Claim Verification in the Age of Large Language Models: A Survey

Alphaeus Dmonte[1], Roland Oruche[2], Marcos Zampieri[1]
Prasad Calyam[2], Isabelle Augenstein[3]

[1]George Mason University, USA
[2]University of Missouri-Columbia, USA
[3]University of Copenhagen, Denmark

Abstract

The large and ever-increasing amount of data available on the Internet, coupled with the laborious task of manual claim and fact verification, has sparked interest in the development of automated claim verification systems.[1] Several deep learning and transformer-based models have been proposed for this task over the years. With the introduction of Large Language Models (LLMs) and their superior performance in several NLP tasks, we have seen a surge of LLM-based approaches to claim verification along with the use of novel methods such as Retrieval Augmented Generation (RAG). In this survey, we present a comprehensive account of recent claim verification frameworks using LLMs. We describe the different components of the claim verification pipeline used in these frameworks in detail, including common approaches to retrieval, prompting, and fine-tuning. Finally, we describe publicly available English datasets for this task.

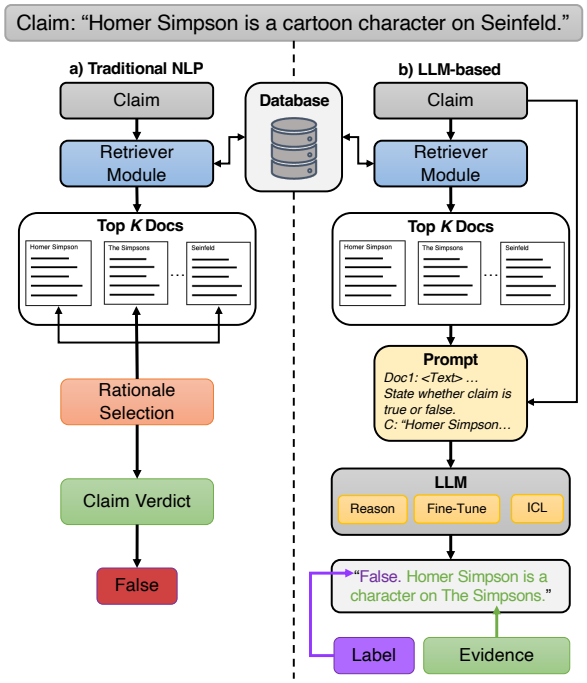

Figure 1: Comparison of claim verification systems between NLP-based (traditional) and LLM-based for claim veracity.

1 Introduction

False information is widely present on social media and on the Web, motivating the development of automated fact verification systems (Guo et al., 2022). The introduction of LLMs has provided malicious actors with sophisticated ways of creating and disseminating false information. Recent election cycles has seen a large number of claims spread across social media and news platforms alike (Dmonte et al., 2024). Similarly, during the COVID-19 pandemic, many factually inaccurate claims were spread (Zhou et al., 2023).

The use of computational models to verify the veracity of information is often defined as *fact verification* or *fact checking* and *claim verification*. While the two terms are often used interchangeably, an important distinction exist between *facts* and *claims*. A *claim* is an assertion with uncertain veracity that requires external evidence for verification, while a fact is an objectively true statement that can be verified through observations or reliable sources. Despite their differences, both are informational statements that can be verified.

Fact-checking potentially false claims is often applied to reduce the spread of misinformation. Organizations such as FactCheck, PolitiFact, NewsGuard, and Full Fact perform manual fact-checking to verify claims in different domains. However, this is a laborious task requiring domain expertise (Adair et al., 2017; Hanselowski, 2020; Warren et al., 2025), which has become increasingly infeasible due to the sheer volume of misinformation generated by humans and AI models. Automated fact-checking has become an increasingly popular

[1]We use the terms *claim verification* and *fact verification* interchangeably given the overlap between the two concepts.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 15–29

approach to verify the veracity of claims in a given text.[2] There are several steps involved in the fact-verification pipelines; the three main components are claim detection, evidence retrieval, and veracity prediction. Models used in these steps have followed the general methodological developments of the field and we have thus observed an increase in the use of LLMs for claim verification (Zhang and Gao, 2023; Wang and Shu, 2023; Quelle and Bovet, 2024). Figure 1 compares the architectural differences between traditional NLP-based vs LLM-based claim verification systems. The latter are less prone to error propagation, and provide additional justifications when verifying claims.

Despite this, LLMs are pre-trained on very large text collections and are prone to hallucinations, often generating texts containing incorrect information (Augenstein et al., 2024). Further, such models can be used to generate misinformation at scale (Chen and Shu, 2023; Zhou et al., 2023; Dmonte et al., 2024) and can therefore be exploited by malicious actors to spread factually incorrect information at an unprecedented rate (Pan et al., 2023c). Furthermore, using these pre-trained models for fact-verification may generate incorrect veracity labels, as the models may also rely on obsolete information. Approaches like RAG (Gao et al., 2023) are used for the model to retrieve the most recent information during fact verification.

A few general automated claim verification surveys have been published (Zeng et al., 2021; Bekoulis et al., 2021; Guo et al., 2022) as well as a couple of surveys on particular aspects of the task, such as explainability (Vallayil et al., 2023) and applications to specialized domains such as scientific texts (Vladika and Matthes, 2023). However, all past related surveys lack consideration for LLM-based approaches or focus on specific sub-tasks of the pipeline (Panchendrarajan and Zubiaga, 2024). In this paper, we fill this important gap by surveying LLM-based frameworks proposed in recent years. To the best of our knowledge, this is the first survey to explore claim verification with LLMs.

[2]We acknowledge that claim verification models can also be applied to other modalities of data (e.g., images). In this survey, we address models applied to text only.

2 Search Criteria

We search popular repositories[3] of scientific articles to collect the papers that serve as primary sources for this survey, using the terms *LLMs*, *claim verification*, *fact verification*, and related keywords. We focus primarily on the *ACL Anthology*, *ACM Digital Library*, *IEEE Xplore*, and proceedings of related conferences such as *AAAI* and *IJCAI*. We further search on *Scopus*, *Springer Link*, *Science@Direct*, and *ArXiv*.

Based on our search terms, we identify over 100 papers, and filter them as follows: (i) since the scope of our review is text-based LLM claim verification systems, we omit papers related to multimodal approaches (e.g., text-to-images), and papers investigating other modalities (e.g., images, graph-based systems); (ii) since we focus on LLM-based systems for claim veracity, we omit papers on claim identification, claim detection, and/or detecting LLM-generated content. This results in a total of 49 papers related to LLM-based approaches for veracity labeling. The papers on topic LLM-based veracity labeling have been published primarily in the ACL Anthology and ACM Digital Library, but also in other repositories such as the IEEE Xplore.

3 Claim Verification Pipeline

Figure 2 shows a typical claim verification pipeline consisting of: claim detection, claim matching, claim check-worthiness, document/evidence retrieval, rationale/sentence selection, veracity label prediction, and explanation/justification generation. Many systems only make use of some of these modules.

Claim Detection Input texts may contain one or more statements, not all of which are claims. Given the input text, a claim detection module is designed to identify all the statements containing a claim. For example, the statement *'I loved the movie Oppenheimer.'* is an opinion, whereas the statement *'The COVID-19 pandemic started in Texas.'* contains a claim.

Check-Worthy Claim Identification Not all identified claims are check-worthy. Given an input claim, the task is to identify the claims that include

[3]ACM: *portal.acm.org*. IEEE Xplore: *ieeexplore.ieee.org*. Scopus: *scopus.com*. ACL: *aclanthology.org*. Web of Science: *isiknowledge.com*. Springer: *link.springer.com*. ArXiv: *arxiv.org*. CEUR: *https://ceur-ws.org/*.

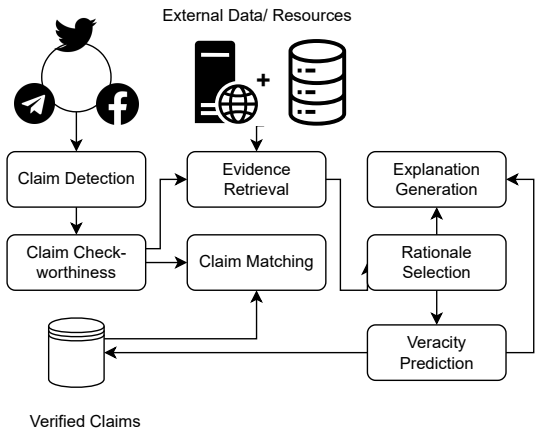

Figure 2: A typical claim verification pipeline

real-world assertions and that may need to be verified (Hassan et al., 2015; Nakov et al., 2021; Wright and Augenstein, 2020). Check-worthy claim identification is a subjective task, and it relies on factors like the popularity of the claim, public interest, to determine the claim's veracity etc. For example, the claim *'The President met the State Governor to discuss the infrastructure deal.'* is less check-worthy than the claim *'Drinking salt water cures COVID.'*

Claim Matching The identified check-worthy claims can be matched to previously fact-checked claims. Given an input claim and a database of previously fact-checked claims, claim matching is used to determine if the input claim is previously fact-checked and exists in the database (Shaar et al., 2020; Nakov et al., 2021). This can help predict the veracity label directly without additional steps.

Document/Evidence Retrieval If the input claim does not exist in the database of fact-checked claims, it needs to be verified. In the evidence retrieval sub-task, all relevant documents related to the input claim are extracted, either from an external database or through an internet search (Chen et al., 2017; Augenstein et al., 2019a). The threshold of the number of documents to be retrieved can be predetermined.

Rationale/Sentence Selection Not all information in the retrieved documents is relevant to the claim. Hence, for the rationale selection task, only the information or evidence most relevant to the claim is selected to predict the veracity label (Thorne et al., 2018a).

Veracity Label Prediction Once a rationale is selected, it is provided to a classifier along with the claim and potentially additional features, to predict a veracity label; often 'SUPPORTED', 'REFUTED', vs 'NOT ENOUGH EVIDENCE' (Thorne et al., 2018a). For some datasets, the labels can be 'TRUE' vs 'FALSE'.

Explanation/Justification Generation Recent works have focused on generating explanations for the veracity labels prediction. This specific task is focused on generating natural language justifications or explanations for the prediction, considering the claim and evidence.

4 LLM Approaches

Figure 3 shows an example pipeline that encapsulates multiple component modules (i.e., Evidence Retrieval, Prompt Creation, Transfer Learning, and LLM Generation) for verifying claims. Different from traditional fact verification pipelines that select evidence sets for verifying claims, LLM-based claim verification conditions generated text based on the concatenated input claim and retrieved evidence. Retrieval-augmented LLMs have been shown to perform well on knowledge-intensive tasks such as text generation.

4.1 Evidence Retrieval Strategies

RAG models, which have been developed to address hallucination in LLMs for knowledge-intensive tasks, have shown success for fact verification (Gao et al., 2023; Guan et al., 2024). Early work, such as (Lewis et al., 2020), develop a framework that retrieves evidence from an external database such as Wikipedia for conditionally generating veracity labels. Izacard et al., (2023) demonstrate that RAGs perform well on the FEVER shared task (Thorne et al., 2018b)) in few-shot settings, showing 5% improvement over large-scale LLMs such as Gopher (Rae et al., 2021) with significantly fewer parameters (i.e., 11B compared to 280B). Other works consider the optimization of either document ranking (Glass et al., 2022; Chen et al., 2022c) or input claims (i.e., queries) (Hang et al., 2024) as a crucial step for improving evidence retrieval for veracity labeling. Hofstätter et al., (2023) use an autoregressive re-ranker to get the most relevant passages from the retriever. These are then passed to the generation model to generate the veracity label.

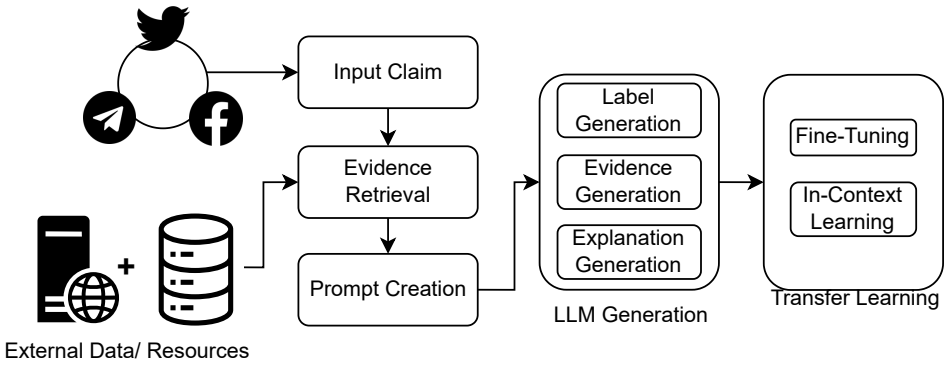

Figure 3: LLM-based claim verification pipeline. This involves creating a prompt from the retrieved evidence and the input claim as input to the LLM to generate a label, sentence evidence, and/or explanation of its response.

Despite this, RAG models often fail when encountering long or complex input claims, resulting in incorrectly generated veracity labels or evidence sentences (Hagström et al., 2025, 2026). Recent works have addressed this by segmenting long claims into smaller sub-claims and performing multiple rounds of retrieval. Khattab et al., (2021) present a pipeline for multi-hop claim verification that uses an iterative retriever and neural methods for effective document retrieval and re-ranking. Shao et al. (2023) show that using a re-ranker to distill knowledge to a retriever helps close the semantic gaps between a query and document passage when verifying claims using iterative RAGs. Other works address the issue of complex claims using fine-grained retrieval techniques based on claim decomposition. Ches et al. (2024) generate sub-questions based on a claim, which a document retriever uses to retrieve relevant documents. A fine-grained retriever retrieves top-k text spans as evidence based on a k-word window and BM25. Zhang and Gao (2023) decompose a claim into sub-claims and generate questions to verify the sub-claims. External knowledge sources are used to retrieve relevant information to verify the sub-claims and generate a final veracity label. Pan et al. (2023b) follow a programming paradigm, where the claim is broken down into subtasks and the final label is the aggregation of the execution of each subtask. The fact-verification subtask uses external knowledge to retrieve relevant evidence for a claim.

Hang et al. (2024) retrieve evidence based on generated knowledge graphs of evidences. They generate a knowledge graph of the user query or input claim and compare it to the database of knowledge graphs to retrieve the most relevant information for claim verification. Hu et al. (2023) propose a latent variable model that allows the retrieval of the most relevant evidence sentences from a document while removing the irrelevant sentences. This approach reduces noisy data during the verification process. Stochastic-RAG proposed by Zamani and Bendersky (2024) uses a stochastic sampling without replacement process for evidence retrieval and selection. This approach overcomes the ranking and selection of the evidence hence optimizing the RAG model. Zhang et al. (2023) optimize the evidence retrieval process by using feedback from the claim verifier. The divergence between the evidence from a retrieved evidence set provided to the verifier and the gold standard evidence, acts as a feedback signal used to train the retriever. Xu et al. (2024b) propose Search-in-the-Chain, where an LLM generates a reasoning chain and based on the answer to each node in the chain, the retrieval can be used to correct an answer or provide additional knowledge, improving the generation accuracy of an LLM.

4.2 Prompt Creation Strategies

Text prompting has shown to be effective for improving the output of LLMs. In the context of claim verification, several works investigate both manual and automated prompting strategies to increase robustness. Zhang and Gao (2023) develop a hierarchical prompting technique to verify multiple sub-claims using a step-by-step approach. Li et al. (2024a) demonstrate a self-sufficient claim verification through prompting instructions on multiple language models.

ProToCo (Zeng and Gao, 2023) demonstrates improved claim verification performance by leveraging a consistency mechanism to construct variants of the original claim-evidence pair prompt based on

three logical relations (i.e., confirmation, negation, uncertainty). Chen et al. (2023) develop a unified retrieval framework that employs discrete, continuous, and hybrid prompt strategies for adjusting to various knowledge-intensive tasks such as claim verification. The FactualityPrompts (Lee et al., 2022) framework tests the output generations of LLMs given an input prompt and uses an external database such as Wikipedia to calculate factuality and quality measures compared to the ground truth. Other works aim to improve an LLM's reasoning abilities by appending the claims with evidence during prompting for claim verification and text generation (Parvez, 2024; Dougrez-Lewis et al., 2024).

4.3 Transfer Learning Strategies

Fine-Tuning. Although recent studies show the success of pre-trained LLMs on zero- or few-shot tasks, they often fail at verifying real-world claims given their limited internal knowledge. The success of fine-tuning has motivated recent work on claim verification. Chen et al. (2022b) fine-tuned an LM on an external corpus for retrieving passage titles and evidence sentences using constrained beam search, finding improved performance on the FEVER dataset. Other work demonstrates that using GPT models to generate synthetic training data improves the performance of LLMs on fact checking (Tang et al., 2024) and claim matching (Choi and Ferrara, 2024).

Pan et al. (2021) develop a pipeline for creating a fact verification dataset and fine-tuning a language model by leveraging passages from Wikipedia to generate QA pairs related to claim veracity, showing improved zero-shot performance. Zeng and Zubiaga (2024) show that unlabelled pairwise data can increase the alignment between claim-evidence pairs, resulting in significant improvement on few-shot claim verification. Other recent works leverage reinforcement learning to fine-tune models for improving the veracity of claims and supporting evidence (Zhang and Gao, 2024; Huang et al., 2024). A document-level and question-level retrieval policy is proposed by Zhang and Gao (2024), where the top-k and top-1 documents for the document and question-level policy respectively, are used as input to a scoring function for label prediction during training. This approach outperforms retrieval and prompting approaches. Chiang et al. (2024) fine-tune LLMs for multi-stage fact verification. They fine-tune a model to generate answers based on claim-evidence pairs and a set of questions, whereas another model is fine-tuned to verify the claim based on the claim-evidence and question-answer pairs. Zhu et al. (2023) fine-tune a generation model to generate counterfactuals for out-of-domain classification.

In-Context Learning. The recent success of pre-trained LLMs in zero- and few-shot settings is largely attributed to its *in-context learning* (ICL) abilities (Kojima et al., 2022; Brown et al., 2020). For claim verification, popular ICL techniques include chain-of-thought (CoT) reasoning (Wei et al., 2022). Zhao et al. (2024) develop a multi-stage verification pipeline based on claim decomposition and self-reflection. An LLM-based verifier module is created using instruction prompting to generate a reasoning analysis among all sub-claims created by the decomposer module. They suggest that zero-shot prompting techniques result in better multi-hop performance on the HOVER and FEVEROUS datasets compared to few-shot prompting and fine-tuning. Kanaani (2024) enables LLMs to generate reasons over retrieved evidence in claim verification using few-shot ICL and the STaR CoT technique inspired by Zelikman et al. (2022). Similar work has leveraged CoT for effectively verifying complex claims using reasoning steps (Yao et al., 2023; Ni et al., 2024). Conversely, HiSS (Zhang and Gao, 2023) demonstrates that prompting for few-shot learning and claim decomposition can substantially improve the performance of CoT models for complex news claim verification. Li et al. (2023b) leverage the ICL capability of LLMs to perform multiple tasks simultaneously. Their approach outperforms or achieves comparable task performance in a zero-shot setting on claim verification datasets.

4.4 LLM Generation Strategies

Label and Evidence Generation. While the majority of claim verification systems predict veracity based on the concatenation of the input claim and evidence sentences (Pradeep et al., 2021), recent work has proposed alternate strategies for determining veracity labels and selecting/generating evidence pieces. Cao et al. (2024) develop SERIf, a claim verification pipeline that features an inference module to predict the veracity label of scientific news articles based on a two-step summarization (i.e., 'Extractive-Abstractive') and evidence retrieval technique. Each summary-evidence

pair is fed into the LLM and produces a binary label, indicating whether the news article is reliable (supported) or unreliable (refuted). Wadden et al. (2022b) leverage the Longformer model (Beltagy et al., 2020) that uses a shared encoding over the claim and document abstracts for rationale identification and claim label prediction.

Li et al. (2024b) propose to select the minimal evidence group within a set of retrieved candidate documents. This aims to minimize redundancy while selecting the most relevant evidence to prompt the language model. Chen et al. (2022b) use BART to encode all candidate sentences from the most relevant retrieved documents. In this, BART serves as an evidence decoder to predict the g-th evidence sentence via generation conditioned on the top k retrieved documents and the input claim. Lee et al. (2022) develop a variant of nucleus sampling called *factual-nucleus sampling*, in which the top-p sampling pool is selected as a set of sub-words whose cumulative probability exceeds p, resulting in improved evidence and label generation without claim verification datasets such as FEVER. Kao and Yen (2024) propose a multi-stage approach, where the evidence sentences are retrieved from articles related to a claim, and arguments are generated by aggregating and reconstructing the evidence. The arguments are refined and passed to an LLM to generate a verification label. Other approaches leverage LLMs reasoning capabilities to generate veracity labels and factual evidence (Cheng et al., 2024; Jafari and Allan, 2024; Li et al., 2024c; Fang et al., 2024; Pan et al., 2023a).

Explainable Generation. Recent studies investigate explainable approaches to improve LLM-based claim verification. Wang and Shu (2023) present the FOLK framework, that leverages the explanation capabilities of LLMs when verifying claims and justifies the prediction through a summary of its decision process. Dammu et al. (2024) proposes a knowledge graph (KG)-based approach for text verification and evidence attribution, where the LLM is fine-tuned on evidence attribution based on the input text and retrieved triplets from the KG, inducing explanations for claim predictions. While explanation techniques can help humans verify facts, LLMs can produce incorrect explanations due to hallucinations, making them unreliable in certain claim verification scenarios (Si et al., 2024; Warren et al., 2025, 2026). Ma et al. (2024) prompt an LLM in a few-shot setting to generate a concise summary of the evidence documents and input claim, serving as an explanation for the verified claim. Sun et al. (2026) propose CLUE, a framework that explains model uncertainty by grounding it in conflicts and agreements between claim–evidence spans, offering uncertainty-aware rationales that better reflect model reasoning. Other works leverage reasoning techniques such as chain-of-thought (CoT) to enable the LLM to be interpretable in its decision-making process when verifying claims (Yao et al., 2023; Pan et al., 2023a; Zhao et al., 2024; Kanaani et al., 2024; Ni et al., 2024; Quelle and Bovet, 2024; Fang et al., 2024). Pan et al. (2023a) and Fang et al. (2024) leverage an LLM's reasoning ability to generate explanations by using question-guided reasoning and minimizing the inherent model biases.

4.5 Agentic Approaches

AI agents using LLMs have been explored for several NLP applications, including claim verification. These systems often include multiple agents tasked to perform different stages of the verification pipeline (Li et al., 2024d; Rosenbaum et al., 2025; Trinh et al., 2025; Ning et al., 2025; Shukla et al., 2025; Fenza et al., 2025; Ma et al., 2025; Lin et al., 2025; Hong et al., 2025). Trinh et al. (2025) proposed a multi-agent system composed of four specialized agents, each concerned with different components of the claim verification pipeline, like claim decomposition, evidence, retrieval, and claim verification. Similarly, the authors in (Rosenbaum et al., 2025) introduced a system of three agents based on knowledge graph retrieval and web-lookup for fact verification. Shukla et al. (2025) and Ma et al. (2025) designed systems consisting of question-answering agents, while authors in (Fenza et al., 2025) used reasoning agents to perform various tasks of within the claim verification process.

Lin et al. (2025) introduced an approach that evaluates the fact-checking capabilities of LLMs and identifies their limitations. The system includes three agents that first generates a dynamic datasets based on a taxonomy followed by fact-verification and justification evaluation, and finally scrutinizing the capabilities of LLMs based on generated justifications to improve their verification capabilities. Apart from these approaches, other multi-agent systems that mimics human claim verification process have been proposed by (Li et al., 2024d) and (Hong et al., 2025). While most of these works have de-

signed systems for claim verification, Ning et al. (2025) proposed a multi-agent debate framework that evaluates the factuality of LLM generations.

5 Evaluation and Benchmarking

5.1 Metrics

The F1 score is the most commonly used metric to measure the performance of automatic claim verification systems, and to a lesser degree Precision, Recall, and Accuracy. Katranidis and Barany (2024) used the error rate between the human and automated fact-verification system to measure verification accuracy. However, these metrics consider a single pipeline component to evaluate the system's overall performance. Hence, Thorne et al. (2018a) introduce the FEVER score, a metric that incorporates both verification accuracy and evidence retrieval accuracy to compute overall system performance. While these metrics are valuable for assessing the performance of the classification tasks, they are inadequate for evaluating the performance of the non-classification components of the pipeline. Hence, metrics like Recall@k are used to measure the performance of the retrieval task (Pan et al., 2023b; Pradeep et al., 2021), while BLEU, METEOR, ROGUE, and BertScore are used to evaluate the quality of explanations or generated questions and answers. Additionally, some works have also evaluated the accuracy, faithfulness, and correctness of the generated explanations (Feher et al., 2025; Xing et al., 2025). Schlichtkrull et al. (2023) propose the new evaluation metric AVeriTeC score that uses METEOR and accuracy for question-answer-based veracity prediction systems. Other metrics like Mean Absolute Error (MAE), Expected Calibration Error (ECE), Area Under ROC Curve (AUC-ROC), and Pearson's Correlation are also used. Most of these metrics are inadequate for evaluating the factual accuracy of LLM-generated text, for which FactScore (Min et al., 2023), SAFE (Wei et al., 2024), and VERISCORE (Song et al., 2024) have been proposed instead. Other metrics and frameworks evaluate factual errors in generated text (Lee et al., 2022; Chern et al., 2023) and their alignment (Zha et al., 2023) and entailment (Lee et al., 2022) considering factuality.

5.2 Datasets

The fundamental resource for training and evaluating claim verification systems is datasets containing annotated texts. As most research in this area deals with English data, we collect information about publicly available English datasets used in the papers discussed in this survey.[4].

General-domain datasets have been created from online data sources and websites, including Wikipedia (Thorne et al., 2018a; Jiang et al., 2020; Diggelmann et al., 2020; Eisenschlos et al., 2021; Schuster et al., 2021; Kamoi et al., 2023), due to the extensive amount of information available spanning various topics and domains, and online fact-checking websites like PolitiFact (Wang, 2017; Augenstein et al., 2019b; Kao and Yen, 2024). Factually incorrect claims are shared through social media channels like X, Reddit, etc. Saakyan et al. (2021) introduce the COVID-Fact dataset consisting of claims extracted from Reddit posts. Several datasets for scientific fact verification have also been introduced (Wadden et al., 2020, 2022a; Lu et al., 2023). Most fact-verification datasets rely on unstructured textual evidence. Hence, a few datasets with structured evidence sources have been introduced (Aly et al., 2021; Lu et al., 2023). While many datasets focus only on veracity labels, some were developed to address explainability too (Chen et al., 2022d; Yang et al., 2022; Ma et al., 2024; Rani et al., 2023; Schlichtkrull et al., 2023).

Most claim-verification datasets include claims extracted from available information sources. LLMs are widely used to generate content, even though they can generate factually incorrect information. Efforts to identify such non-factual text have been undertaken. However, claim-verification datasets consisting of human-written claims can be inadequate for this task due to the linguistic discrepancy between human- and LLM-generated text. As such, LLM-generated claim verification datasets have been introduced (Li et al., 2024a; Cao et al., 2024). While these are used to evaluate automatic claim verification systems, there is a need to evaluate the factual accuracy of LLM-generated text, leading to the introduction of evaluation datasets (Lee et al., 2022; Wang et al., 2024; Malaviya et al., 2024).

5.3 Shared Tasks

Shared tasks are competitions where teams develop systems for tasks using a common benchmark dataset. Multiple shared tasks on claim and fact-checking have been organized, including the

[4] https://github.com/LanguageTechnologyLab/Claim-Verification-Papers.git

Fact Extraction and Verification (FEVER) challenge (Thorne et al., 2018b), TabFact (Wang et al., 2021) , CLEF 2020 CheckThat! (Barrón-Cedeno et al., 2020), SCIVER (Wadden and Lo, 2021), SEM-TAB-FACT (Wang et al., 2021), FACTIFY-5WQA (Suresh et al., 2024) , and more recently AVeriTeC (Schlichtkrull et al., 2024) . While there have been various techniques like question-answer generation as a precursor to label prediction as in AVeriTeC, all these shared tasks are centered on predicting veracity labels. Given an LLM's tendency to generate plausible yet factually inaccurate text, there is a need to organize shared tasks to evaluate the factual accuracy of LLM-generated content.

6 Open Challenges and Opportunities

LLM Biases Despite their strong reasoning capabilities, LLMs often exhibit implicit and explicit biases that can lead to harmful or inaccurate responses (Li et al., 2023c; Gallegos et al., 2024; Singhal et al., 2025). These biases arise at multiple stages of the claim verification cycle, including dataset construction, model architecture and reward design, and prompting or inference strategies. Prior work demonstrate that bias can propagate through LLM-based verification pipelines, where biased retrieval or parametric knowledge affects downstream reasoning and final veracity judgments (Iqbal et al., 2024; Gallegos et al., 2024; Bakke and Heggelund, 2025). Recent studies seek to address this issue using decomposition-based verification (Zhang and Gao, 2023; Wang and Shu, 2023; Huang et al., 2025), self-verification mechanisms (Weng et al., 2023; Kumar et al., 2025), and external frameworks such as "LLM-as-a-Judge" (Seo et al., 2025; Zheng et al., 2023). However, these methods remain limited in systematically identifying and mitigating bias across diverse claim types, highlighting the need for more robust claim verification frameworks.

Handling Irrelevant Context Retrieved evidence may be irrelevant, which is a challenge for LLMs, as they may not be trained to ignore such evidence. The lack of robustness to noise can lead an LLM to produce misinformation and incorrect verification. Recent research on open-domain question answering shows that external knowledge relevant to the task can improve model performance, however, irrelevant context can also lead to inaccurate predictions (Petroni et al., 2020; Shi et al., 2023; Li et al., 2023a; Yu et al., 2023). For the fact-verification, recent work has proposed techniques to identify the most relevant context for improving veracity label prediction, thus improving the overall system performance (Wang et al., 2023; Yoran et al., 2024; Xia et al., 2024). Hagström et al. (2025) further demonstrate that RAG systems frequently fail to exploit useful context and are easily misled by irrelevant information, highlighting persistent challenges in robust context utilisation across domains. However, more work is required to verify the effectiveness across domains.

Handling Knowledge Conflicts The reliance of fact-verification approaches on retrieved evidence can cause knowledge conflicts in LLMs, where retrieved external evidence may conflict with the internal parameters of the pre-trained LLM. This causes the LLM to ignore the retrieved evidence and produces hallucinations (Xu et al., 2024a). Approaches for avoiding knowledge conflicts have been introduced for question answering (Li et al., 2023a; Neeman et al., 2023; Mallen et al., 2023; Longpre et al., 2021; Chen et al., 2022a; Marjanović et al., 2024). For claim verification, recent work highlights that LLMs often struggle to reconcile conflicting evidence, leading to unreliable predictions (Warren et al., 2025, 2026), and that explicitly modelling evidence conflicts and agreements can improve robustness and interpretability (Sun et al., 2026). However, current approaches to addressing knowledge conflicts do not yet provide robust, generalisable solutions for LLM-based fact-verification across diverse domains and evidence conditions.

Multilinguality Most automated claim verification approaches rely on English datasets. Furthermore, there are limited multilingual fact-verification datasets (Gupta and Srikumar, 2021; Kazemi et al., 2022; Pikuliak et al., 2023; Singh et al., 2023; Resck et al., 2025). This hinders the development of approaches for multilingual fact-verification, which achieve the best performance when trained on language-specific datasets (Panchendrarajan and Zubiaga, 2024).

Community Notes Recently, platforms such as X and Facebook have introduced *community notes*. In this approach, platform users can add context to a post helping the community infer the veracity of information. According to platforms, this is intended to serve as a community-driven

moderation program as opposed to traditional moderation approaches that rely on moderators aided by computational approaches (X, 2025; Inc., 2025). Recent research has examined the epistemic and practical implications of such systems, showing how community-driven moderation reshapes fact-checking workflows and the distribution of trust online (Augenstein et al., 2025), and raising questions about whether community notes can effectively substitute for professional fact-checkers in practice (Borenstein et al., 2025). Future works is needed to better understand how community-generated signals can be reliably integrated with automated verification systems.

7 Conclusion

We presented a survey on LLM approaches to claim verification. To the best of our knowledge, this is the first claim verification survey to focus exclusively on LLM approaches, thus filling an important gap in the literature. We have described the traditional claim verification pipeline including its component tasks and discussed various LLM-based approaches used in this task. Finally, we have also described publicly available English datasets providing important information to new and seasoned researchers on this topic.

Advances in LLM development will likely continue to improve the quality of claim verification systems. We hope this survey motivates future research on this topic taking advantage of recently-proposed LLMs, RAG methods, and others. Claim verification is a vibrant research topic and we envisage multiple open research directions such as handling irrelevant context, knowledge conflicts and multilingualism.

Limitations

We have attempted to write the most comprehensive survey possible given the constraints of an 8-page paper. We carried out thorough searches through the ACL Anthology and similar repositories such as the ACM Digital Library. It is likely, however, that we may not have covered some relevant recent work.

Ethical Considerations

This survey reviews existing systems for automated claim verification and does not introduce any new models, frameworks, or datasets. Claim verification datasets often use data from news sources and social media and may include sensitive information as well as several societal biases. Such datasets that are used to create the automated verification models can embed the biases in the models, and this may affect the model evaluations and interpretations. Automated claim verification systems, especially systems that use LLMs, should therefore be used as decision-support tools rather than sources of truth, especially in a highly sensitive setting.

Acknowledgments

This research was co-funded by the European Union (ERC, ExplainYourself, 101077481), and supported by the Pioneer Centre for AI, DNRF grant number P1. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward "the holy grail": The continued quest to automate fact-checking. In *C+J*.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *NeurIPS*.

Isabelle Augenstein, Michiel Bakker, Tanmoy Chakraborty, David Corney, Emilio Ferrara, Iryna Gurevych, Scott Hale, Eduard Hovy, Heng Ji, Irene Larraz, Filippo Menczer, Preslav Nakov, Paolo Papotti, Dhruv Sahnan, Greta Warren, and Giovanni Zagni. 2025. Community Moderation and the New Epistemology of Fact Checking on Social Media.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David P. A. Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Eduard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram A. Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mac. Intell.*, 6.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019a. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *EMNLP-IJCNLP*.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019b. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *EMNLP-IJCNLP*.

Eivind Morris Bakke and Nora Winger Heggelund. 2025. (fact) check your bias. In *FEVER*.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, and 1 others. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *CLEF*.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Computing Surveys (CSUR)*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv*.

Nadav Borenstein, Greta Warren, Desmond Elliott, and Isabelle Augenstein. 2025. Can Community Notes Replace Professional Fact-Checkers? In *ACL*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Yupeng Cao, Aishwarya Muralidharan Nair, Elyon Eyimife, Nastaran Jamalipour Soofi, KP Subbalakshmi, John R Wullert II, Chumki Basu, and David Shallcross. 2024. Can large language models detect misinformation in scientific news reporting? *arXiv*.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *ACL*.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022a. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *EMNLP*.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2023. A unified generative retriever for knowledge-intensive language tasks via prompt learning. In *SIGIR*.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022b. GERE: generative evidence retrieval for fact verification. In *SIGIR*.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2022c. Corpusbrain: Pre-train a generative retrieval model for knowledge-intensive language tasks. In *CIKM*.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *NAACL*.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022d. Generating literal and implied subquestions to fact-check complex claims. In *EMNLP*.

Xiaoxia Cheng, Zeqi Tan, Wei Xue, and Weiming Lu. 2024. Information re-organization improves reasoning in large language models. In *NeurIPS*.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, and 1 others. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv*.

Shang-Hsuan Chiang, Ming-Chih Lo, Lin-Wei Chao, and Wen-Chih Peng. 2024. Team trifecta at factify5wqa: Setting the standard in fact verification with fine-tuning. *arXiv*.

Eun Cheol Choi and Emilio Ferrara. 2024. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. In *WWW*.

Preetam Prabhu Srikar Dammu, Himanshu Naidu, Mouly Dewan, YoungMin Kim, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. Claimver: Explainable claim-level verification and evidence attribution of text through knowledge graphs. *arXiv*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020*.

Alphaeus Dmonte, Marcos Zampieri, Kevin Lybarger, and Massimiliano Albanese. 2024. Classifying human-generated and ai-generated election claims in social media. In *SECRYPT*.

John Dougrez-Lewis, Mahmud Elahi Akhter, Yulan He, and Maria Liakata. 2024. Assessing the reasoning abilities of chatgpt in the context of claim verification. *arXiv*.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *NAACL*.

Yi Fang, Moxin Li, Wenjie Wang, Hui Lin, and Fuli Feng. 2024. Counterfactual debating with preset stances for hallucination elimination of llms. *arXiv*.

Darius Feher, Abdullah Khered, Hao Zhang, Riza Batista-Navarro, and Viktor Schlegel. 2025. Learning to generate and evaluate fact-checking explanations with transformers. *Engineering Applications of Artificial Intelligence*.

Giuseppe Fenza, Domenico Furno, Vincenzo Loia, and Pio Pasquale Trotta. 2025. Multi-llm agents architecture for claim verification. In *ITASEC*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *ACL*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *NAACL*.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *NAACL*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *TACL*, 10.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *ACL-IJCNLP*.

Lovisa Hagström, Youna Kim, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, and Isabelle Augenstein. 2026. CUB: Benchmarking Context Utilisation Techniques for Language Models. In *ACL*.

Lovisa Hagström, Sara Vera Marjanovic, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2025. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In *ACL*.

Ching Nam Hang, Pei-Duo Yu, and Chee Wei Tan. 2024. Trumorgpt: Query optimization and semantic reasoning over networks for automated fact-checking. In *CISS*.

Andreas Hanselowski. 2020. *A machine-learning-based pipeline approach to automated fact-checking*. Ph.D. thesis, Technical University of Darmstadt.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *CIKM*.

Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2023. Fid-light: Efficient and effective retrieval-augmented text generation. In *SIGIR*.

Spencer Hong, Meng Luo, and Xinyi Wan. 2025. EMULATE: A multi-agent framework for determining the veracity of atomic claims by emulating human actions. In *FEVER*.

Xuming Hu, Junzhe Chen, Zhijiang Guo, and Philip S Yu. 2023. Give me more details: Improving fact-checking with latent retrieval. *arXiv*.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024. Training language models to generate text with citations via fine-grained rewards. *arXiv*.

Yani Huang, Richong Zhang, Zhijie Nie, Junfan Chen, and Xuefeng Zhang. 2025. A graph-based verification framework for fact-checking. *arXiv*.

Meta Platforms Inc. 2025. Community notes: A new way to add context to posts. `https://transparency.meta.com/features/community-notes`.

Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Nenkov Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. OpenFactCheck: A unified framework for factuality evaluation of LLMs. In *EMNLP*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*.

Nazanin Jafari and James Allan. 2024. Robust claim verification through fact detection. *arXiv*.

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 *EMNLP Findings*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *EMNLP*.

Mohammadamin Kanaani, Sajjad Dadkhah, and Ali A. Ghorbani. 2024. Triple-R: Automatic reasoning for fact verification using language models. In *LREC-COLING*.

Wei-Yu Kao and An-Zi Yen. 2024. MAGIC: Multi-argument generation with self-refinement for domain generalization in automatic fact-checking. In *LREC-COLING 2024*.

Vasileios Katranidis and Gabor Barany. 2024. Faaf: Facts as a function for the evaluation of rag systems. *arXiv*.

A Kazemi, Z Li, V Peréz-Rosas, SA Hale, and R Mihalcea. 2022. Matching tweets with applicable fact-checks across languages. In *CEUR Workshop*.

Omar Khattab, Christopher Potts, and Matei A. Zaharia. 2021. Baleen: Robust multi-hop reasoning at scale via condensed retrieval. In *NeurIPS*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Adarsh Kumar, Hwiyoon Kim, Jawahar Sai Nathani, and Neil Roy. 2025. Improving the reliability of llms: Combining cot, rag, self-consistency, and self-verification. *arXiv*.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *NeurIPS*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *ACL Findings*.

Jiazheng Li, Runcong Zhao, Yongxin Yang, Yulan He, and Lin Gui. 2023b. Overprompt: Enhancing chatGPT through efficient in-context learning. In *R0-FoMo*.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024a. Self-checker: Plug-and-play modules for fact-checking with large language models. In *NAACL Findings*.

Xiangci Li, Sihao Chen, Rajvi Kapadia, Jessica Ouyang, and Fan Zhang. 2024b. Minimal evidence group identification for claim verification. *arXiv*.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024c. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *ICLR*.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024d. Large language model agentic approach to fact checking and fake news detection. In *ECAI*.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023c. A survey on fairness in large language models. *CoRR*.

Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See Kiong Ng, and Tat-Seng Chua. 2025. Fact-audit: An adaptive multi-agent framework for dynamic fact-checking evaluation of large language models. In *ACL*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *EMNLP*.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *EMNLP*.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, and Shu Wu. 2024. Ex-fever: A dataset for multi-hop explainable fact verification. In *ACL Findings*.

Jiatong Ma, Linmei Hu, Rang Li, and Wenbo Fu. 2025. Local: Logical and causal fact-checking with llm-based multi-agents. In *ACM Web Conference*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *NAACL*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *ACL)*.

Sara Vera Marjanović, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. Dynamicqa: Tracing internal knowledge conflicts in language models. In *EMNLP*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, and 1 others. 2021. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *ECIRl*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *ACL*.

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. Afacta: Assisting the annotation of factual claim detection with reliable llm annotators. *arXiv*.

Yucheng Ning, Xixun Lin, Fang Fang, and Yanan Cao. 2025. Mad-fact: A multi-agent debate framework for long-form factuality evaluation in llms. *arXiv*.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *ACL-IJCNLP*.

Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. 2023a. QACheck: A demonstration system for question-guided multi-hop fact-checking. In *EMNLP: System Demonstrations*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023b. Fact-checking complex claims with program-guided reasoning. In *ACL*.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023c. On the risk of misinformation pollution with large language models. In *ACL Findings*.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*.

Md Rizwan Parvez. 2024. Evidence to generate (e2g): A single-agent two-step prompting for context grounded and retrieval augmented reasoning. *arXiv*.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *AKBC*.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *EMNLP*.

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

Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, and 1 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv*.

Anku Rani, S.M Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY-5WQA: 5W aspect-based fact verification through question answering. In *ACL*.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and Interpretability of Multilingual Large Language Models: A Survey. In *EMNLP*.

Richard Rosenbaum, Timo Cavelius, Lasse Strothe, Andrii Lata, Jana Diesner, and 1 others. 2025. Hybrid fact-checking that integrates knowledge graphs, large language models, and search-based retrieval agents improves interpretable claim verification. In *9th Widening NLP Workshop*.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *ACL-IJCNLP*.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, and 1 others. 2024. The automated verification of textual claims (averitec) shared task. In *FEVER*.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *NeurIPS*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *NAACL*.

Wooseok Seo, Seungju Han, Jaehun Jung, Benjamin Newman, Seungwon Lim, Seungbeen Lee, Ximing Lu, Yejin Choi, and Youngjae Yu. 2025. Verifying the verifiers: Unveiling pitfalls and potentials in fact verifiers. *arXiv*.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *ACL*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *EMNLP Findings*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *ICML*, volume 202.

Satyam Shukla, Himanshu Dutta, and Pushpak Bhattacharyya. 2025. Recon, answer, verify: Agents in search of truth. In *EMNLP: Industry Track*.

Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. In *NAACL*.

Iknoor Singh, Carolina Scarton, Xingyi Song, and Kalina Bontcheva. 2023. Finding already debunked narratives via multistage retrieval: Enabling cross-lingual, cross-dataset and zero-shot learning. *arXiv*.

Aryan Singhal, Veronica Shao, Gary Sun, Ryan Ding, Jonathan Lu, and Kevin Zhu. 2025. A comparative study of translation bias and accuracy in multilingual large language models for cross-language claim verification. In *NeurIPS*.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. *arXiv*.

Jingyi Sun, Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2026. Explaining Sources of Uncertainty in Automated Fact-Checking. In *ACL*.

Suryavardan Suresh, Anku Rani, Parth Patwa, Aishwarya Reganti, Vinija Jain, Aman Chadha, Amitava Das, Amit Sheth, and Asif Ekbal. 2024. Overview of factify5wqa: Fact verification through 5w question-answering. *arXiv*, abs/2410.04236.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *FEVER*.

Tam Trinh, Manh Nguyen, and Truong-Son Hy. 2025. Towards robust fact-checking: A multi-agent system with advanced evidence retrieval. *arXiv*.

Manju Vallayil, Parma Nand, Wei Qi Yan, and Héctor Allende-Cid. 2023. Explainability of automated fact verification systems: A comprehensive review. *Applied Sciences*.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *ACL Findings*.

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 *EMNLP*.

David Wadden and Kyle Lo. 2021. Overview and insights from the SCIVER shared task on scientific claim verification. In *Second Workshop on Scholarly Document Processing*.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. SciFact-open: Towards open-domain scientific claim verification. In *EMNLP Findings*.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *NAACL Findings*.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *EMNLP Findings*.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *SemEval*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *EMNLP Findings*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *arXiv*.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking. In *CHI*.

Greta Warren, Jingyi Sun, Irina Shklovski, and Isabelle Augenstein. 2026. Show me the evidence: Evaluating the role of evidence and natural language explanations in ai-supported fact-checking. In *ACM FAccT*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

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. In *NeurIPS*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *EMNLP Findings*.

Dustin Wright and Isabelle Augenstein. 2020. Claim check-worthiness detection as positive unlabelled learning. In *EMNLP Findings*.

X. 2025. Community notes: a collaborative way to add helpful context to posts and keep people better informed. `https://communitynotes.x.com/guide/en/about/introduction`.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Jun Chen, and Haifeng Huang. 2024. Improving retrieval augmented language model with self-reasoning. *arXiv*.

Rui Xing, Timothy Baldwin, and Jey Han Lau. 2025. Evaluating evidence attribution in generated fact checking explanations. In *NAACL*.

Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024a. Knowledge conflicts for llms: A survey. *arXiv*.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024b. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In *WWW*.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *COLING*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *ICLR*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *ICLR*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv*.

Hamed Zamani and Michael Bendersky. 2024. Stochastic RAG: end-to-end retrieval-augmented generation through expected utility maximization. In *SIGIR*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*.

Fengzhu Zeng and Wei Gao. 2023. Prompt to be consistent is better than self-consistent? few-shot and zero-shot fact verification with pre-trained language models. In *ACL Findings*.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*.

Xia Zeng and Arkaitz Zubiaga. 2024. MAPLE: Micro analysis of pairwise language evolution for few-shot claim verification. In *EACL Findings*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *ACL*.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. From relevance to utility: Evidence retrieval with feedback for fact verification. In *EMNLP Findings*.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *AACL-IJCNLP*.

Xuan Zhang and Wei Gao. 2024. Reinforcement retrieval leveraging fine-grained feedback for fact checking news claims with black-box LLM. In *LREC-COLING 2024*.

Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. 2024. PACAR: Automated fact-checking with planning and customized action reasoning using large language models. In *LREC-COLING 2024*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G. Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *CHI*.

Yingjie Zhu, Jiasheng Si, Yibo Zhao, Haiyang Zhu, Deyu Zhou, and Yulan He. 2023. EXPLAIN, EDIT, GENERATE: Rationale-sensitive counterfactual data augmentation for multi-hop fact verification. In *EMNLP*.