret}}$ (Abdelnabi and Fritz, 2023)	FEVER 1.0	Mask tokens for lowest retrieval scores; T5 reconstructs masked supporting evidence; Select evidence maximizing retrieval scores	✓		✓		Disrupted retrieval
Supporting Generation (Abdelnabi and Fritz, 2023)	FEVER 1.0	Fine-tune GPT-2 to generate supporting evidence; Select evidence maximizing SUP probability of BERT	✓		✓		REF → SUP NEI → SUP
Neutral Noise (Samarinas et al., 2021)	FEVER 1.0 Factual-NLI	Add top-scoring entailment-neutral documents retrieved by BM25	✓		✓		Disrupted retrieval
Omission Generation (Atanasova et al., 2022)	FEVER 1.0 VitaminC HoVer	Remove sentence constructs from evidence text and preserve evidence's stance towards the claim	✓		✓		Disrupted retrieval
Fact2Fiction (He et al., 2025)	AVeriTeC	A Planner generates a targeted adversarial answer for each sub-question; An Executor crafts tailored malicious evidence corpora to compromise each sub-question.	✓		✓		SUP → REF REF → SUP

BB-Ret: Black-box attack without access to the FC retrieval model's architecture; **WB-Ver**: White-box attack with access to the FC retrieval model's architecture; **BB-Ver**: Black-box attack without access to the FC verification model's architecture; **WB-Ver**: White-box attack with access to the FC verification model's architecture; "✓" signifies that the corresponding access setting has been explored for the attack; a blank indicates it has not. **NEI**: NotEnoughInfo; **SUP**: Supported; **REF**: Refuted.

[1] Regarding the target attack, the notation "A → B" denotes a corrupted verdict in which the gold label A is intentionally altered to the target label B, and "Generic" means no specified verdict shift for corrupted verdict.

[2] The AdvAdd attack is implemented as the Claim-conditioned Article Generation attack in (Abdelnabi and Fritz, 2023), disrupting the retrieval process.

Table 4: A summary of adversarial evidence attack techniques.

Attack Technique	Target Dataset	Attack Settings					Attack Target[1]
		Attack Method	BB-Ret	WB-Ret	BB-Ver	WB-Ver	
Rule-based Approaches							
Symmetric (Schuster et al., 2019)	FEVER 1.0	Manually generate pairs holding the same relation but expressing a different, contrary, fact	✓		✓		Generic
LM-based Approaches							
GPT-4 Symmetric (Zhang et al., 2024a)	CHEF	Use GPT-4 to generate Chinese pairs holding the same relation but expressing a different, contrary, fact	✓		✓		Generic

BB-Ret: Black-box attack without access to the FC retrieval model's architecture; **WB-Ver**: White-box attack with access to the FC retrieval model's architecture; **BB-Ver**: Black-box attack without access to the FC verification model's architecture; **WB-Ver**: White-box attack with access to the FC verification model's architecture; "✓" signifies that the corresponding access setting has been explored for the attack; a blank indicates it has not.

[1] "Generic" means no specified verdict shift for corrupted verdict.

Table 5: A summary of adversarial claim-evidence pair attack techniques, categorized into rule-based and LM-based approaches.

Attack Technique	Test FC System	FEVER		Acc		Res	Macro-F1	Rec_{Evd}		Rec_{Doc}		Corr[2]	Pot[2]	Adj[2]
		Org	Mod	Org	Mod			Org	Mod	Org	Mod			
Model-targeting	Papelo	73.87	58.26	–	–	58.66[1]	–	–	–	–	–	62.5	57.84	36.15
	NSMN	68.77	47.85	–	–	51.09[1]	–	–	–	–	–			
	HexaF	67.37	50.15	–	–	50.06[1]	–	–	–	–	–			
	Enhanced ESIM	62.76	46.75	–	–	43.98[1]	–	–	–	–	–			
	TF-IDF + ESIM	32.83	28.13	–	–	26.86[1]	–	–	–	–	–			
	TF-IDF + DA	27.73	21.82	–	–	22.28[1]	–	–	–	–	–			
Dataset Bias	Oracle + DA	–	–	82.38	62.56	–	–	–	–	–	–	89.5	63.16	56.53
	Oracle + ESIM	–	–	83.58	60.66	–	–	–	–	–	–			
	Papelo	73.87	56.36	75.58	57.26	58.66[1]	–	71.47	50.00	–	–			
	NSMN	68.77	48.85	70.27	50.35	51.09[1]	–	78.07	71.62	–	–			
	HexaF	67.37	45.25	74.67	51.35	50.06[1]	–	80.78	75.98	–	–			
	Enhanced ESIM	62.76	31.53	67.56	36.84	43.98[1]	–	86.04	80.93	–	–			
	TF-IDF + ESIM	32.83	20.72	49.75	37.34	26.86[1]	–	45.50	39.20	–	–			
	TF-IDF + DA	27.73	18.32	46.49	37.44	22.28[1]	–	45.50	39.20	–	–			
Lexically-informed	Papelo	73.87	44.24	–	–	58.66[1]	–	–	–	–	–	34.0	65.64	22.32
	NSMN	68.77	39.44	–	–	51.09[1]	–	–	–	–	–			
	HexaF	67.37	35.64	–	–	50.06[1]	–	–	–	–	–			
	Enhanced ESIM	62.76	38.24	–	–	43.98[1]	–	–	–	–	–			
	TF-IDF + ESIM	32.83	27.83	–	–	26.86[1]	–	–	–	–	–			
	TF-IDF + DA	27.73	20.72	–	–	22.28[1]	–	–	–	–	–			
EntityLess	FEVER 2.0 Shared Task Systems	–	–	–	–	–	–	–	–	–	–	64.71	79.66	51.54
EntityLinking		–	–	–	–	–	–	–	–	–	–			
Controversy		–	–	–	–	–	–	–	–	–	–			
NotClear		–	–	–	–	–	–	–	–	–	–			
FiniteSet		–	–	–	–	–	–	–	–	–	–			
SubsetNum		–	0.00[3]	–	16.12[3]	–	–	–	–	–	–			
NotEnoughInfo Add		–	76.39[3]	–	76.39[3]	–	–	–	–	–	–			
Conjunction	FEVER 2.0 Shared Task Systems	–	38.25[3]	–	42.50[3]	–	–	–	–	–	–	81.44	68.51	55.79
Multi-hop		–	31.54[3]	–	51.64[3]	–	–	–	–	–	–			
Add Unver		–	55.63[3]	–	55.63[3]	–	–	–	–	–	–			
Date Manip		–	27.53[3]	–	34.18[3]	–	–	–	–	–	–			
Multi-hop Temp		–	8.33[3]	–	24.48[3]	–	–	–	–	–	–			
Entity Disamb		–	–	–	–	–	–	–	–	–	–			
Lexical Sub		–	28.87[3]	–	29.08[3]	–	–	–	–	–	–			
Fact Mixing	FEVER 2.0 Shared Task Systems	–	38.07[3]	–	40.63[3]	–	–	–	–	–	–	84.81	78.80	66.83
Adv Trigger (w STS)	RoBERTa	–	–	–	–	–	63.5[4]	–	–	–	–	–	–	–
Adv Trigger (w/o STS)		–	–	–	–	–	53.4[4]	–	–	–	–			
Colloquial	KGAT (BERT) + Evidence Oracle	–	–	69.7	57.3	–	–	–	–	–	–	–	–	–
	KGAT (BERT) + WikiAPI + BERT	62.4	43.6	67.5	53.2	–	–	85.3	73.4	90.0	72.2			
	KGAT (BERT) + DPR + BERT	55.4	41.5	62.9	51.2	–	–	81.8	77.4	84.0	79.6			
	KGAT (CorefBERT) + Evidence Oracle	–	–	77.5	67.7	–	–	–	–	–	–			
	KGAT (CorefBERT) + WikiAPI + BERT	69.5	52.4	73.8	60.9	–	–	85.3	73.4	90.0	72.2			
	KGAT (CorefBERT) + DPR + BERT	52.4	55.4	61.1	61.0	–	–	81.8	77.4	84.0	79.6			

Acc: Accuracy; **Res**: Resilience; $\textbf{Rec}_{\textbf{Evd}}$: Evidence Sentence Recall; $\textbf{Rec}_{\textbf{Doc}}$: Document Recall; **Pot:** Potency; **Corr:** Correctness Rate; **Adj**: Adjusted Potency; **Org**: Results on original target claims; **Mod**: Results on modified claims; **STS**: Semantic Textual Similarity; **DPR**: Dense Passage Retrieval; **–**: This metric was not assessed.
Test FC systems: Papelo (Malon, 2018); NSMN (Nie et al., 2019); HexaF (Yoneda et al., 2018); Enhanced ESIM, WikiAPI (Hanselowski et al., 2018); TF-IDF (Chen et al., 2017); ESIM (Gardner et al., 2018); DA (Parikh et al., 2016); FEVER 2.0 Shared Task Systems (Thorne et al., 2019b); RoBERTa (Liu et al., 2019); KGAT (Liu et al., 2020); BERT (Devlin et al., 2019), CorefBERT (Ye et al., 2020); DPR (Karpukhin et al., 2020).

[1] Statistics for Resilience are with a mixture of adversarial claims generated by Model-targeting, Dataset Bias, and Lexically-informed attacks across test FC systems.

[2] In (Kim and Allan, 2019), (Hidey and Diab, 2018), and (Niewinski et al., 2019), statistics for Corr, Pot, and Adj are with their respective attacks combined.

[3] Statistics for average FEVER and Acc are reported for adversarial claims annotated as "correct" and attack types with over five generated claims, referring to (Thorne et al., 2019b).

[4] Statistics for Macro-F1 score are with verification results over generated claims.

Table 6: Evaluation of adversarial claim attacks' performance against test FC systems on FEVER 1.0.

Attack Technique	Test FC System	Acc on All		Acc on SUP		Acc on REF		Acc on NEI		Macro-F1		$\text{Rec}_{\text{Evd}}^{\text{Adv}}$	$\text{Rec}_{\text{Evd}}^{\text{Gold}}$	→ NEI
		Org	Mod	Org	Mod	Org	Mod	Org	Mod	Org	Mod			
AdvAdd	KGAT	70.76	29.02	–	–	72.50	42.45	69.01	15.59	–	–	–	–	–
	CorefBERT	73.05	28.59	–	–	74.03	39.63	72.07	17.54	–	–	–	–	–
	MLA	75.92	51.86	–	–	78.71	71.84	73.13	31.87	–	–	–	–	–
AdvMod	KGAT	70.76	37.22	–	–	72.50	51.74	69.01	22.70	–	–	–	–	–
	CorefBERT	73.05	32.62	–	–	74.03	36.66	72.07	28.58	–	–	–	–	–
	MLA	75.92	52.72	–	–	78.71	70.57	73.13	34.86	–	–	–	–	–
Imperceptible														
Homoglyph ($\epsilon = 5$)	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	89.0	39.6	71.2	50.3	–	–	–	–	55.2	–	83.6
	$\text{CorefBERT}_{\text{BASE}}$	–	–	87.5	39.9	72.8	52.5	–	–	–	–	–	–	–
	KGAT ($\text{RoBERTa}_{\text{LARGE}}$)	–	–	91.5	60.6	74.7	60.1	–	–	–	–	–	–	–
	$\text{CorefRoBERTa}_{\text{LARGE}}$	–	–	92.2	65.1	77.5	60.7	–	–	–	–	–	–	–
Reorder ($\epsilon = 5$)	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	89.0	37.8	71.2	49.5	–	–	–	–	55.1	–	81.8
	$\text{CorefBERT}_{\text{BASE}}$	–	–	87.5	37.4	72.8	52.3	–	–	–	–	–	–	–
	KGAT ($\text{RoBERTa}_{\text{LARGE}}$)	–	–	91.5	47.7	74.7	54.8	–	–	–	–	–	–	–
	$\text{CorefRoBERTa}_{\text{LARGE}}$	–	–	92.2	49.9	77.5	51.4	–	–	–	–	–	–	–
Delete ($\epsilon = 5$)	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	89.0	38.9	71.2	49.7	–	–	–	–	60.5	–	79.4
	$\text{CorefBERT}_{\text{BASE}}$	–	–	87.5	39.2	72.8	52.5	–	–	–	–	–	–	–
	KGAT ($\text{RoBERTa}_{\text{LARGE}}$)	–	–	91.5	60.8	74.7	60.7	–	–	–	–	–	–	–
	$\text{CorefRoBERTa}_{\text{LARGE}}$	–	–	92.2	66.1	77.5	59.8	–	–	–	–	–	–	–
$\text{Imperceptible}_{\text{Ret}}$														
Homoglyph ($\epsilon = 5$)	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	89.0	62.3	71.2	60.5	–	–	–	–	31.5	–	88.9
Lexical Variation	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	89.0	68.9	71.2	65.4	–	–	–	–	42.1	–	73.6
	$\text{CorefBERT}_{\text{BASE}}$	–	–	87.5	67.7	72.8	66.6	–	–	–	–	–	–	–
	KGAT ($\text{RoBERTa}_{\text{LARGE}}$)	–	–	91.5	73.6	74.7	69.9	–	–	–	–	–	–	–
	$\text{CorefRoBERTa}_{\text{LARGE}}$	–	–	92.2	74.5	77.5	71.6	–	–	–	–	–	–	–
Contextualized Replace	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	89.0	50.7	71.2	59.7	–	–	–	–	30.3	–	69.3
	$\text{CorefBERT}_{\text{BASE}}$	–	–	87.5	50.0	72.8	60.8	–	–	–	–	–	–	–
	KGAT ($\text{RoBERTa}_{\text{LARGE}}$)	–	–	91.5	55.8	74.7	63.9	–	–	–	–	–	–	–
	$\text{CorefRoBERTa}_{\text{LARGE}}$	–	–	92.2	55.6	77.5	64.0	–	–	–	–	–	–	–
Omitting Paraphrase	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	89.0	51.0	71.2	54.3	–	–	–	–	54.4	–	83.8
	$\text{CorefBERT}_{\text{BASE}}$	–	–	87.5	50.7	72.8	55.8	–	–	–	–	–	–	–
	KGAT ($\text{RoBERTa}_{\text{LARGE}}$)	–	–	91.5	56.8	74.7	60.7	–	–	–	–	–	–	–
	$\text{CorefRoBERTa}_{\text{LARGE}}$	–	–	92.2	55.5	77.5	58.4	–	–	–	–	–	–	–
Omitting Generate	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	89.0	29.9	71.2	46.8	–	–	–	–	30.9	–	87.9
	$\text{CorefBERT}_{\text{BASE}}$	–	–	87.5	30.2	72.8	48.9	–	–	–	–	–	–	–
	KGAT ($\text{RoBERTa}_{\text{LARGE}}$)	–	–	91.5	33.8	74.7	51.9	–	–	–	–	–	–	–
	$\text{CorefRoBERTa}_{\text{LARGE}}$	–	–	92.2	31.6	77.5	47.4	–	–	–	–	–	–	–
Claim-aligned Re-writing	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	–	–	71.2	38.4[2]	–	–	–	–	94.4[2]	–	1.8[2]
	$\text{CorefBERT}_{\text{BASE}}$	–	–	–	–	72.8	36.9	–	–	–	–	–	–	–
	KGAT ($\text{RoBERTa}_{\text{LARGE}}$)	–	–	–	–	74.7	44.9	–	–	–	–	–	–	–
	$\text{CorefRoBERTa}_{\text{LARGE}}$	–	–	–	–	77.5	42.1	–	–	–	–	–	–	–
Claim-aligned $\text{Re-writing}_{\text{ret}}$	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	–	–	71.2	43.7[2]	–	–	–	–	99.1[2]	–	1.8[2]
	$\text{CorefBERT}_{\text{BASE}}$	–	–	–	–	72.8	43.1	–	–	–	–	–	–	–
	KGAT ($\text{RoBERTa}_{\text{LARGE}}$)	–	–	–	–	74.7	48.8	–	–	–	–	–	–	–
	$\text{CorefRoBERTa}_{\text{LARGE}}$	–	–	–	–	77.5	47.6	–	–	–	–	–	–	–
Supporting Generation	KGAT ($\text{BERT}_{\text{BASE}}$)	–	–	–	–	71.2	42.0[2]	72.4	32.2[2]	–	–	85.7[2]	–	3.8[2]
	$\text{CorefBERT}_{\text{BASE}}$	–	–	–	–	72.8	39.7	72.8	34.5	–	–	–	–	–
	KGAT ($\text{RoBERTa}_{\text{LARGE}}$)	–	–	–	–	74.7	43.2	68.8	25.6	–	–	–	–	–
	$\text{CorefRoBERTa}_{\text{LARGE}}$	–	–	–	–	77.5	45.2	70.0	30.0	–	–	–	–	–
Omission Generation	BERT	–	–	–	–	–	–	–	–	87.16[3]	59.51[3]	–	–	–
	RoBERTa	–	–	–	–	–	–	–	–	88.69[3]	59.10[3]	–	–	–
	ALBERT	–	–	–	–	–	–	–	–	86.67[3]	63.00[3]	–	–	–
	Ensemble[1]	–	–	–	–	–	–	–	–	88.81[3]	61.36[3]	–	–	–
Neutral Noise	BM25 retriever	–	–	–	–	–	–	–	–	–	–	–	35.17[4]	–
	QR-BERT retriever	–	–	–	–	–	–	–	–	–	–	–	54.10[4]	–

Acc: Accuracy; $\textbf{Rec}_{\textbf{Evd}}^{\textbf{Adv}}$: Recall of adversarial evidence; $\textbf{Rec}_{\textbf{Evd}}^{\textbf{Gold}}$: Recall of gold evidence after attack; **→ NEI**: the ratio of verdict predictions that shift to NEI;
Org: Original dataset (before attack); **Mod**: Modified dataset (after attack); **–**: This metric was not assessed.
Test FC systems: KGAT (Liu et al., 2020); BERT (Devlin et al., 2019), CorefBERT, CoreRoBERTa (Ye et al., 2020); MLA (Kruengkrai et al., 2021); RoBERTAa (Liu et al., 2019); ALBERT (Lan et al., 2020); BM25 (Robertson et al., 1994); QR-BERT (Samarinas et al., 2020).

[1] An ensemble of three FC models, i.e., BERT, RoBERTa, and ALBERT, in (Atanasova et al., 2022).

[2] Statistics for Acc are reported under stance filtering for both Claim-aligned Re-writing and Supporting Generation attacks, and under retrieval filtering for the Claim-aligned $\text{Re-writing}_{\text{Ret}}$ attack.

[3] Statistics for Macro-F1 score on claim verification results before and after attack.

[4] Statistics for Recall@5 on gold evidence retrieval with adversarial evidence included.

Table 7: Evaluation of adversarial evidence attacks' performance against test FC systems on FEVER 1.0.