Half Truth Detection and Mitigation: A Survey

Satyam Shukla **Pushpak Bhattacharyya**
Computation for Indian Language Technology
Department of Computer Science and Engineering
Indian Institute of Technology Bombay, India
{satyamshukla, pb}@cse.iitb.ac.in

Abstract

In the digital age, misinformation has taken complex forms, including half-truths—statements that are factually accurate but deliberately omit critical context, leading to deception. Unlike outright fake news, half-truths are subtler and harder to detect, often blending facts with misleading implications. This survey explores recent research advances in the detection and mitigation of half-truths. We analyze foundational studies on the linguistic and computational complexity of half-truths, as well as emerging approaches using machine learning, large language models (LLMs), and claim rewriting. In addition, we present a detailed comparison of datasets designed for fine-grained veracity detection, particularly those annotating partial truths. The survey also highlights recent mitigation strategies such as claim editing and reinforcement learning-based debunking systems. We conclude by identifying open challenges and potential future directions to combat half-truths in automated fact-checking systems.

1 Introduction

In today's digitally connected world, the spread of deceptive information has reached alarming levels. While considerable attention has been paid to fake news and outright lies, a subtler and more insidious form of misinformation has emerged—**half-truths**. Unlike completely fabricated claims, half-truths present factual information in isolation or omit critical context, leading to distorted public understanding. These statements often evade traditional fact-checking systems due to their partial correctness, making them more persuasive and harder to detect.

Half-truths are especially dangerous because they blend truth with deception, creating a false sense of credibility. For example, the claim, "100% donated to charity," sounds trustworthy but becomes misleading without clarifying whether the 100% refers to gross income, net profit, or some other basis. Similarly, political statements like "The GOP budget would cut nearly a million jobs" may stem from speculative economic models but are framed with unwarranted certainty to induce fear or support. These statements are not entirely false, but their framing misguides the audience.

The proliferation of half-truths poses a significant threat to public trust, democratic integrity, and informed decision-making. Social media platforms amplify these messages, exploiting confirmation bias and echo chambers to rapidly disseminate misleading content. Moreover, half-truths are often employed deliberately as propaganda tools, allowing entities—political, commercial, or ideological—to manipulate narratives while retaining plausible deniability.

Recent advancements in machine learning, particularly the emergence of **Large Language Models (LLMs)** and evidence-aware reasoning techniques, have opened new avenues for addressing this complex problem. Unlike binary fact-checking systems that categorize claims as simply "true" or "false," modern approaches aim to detect nuances in veracity, including labels like *mostly true*, *half true*, and *mostly false*. This fine-grained classification is crucial for detecting half-truths that lie between factual and fabricated content.

Researchers have proposed diverse pipelines that go beyond claim verification to include evidence retrieval, semantic editing, and explanation generation. For instance, models such as CofCED (Yang et al., 2022) employ evidence distillation for accurate fact classification, while claim rewriting techniques seek to mitigate the misleading nature of half-truths by generating fully truthful alternatives.

Furthermore, **temporal leakage prevention** has emerged as a key concern in dataset construction. Datasets like WatClaimCheck (Khan et al., 2022), LIAR-RAW, and RAW-FC address this by ensuring that evidence predates the claim publication,

providing a more realistic setup for training and evaluation.

This survey explores the landscape of half-truth detection and mitigation through three core lenses:

1. **Detection Techniques**: We cover supervised, unsupervised, and prompt-based methods to classify partial truths.

2. **Mitigation Strategies**: This includes claim editing, summarization, and explanation generation to correct or contextualize misleading content.

3. **Datasets and Evaluation**: We compare datasets that support fine-grained veracity labels and discuss challenges in annotation, temporal alignment, and originates from multiple domains.

By examining current literature, datasets, and techniques, we aim to present a comprehensive picture of the field and highlight emerging directions. Half-truth detection remains a highly relevant and underexplored subdomain of misinformation research. As misinformation tactics grow more sophisticated, the ability to distinguish and mitigate partial truths will become increasingly central to maintaining public trust and societal resilience.

2 Motivation

Manual detection of misinformation, especially half-truths, is not scalable in the face of the sheer volume and speed of online content generation. While fully false claims can be debunked with known facts, partially true claims require contextual and semantic understanding of what's omitted or distorted.

Automated systems offer the promise of real-time intervention, especially when enhanced with retrieval-augmented generation (RAG), multi-hop reasoning, or claim rewriting. Additionally, editing-based approaches show promise for not just identifying but also correcting misleading claims by rephrasing them into fully truthful versions, using evidence as guidance.

Recent research emphasizes the importance of this dual capability—detection and mitigation—in fact-checking pipelines. Claim editing, when controlled and evidence-guided, can be deployed to counter misinformation with revised content that retains semantic structure but restores truthfulness.

3 Problem Statement

Detection Task: Given a claim C and corresponding evidence E, classify C into one of the veracity classes: True, Half-True, or False. This is framed as a multi-class classification task.

Mitigation Task (Claim Editing): Given a claim C labelled as Half-True or False, along with evidence E, generate a new claim C^* that is fully true and semantically faithful. The task involves controlled text generation, aiming to preserve relevant parts while editing the misleading portions based on evidence.

These two tasks reflect the complementary goals of truth detection and information correction, both of which are essential to automated fact-checking systems. The complexity of this pipeline is illustrated in Figure 1.

4 Background and Terminologies

True and Mostly True Information. A **true** statement is entirely accurate, supported by reliable evidence, and presented with sufficient context to avoid misinterpretation. In contrast, a **mostly true** statement is generally accurate but may omit minor details or require slight clarification.

Example (True): "The Earth revolves around the Sun once every 365.25 days." This is a well-established scientific fact, requiring no further context.

Example (Mostly True): "India is the world's largest democracy." This is broadly true based on population and electoral participation, though some political and systemic nuances may be omitted.

Half-Truth is Not the Same as Half-Lie. A **half-truth** is a statement that is factually correct in parts but misleading due to omission of crucial context or selective presentation. It is not the same as a **half-lie**, which implies a literal mixture of truth and falsehood.

Example (Half-Truth): "The minister attended the relief camp." While the minister may have visited, the statement is misleading if they stayed only briefly for a photo opportunity without engaging meaningfully. The omission of this detail distorts the truth.

Key distinction: A half-truth *omits* context to mislead, whereas a half-lie *adds* falsehood to mix with truth.

Mostly False and False Information. Mostly false statements contain a small grain of truth but

are largely inaccurate. **False** statements are entirely untrue and unsupported by evidence.

Example (Mostly False): "Vaccines often cause the diseases they are meant to prevent." This exaggerates rare side effects, misrepresenting the broader medical consensus.

Example (False): "The Moon is made of cheese." This is entirely fabricated with no basis in reality.

Leakage. Leakage refers to the use of evidence that was published *after* the claim was made. This creates a temporal inconsistency, as future knowledge would not have been available to verify the claim at the time it was originally encountered. Models trained or evaluated with such leaked evidence may achieve artificially high performance by relying on information that violates the chronological integrity of real-world fact-checking.

Example: If a claim was made on January 1, 2020, and the evidence used to verify it includes an article from March 2020, this would be considered leakage. Such setups do not reflect real-time fact-checking scenarios and can mislead evaluation results.

Evidence Access Settings. Researchers typically categorize evidence usage into three main settings: open, closed, and gold. In the open setting, the system actively retrieves supporting information or answers using external tools such as search engines or APIs. The closed setting relies solely on the language model's internal (parametric) knowledge to generate responses without external input. In the gold setting, the relevant evidence is provided in advance, and the model's task is to evaluate or reason over this given context in relation to the claim.

5 Related Work

In this section, we review the body of literature that forms the basis for research in half-truth detection and mitigation. We categorize the work into foundational research, which established the theoretical or architectural basis, and competitive/modern approaches, which push the boundary in detection and debunking.

Deception via Half-Truths. Estornell et al. (2020) introduced a game-theoretic analysis of deception by omission in their paper "Deception through Half-Truths." They demonstrated that half-truths can be computationally more deceptive than outright lies, especially in decision-making systems. The paper mathematically proves that finding an optimal half-truth attack is NP-hard, motivating the need for specialized detection algorithms that go beyond surface-level fact verification.

Linguistic Cues and Deceptive Speech. (Hazra and Majumder, 2024) curated the T4TEXT dataset from transcripts of a deception-based game show to train LLMs on recognizing deceptive cues including half-truths. Their work shows that models can identify ambiguity, omission, and semantic evasion—hallmarks of half-truths—better than human judges. This work underlines the growing capability of language models to parse nuanced veracity.

5.1 Text-Based Datasets

Many text-based datasets focus on Wikipedia as a single source of truth. For example, Schuster et al. (2021) relied solely on Wikipedia, which, while useful, fails to capture misinformation spread beyond that domain. Datasets such as HOVER (Jiang et al., 2020) and FEVER (Thorne et al., 2018) similarly use only Wikipedia as their knowledge base. Although these datasets provide large-scale examples, they limit their scope to a single source and ignore the complexities of real-world fact-checking, which often involves information from multiple, disparate sources.

To address these limitations, other datasets have been proposed that incorporate evidence from a broader range of real-world sources. These include works by (Hanselowski et al., 2019), (Wadden et al., 2020), (Kotonya and Toni, 2020), and (Vlachos and Riedel, 2014), which provide claims grounded in natural domains beyond politics. However, many of these datasets are either limited in scope or provide fewer instances of claims from political domains, reducing their effectiveness for comprehensive fact-checking for the political domain.

A significant limitation of Alhindi et al. (2018) is its failure to capture the complete evidence for a claim from its source. In case human justification is not present, it uses the last 5 sentences of the source article. This shortcoming limits the dataset's ability to support accurate fact-checking in complex cases.

In contrast, Misra (2022) introduced a dataset that includes only the claim and metadata from PolitiFact, without incorporating the corresponding retrieved evidence. Building on this, we propose a more comprehensive **PolitiFact-PLUS** dataset, which includes not only claims and meta-

data but also the complete evidence from the original sources. This enriched dataset provides a broader and more detailed basis for fact-checking, making it highly suitable for testing machine learning models on real-world claims across various domains.

5.2 Competitive and Modern Approaches

Half-Truth Detection Pipelines. Singamsetty et al. (2023) proposed an end-to-end pipeline for half-truth detection using the LIAR-PLUS dataset. They developed a BERT-based classification system for multi-class veracity labeling (True, Half-True, False), achieving state-of-the-art performance in the half-true class. Their work integrates evidence extraction with justification-aware modeling, offering a practical foundation for automated fact-checking pipelines.

Claim Editing for Mitigation. Singamsetty et al. (2023) implemented a claim editing model using T5, which rewrites misleading claims into truthful ones based on evidence. This is a controlled generation task that maintains meaning while correcting misinformation. Their method achieves high BLEU and semantic similarity scores, establishing claim rewriting as a strong candidate for half-truth mitigation.

Supervised Approaches. Supervised learning remains the dominant paradigm for automated fact-checking, particularly when ample labeled data is available. These approaches typically involve a pipeline of claim classification, evidence retrieval, and veracity labeling trained on datasets like FEVER or LIAR. One notable example is CofCED by Yang et al. (2022), which introduces a coarse-to-fine evidence distillation framework. CofCED first retrieves a broad set of candidate evidence using dense retrieval, then applies a fine-grained refinement stage to filter only the most relevant pieces before classification. This hierarchical approach reduces noise and improves claim verification accuracy.

Beyond CofCED, Guo et al. (2022) surveyed various supervised fact-checking architectures, emphasizing the importance of multi-hop reasoning and attention-based mechanisms. Another contribution by Yoneda et al. (2018) proposed a pipeline that combines document retrieval with sentence-level entailment classification using supervised learning, highlighting how effective evidence segmentation plays a critical role in downstream prediction. These approaches establish a foundation for claim verification but often assume high-quality, static evidence—which becomes problematic in temporally evolving scenarios.

Leakage Prevention. Preventing information leakage, particularly **temporal leakage**, is essential to ensure that fact-checking models mirror real-world conditions. Temporal leakage occurs when the evidence used for verification was published after the claim itself, which would not have been available at the time of fact-checking.

Khan et al. (2022) introduced the **WatClaimCheck** dataset, which links each claim to the original *premise article* that the fact-checkers consulted when producing the verdict. This makes the task more realistic by using the actual reference materials used in professional verification. To go a step further, Yang et al. (2022) curated the **LIAR-RAW** and **RAW-FC** datasets. These resources include only web documents that were publicly available **before the claim was published**, strictly excluding fact-checking reports or articles published after the claim date. By relying solely on contemporaneous evidence, these datasets provide a more temporally accurate setting for evaluating verification models. This setup ensures that models mimic the constraints of real-world fact-checkers who must assess a claim without access to future information or post-verdict commentary.

Additionally, Spangher et al. (2021) presented a framework that models evidence sourcing as a multitask learning problem, where one task involves identifying sources within a temporal window relevant to the claim. These works collectively emphasize the critical role of chronology in maintaining fact-checking fidelity and preventing artificial performance inflation.

Reasoning and Prompting Strategies. Recent advancements in large language models (LLMs) have led to growing interest in prompt-based and multi-agent reasoning systems for fact verification. A key example is **HiSS** (Hierarchical Self-Supervision) proposed by Zhang and Gao (2023), where LLMs are fine-tuned with intermediate reasoning supervision, enabling step-by-step decomposition of fact-checking tasks. Another notable system is **FIRE** (*Fact-checking with Iterative Retrieval and Evidence*) introduced by Xie et al. (2025), which combines prompting with retrieval to iteratively refine evidence selection before an-

swering. FIRE demonstrates improved robustness and factuality across several benchmarks.

SAFE proposed by Wei et al. (2024) operates by progressively generating web search queries and then assessing whether the gathered evidence collectively supports the given claim. However, it separates evidence retrieval from verification and enforces a fixed number of searches, irrespective of how complex or simple the claim might be, showing limited flexibility. In a similar vein, **Factool** by Chern et al. (2023) uses a plug-and-play prompting framework to evaluate the factual consistency of LLM-generated responses. It leverages retrieval-augmented generation and a verifier agent to rank and cross-check claims. Moreover, Cohen and Manning (2023) proposed a cross-examination method where two LLMs interact: one acts as a respondent generating factual statements, and the other as a verifier that challenges and refines the initial output. This adversarial prompting strategy enhances factual consistency and reduces hallucinations.

Other relevant prompting-based reasoning techniques include **Self-Ask** (Press et al., 2022), which allows LLMs to pose sub-questions before final judgment, and **Chain-of-Verification** (Liang et al., 2023), which models verification as a sequence of entailment steps. These frameworks are especially relevant for half-truth detection, where nuanced reasoning and context understanding are essential to flag omissions and distortions.

Counter-Misinformation via Response Generation. He et al. (2023) proposed MisinfoCorrect, a reinforcement-learning (RL) framework to generate ideal corrective responses for social media misinformation. They optimize for factuality, politeness, and refutation tone. Though not explicitly about half-truths, it provides a mitigation strategy to counter misleading narratives that blend truth with distortion.

Query Rewriting for Evidence Retrieval. Kazemi et al. (2023) introduced a novel RL-based query rewriting system to help retrieve better evidence for vague or misleading claims. The model learns how to replace or rephrase ambiguous parts of a query (often due to half-truth phrasing), improving retrieval effectiveness. This can be used upstream in any fact-checking pipeline.

6 Datasets

Fact-checking datasets differ widely in their structure, quality of evidence, and susceptibility to leakage, which can significantly affect their applicability to real-world verification tasks. Leakage refers to unintended cues within the evidence—such as annotator commentary, explicit verdict indicators, or biased language—that reveal the label and thus compromise the realism of automated claim verification. Early datasets like **LIAR** (Wang, 2017) and **PunditFact** (Rashkin et al., 2017) provide only meta-information or brief snippets without any explicit supporting evidence, limiting their value for evidence-based fact-checking models. On the other hand, synthetic datasets such as **FEVER** (Thorne et al., 2018), **VitaminC** (Schuster et al., 2021), and **HOVER** (Jiang et al., 2020) are constructed using curated or manipulated content from Wikipedia, ensuring minimal leakage and supporting more controlled experiments in claim verification.

In contrast, natural datasets often incorporate evidence from less curated or real-world sources, increasing the likelihood of leakage. For example, **MultiFC** (Augenstein et al., 2019) and **X-Fact** (Gupta and Srikumar, 2021) aggregate claims and evidence from multiple fact-checking websites, where verdict-related cues—such as labels embedded in headlines or evaluator commentary—are frequently entangled with the evidence text. Similarly, **LIAR-PLUS** (Alhindi et al., 2018), an extension of **LIAR**, further amplify the scope of textual evidence and introduce entailment-based variations of claims, but often inherit the same leakage issues from the original meta-sourced annotations. Datasets like **RU22fact** (Zeng et al., 2024), based on Russian political claims, also exhibit label leakage due to embedded verdict cues in the retrieved evidence. Further, Russo et al. (2023) proposed **LIAR++**, an extension of LIAR-PLUS (Alhindi et al., 2018) and **FullFact**, both dataset contains claim, corresponding evidence and verdict labels. The **LIAR++** dataset was developed by refining the LIAR-PLUS dataset, which originally consisted of political claims and corresponding articles sourced from PolitiFact between 2007 and 2016. In contrast to LIAR-PLUS, LIAR++ excludes verdicts that were artificially constructed and instead retains only high-quality, gold-standard verdicts extracted from specific sections like "Our ruling." It also preserves explicit verdict language that was previously removed. As a result, LIAR++ comprises 6,451

Dataset	Type	Domain	#Claims	Evidence Type	#Classes
HOVER (Jiang et al., 2020)	Synthetic	Multiple	26,171	Text	2
FEVER (Thorne et al., 2018)	Synthetic	Multiple	185,445	Text	3
VitaminC (Schuster et al., 2021)	Synthetic	Multiple	488,904	Text	3
PunditFact (Rashkin et al., 2017)	Natural	Multiple	4,361	Meta	2/6
Snopes (Hanselowski et al., 2019)	Natural	Multiple	6,422	Text	3
SciFact (Wadden et al., 2020)	Natural	Science	1,409	Text	3
PUBHEALTH (Kotonya and Toni, 2020)	Natural	Health	11,832	Text	4
PolitiFact (Vlachos and Riedel, 2014)	Natural	Politics	106	Text	5
LIAR (Wang, 2017)	Natural	Politics	12,836	Meta	6
LIAR-PLUS (Alhindi et al., 2018)	Natural	Politics	12,836	Both	6
LIAR++ (Russo et al., 2023)	Natural	Politics	6,451	Both	3
FullFact (Russo et al., 2023)	Natural	Multiple	1,838	Both	3
AnswerFact (Zhang et al., 2020)	Natural	Product	60,864	Both	5
T4TEXT (Hazra and Majumder, 2024)	Natural	Dialogue	150 episodes	Text	2
HAD (Yi et al., 2021)	Natural	Audio	88035	Text	3
COVID Misinfo (Ou et al., 2022)	Natural	Health	14,384	Text	3

Table 1: Comparison of major datasets used in claim verification, misinformation detection, and evidence-based fact-checking.

claim-article-verdict instances. Similarly, the FullFact dataset was curated from the FullFact website, covering a broader range of domains such as health, law, and education, with data spanning from 2010 to 2021. Unlike LIAR++, this dataset required minimal filtering since verdicts were consistently provided as distinct components on each web page, resulting in 1,838 high-quality triples. Importantly, these datasets provide fine-grained half-truth annotations, allowing for nuanced multi-class classification beyond simple binary verdicts, which is crucial for understanding misinformation on a spectrum.

The broader landscape of datasets includes both synthetic and naturally occurring claim scenarios across diverse domains. While synthetic datasets like FEVER, VitaminC, and HOVER emphasize curated evidence to support scalable training, natural datasets cover real-world claim types: LIAR, PolitiFact (Vlachos and Riedel, 2014), and its variants focus on political discourse; PUBHEALTH (Kotonya and Toni, 2020) and **SciFact** (Wadden et al., 2020) target health and scientific misinformation, respectively; **AnswerFact** (Zhang et al., 2020) captures consumer product claims in QA contexts; and **T4TEXT** (Hazra and Majumder, 2024) addresses deceptive strategies in dialogue, including omission and ambiguity. Additionally, Yi et al. (2021) published the Half-truth Audio Detection (HAD) dataset, which is built upon the publicly available AISHELL-3 corpus (Yao Shi, 2015). The corpus comprises 88,035 utterances totalling approximately 85 hours of speech from 218 native Mandarin speakers (175 female, 43 male). While the COVID Misinfo Dataset (COVMIS) (Ou et al., 2022) is a dataset compiled between November 2019 and March 2021, consisting of 14,384 claims, 134,320 associated articles, and various metadata attributes. These include details such as the claimant, source, publication date, assigned truth label (true, partly true, or false), and corresponding justification. Each claim is linked to a set of relevant articles gathered from credible sources, which are used as reference material to evaluate the factual accuracy of the claims.

7 Conclusion and Future Work

Half-truth detection and mitigation represent a critical and evolving area within the broader fact-checking and misinformation detection landscape. As this survey highlights, half-truths are particularly challenging because they blend factual accuracy with contextual omissions or distortions, making them harder to detect than outright false claims. Existing approaches—including supervised learning pipelines, multi-agent prompting strategies, and claim rewriting techniques—have shown considerable promise. However, most solutions remain limited in terms of domain generalization, fine-grained veracity classification, and scalable mitigation strategies.

Recent advancements in large language mod-

els (LLMs), evidence-aware reasoning frameworks, and structured claim decomposition have pushed the boundaries of what automated fact-checking systems can achieve. Despite these improvements, several open challenges remain. Current datasets suffer from temporal leakage, limited domain coverage, and incomplete evidence representation, making it difficult to train truly robust systems. Moreover, while prompt-based reasoning has introduced interpretability into fact-checking, these approaches often rely on handcrafted prompts or domain-specific tuning, reducing their generalizability across different misinformation scenarios.

Looking forward, future research should focus on developing robust, domain-agnostic pipelines capable of decomposing complex claims into structured reasoning steps, applicable across multiple domains and languages. Improved datasets with temporally aligned evidence, richer annotations for fine-grained labels such as "half-true" and "mostly false," and context-aware evidence retrieval mechanisms are needed to support these efforts. Additionally, interactive multi-agent systems that integrate claim questioning, evidence verification, and mitigation (such as claim rewriting or explanation generation) in a seamless pipeline present a promising direction.

Another critical area for future work is the mitigation of half-truths. Rather than merely detecting misleading claims, systems should be able to automatically generate corrected or contextually complete claims, preserving the truthful elements while removing ambiguity and deception. Reinforcement learning and human-in-the-loop strategies can further enhance the quality and ethical considerations of these mitigation processes.

Finally, the real-world deployment of half-truth detection systems will require addressing user trust, explainability, and integration with content moderation workflows on social media and news platforms. Future systems must not only be technically sound but also socially responsible, ensuring transparency in how veracity judgments are made and presented to users.

References

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, Brussels, Belgium. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Karmel Cohen and Christopher D Manning. 2023. Lm cross-examination: A two-agent framework for verifying generated content. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. 2020. Deception through half-truths. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10110–10117.

Yujing Guo, Wenjie Wang, Ming Tan, Wayne Xin Li, and Min-Yen Kan. 2022. A survey on automated fact-checking. In *ACM Computing Surveys*.

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, Online. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Sanchaita Hazra and Bodhisattwa Prasad Majumder. 2024. To tell the truth: Language of deception and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 8498–8512. Association for Computational Linguistics.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: A case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*.

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

Ashkan Kazemi, Artem Abzaliev, Naihao Deng, Rui Hou, Scott A. Hale, Veronica Perez-Rosas, and Rada Mihalcea. 2023. Query rewriting for effective misinformation discovery. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 398–407, Nusa Dua, Bali. Association for Computational Linguistics.

Moin Khan, M Saad Qureshi, Shafiq Joty, and Preslav Nakov. 2022. Watclaimcheck: Temporal-aware evidence retrieval for fact checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Haoyu Liang, Yinghui Wu, Yichao Zheng, and Caiming Xiong. 2023. Encouraging chain-of-verification for fact verification with large language models. In *arXiv preprint arXiv:2308.06792*.

Rishabh Misra. 2022. Politifact fact check dataset.

Jia Ying Ou, Uyen Trang Nguyen, and Tayzoon Ismail. 2022. Covmis: A dataset for research on covid-19 misinformation. In *2022 5th International Conference on Data Science and Information Technology (DSIT)*, pages 1–11.

Ofir Press, Lior Barak, Yuval Shoham, and Stuart M. Shieber. 2022. Measuring and narrowing the compositionality gap in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

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

Sandeep Singamsetty, Nishtha Madaan, Sameep Mehta, Varad Bhatnagar, and Pushpak Bhattacharyya. 2023. " beware of deception": Detecting half-truth and debunking it through controlled claim editing. *arXiv preprint arXiv:2308.07973*.

Alexander Spangher, Yixin Dang, Kateryna Tymoshenko, Sebastian Martschat, and Joel Tetreault. 2021. Multitask evidence retrieval for fact checking articles. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

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

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2025. FIRE: Fact-checking with iterative retrieval and verification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2901–2914, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhen Yang, Liangming Cao, Jing Zhang, Yue Zhang, and Yufang Fan. 2022. Cofced: Evidence distillation and refinement for fact verification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xin Xu Shaoji Zhang Ming Li Yao Shi, Hui Bu. 2015. Aishell-3: A multi-speaker mandarin tts corpus and the baselines.

Jiangyan Yi, Ye Bai, Jianhua Tao, Haoxin Ma, Zhengkun Tian, Chenglong Wang, Tao Wang, and Ruibo Fu. 2021. Half-truth: A partially fake audio detection dataset. *arXiv preprint arXiv:2104.03617*.

Takuma Yoneda, Jonathan Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Ucl machine reading group: Four factor framework for fact finding (fever). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*.

Yirong Zeng, Xiao Ding, Yi Zhao, Xiangyu Li, Jie Zhang, Chao Yao, Ting Liu, and Bing Qin. 2024. RU22Fact: Optimizing evidence for multilingual explainable fact-checking on Russia-Ukraine conflict. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14215–14226, Torino, Italia. ELRA and ICCL.

Wenxuan Zhang, Yang Deng, Jing Ma, and Wai Lam. 2020. AnswerFact: Fact checking in product question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2407–2417, Online. Association for Computational Linguistics.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method.