Explainable Automated Fact-Checking: An Overview

Hoang Long Nguyen[1,2,*], Frédéric Rayar[1,*] and Nicolas Ragot[1]

[1]Université de Tours, LIFAT (UR6300), Tours, 37200, France
[2]Hong Duc University, Thanh Hoa, 440000, Vietnam

Abstract

Recently, with the rapid development of AI-generated content, misinformation and disinformation have been spreading widely across various communication platforms, especially social networking services. This surge in misleading information has amplified the demand for robust and reliable automated fact-checking (AFC) systems. Although numerous approaches have been proposed, these systems have yet to gain users' trust, as they often provide predictions without accompanying explanations, limiting the transparency of these approaches. To address this challenge, explainable AFC has emerged as an essential research direction that aims to provide the explainability and trustworthiness of the AFC approaches. In this study, we present a comprehensive overview of explainable AFC. We first introduce conventional explainable AFC methods and discuss recent approaches that leverage Large Language Models and Vision–Language Models. Then, we review state-of-the-art datasets and evaluation metrics to provide a broader understanding of the available resources and methodologies. Finally, we highlight the limitations of current studies and provide research directions to support explainable AFC.

Keywords

Automated Fact-Checking, xAI, Explainability, Multimodal, Datasets

1. Introduction

The rapid spread of multimodal misinformation and disinformation, especially AI-generated content, across social media without moderation poses serious challenges to digital information reliability [1]. Disinformation refers to intentionally manipulated content created to mislead, raising concerns about its potential to shape public opinion, influence elections, and erode trust in institutions and scientific evidence.

This challenge requires the process to moderate the information in the media era, called fact-checking, to reduce the spread of disinformation. In the early stages, this work was carried out by journalists. However, with the rapid expansion of the Internet and the widespread adoption of smart devices, the volume of information has grown so large that manual fact-checking is no longer feasible [2]. To address this problem, researchers have proposed automated fact-checking (AFC) systems that leverage machine learning and deep learning techniques. AFC has become an increasingly prominent approach for verifying claims and generating explanations to support the verification process. A conventional automated fact-checking pipeline typically consists of three main stages: (i) claim detection and extraction, (ii) evidence retrieval, and (iii) verification. The verification stage can be further divided into subtasks such as out-of-context classification, manipulated content classification, and veracity classification, which are often followed by the generation of explanations. Consequently, there has been a growing effort toward developing efficient and transparent approaches for explainable automated fact-checking systems.

Nevertheless, existing explainable automated fact-checking surveys have several drawbacks such as: (i) a lack of discussion on LLM- and VLM-based approaches, and (ii) limited coverage of evaluation metrics for generated explanations.

The *ROMCIR 2026 – The 6th Workshop on Reducing Online Misinformation through Credible Information Retrieval* [3] provides an opportunity to address these limitations by advancing information access

ROMCIR 2026: The 6th Workshop on Reducing Online Misinformation through Credible Information Retrieval (held as part of ECIR 2026: The 48th European Conference on Information Retrieval). April 2, 2026. Delft, The Netherlands.
*Corresponding author.
hnguyen@univ-tours.fr (H. L. Nguyen); frederic.rayar@univ-tours.fr (F. Rayar); nicolas.ragot@univ-tours.fr (N. Ragot)
0009-0003-2327-4178 (H. L. Nguyen); 0000-0003-1927-8400 (F. Rayar); 0000-0003-2321-942X (N. Ragot)


systems designed to mitigate both human- and AI-generated information disorder. The workshop particularly emphasizes on fact-checking methodologies and the assessment of factual accuracy and reliability across multimodality, while also tackling AI-related challenges such as explainability, hallucination detection, and the evaluation of automatically generated information.

In this paper, we present a comprehensive survey on explainable automated fact-checking that synthesizes both prior studies and recent advances employing LLMs and VLMs. We also outline the commonly used evaluation metrics for this task, which give additional values to our study in comparison with previous one(s) (see Table 1).

Surveys	Technique					Multimodal	Dataset	Metric
	Attention-based	Rule-based	Summarization-based	LLM-based	VLM-based			
Kotonya et al. [4]	✓	✓	✓	–	–	–	✓	–
Ours	✓	✓	✓	✓	✓	✓	✓	✓

Table 1
Comparison between existing explainable fact-checking survey and ours.

We conduct a systematic search of major scientific repositories to identify the primary studies included in this survey. Specifically, we query the ACL Anthology, ACM Digital Library, IEEE Xplore, SpringerLink, ScienceDirect, arXiv, and Google Scholar. The search is performed using combinations of keywords such as automated fact-checking, minsinformation detection, claim verification, fact-checking datasets, explainable fact-checking, explanation generation, and related terms.

The initial search yields more than 180 candidate papers. We then apply filtering criteria aligned with the scope of this survey. In particular, a study is retained only if it substantially addresses the intersection between automated fact-checking and explainability or interpretability techniques, either by explicitly generating natural-language rationales, highlighting supporting evidence, or exposing intermediate reasoning processes.

The selected papers are subsequently categorized by building upon the taxonomy proposed by Kotonya et al. [4], while extending it to incorporate recent LLM- and VLM-based approaches. Finally, we identify and analyze the state-of-the-art datasets and evaluation metrics adopted in the selected studies, which form the empirical foundation of this survey.

The remainder of this paper is structured as follows: Section 2 presents conventional approaches as well as recent methods based on LLMs and VLMs. Section 3 describes widely-used datasets and evaluation metrics for explainable automated fact-checking. Section 4 discusses current limitations and outlines key research directions, and Section 5 concludes the paper.

2. Explainable Automated Fact-Checking

In this section, we present an overview of studies on explainable automated fact-checking, focusing on two main categories of approaches: conventional methods and those based on LLMs and VLMs. Figure 1 illustrates the differences in the explanation-generation procedures between these two paradigms.

The transparency of an automated fact-checking system could be split into two main directions: interpretability and explainability [4]. Interpretability refers to how easily an expert/researcher can understand a model's internal structure and decision process, for example in linear models or decision trees where the contribution of each feature is explicit, meanwhile, explainability refers to the capacity of providing understandable rationales by natural languages for the predictions in the verification stage.

2.1. Conventional approaches

Prior studies introduced various specific task formulations for generating these explanations, which are summarized in the Table. 2 below. Analyzing multimodal data is particularly essential, especially in non-English contexts, where users may rely significantly on multimodal cues and where cultural background

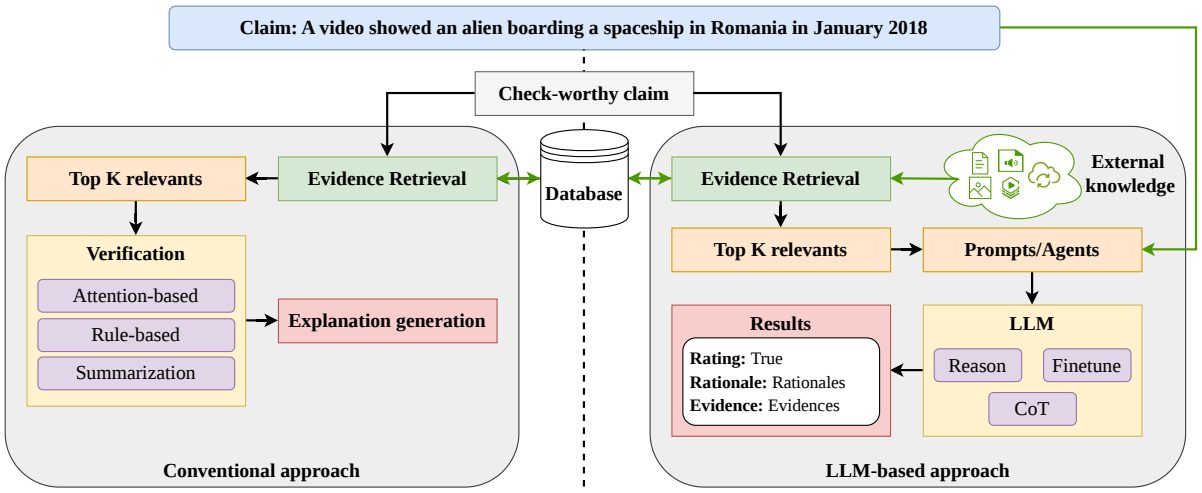

Figure 1: Comparison between conventional and LLM-based explanable automated fact-checking.

knowledge shapes the interpretation of visual, textual, and audio content, thereby influencing both claim understanding and verification outcomes.

Conventional explainable fact-checking could be categorized into attention-based highlighting, rule-based, and summarization techniques.

Task	Method	Year	Technique	Dataset	Lang.	Modality			
						Text	Image	Audio	Video
Attention-based explanations	DeClarE [5]	2018	BiLSTM + attention	Snopes[1], PolitiFact[2], NewsTrust [6], RumourEval [7]	En	✓	–	–	–
	dEFEND [8]	2019	RNN + GRU + attention	FakeNewsNet [9]	En	✓	–	–	–
	XFake [10]	2019	CNN + attention + XGBoost	PolitiFact[2]	En	✓	–	–	–
	GCAN [11]	2020	CNN + RNN + graph convolution	Twitter (2015) [12], Twitter (2016) [13]	En	✓	–	–	–
	DTCA [14]	2020	Decision tree + attention	RumourEval [7], PHEME [15]	En	✓	–	–	–
Rule-based explanations	ExFaKT [16]	2019	KG + Horn rules	YAGO [17], DBpedia [18]	En	✓	–	–	–
	Ahmadi et al. [19]	2019	KG + logical rules	DBpedia [18]	En	✓	–	–	–
Explanation as summarisation	Liu et al. [20]	2019	BERT-based	CNN/DailyMail [21], NYT [22], XSum [23]	En	✓	–	–	–
	Atanasova et al. [24]	2020	BERT-based	LIAR [25]	En	✓	–	–	–
	Kotonya & Toni [26]	2020	BERT-based	PUBHEALTH [26]	En	✓	–	–	–
	Biased TextRank [27]	2021	Graph-based	LIAR-PLUS [28], HNR [27]	En	✓	–	–	–

Table 2
Overview of representative conventional explainable fact-checking approaches.

2.1.1. Attention-based Explanation

Attention-based explanations primarily target the verification stage by highlighting input tokens, sentences, or evidence fragments that most strongly influence the model's veracity prediction. DeClarE [5] introduced an end-to-end neural network model built on a biLSTM architecture that incorporates external supporting or refuting textual evidence from the Web, along with linguistic style cues and assessments of source trustworthiness. The explanations are then presented as tokens highlighted in the articles, guided by attention weights. dEFEND [8] combined RNN and GRU encoders to jointly model news articles and user comments from the FakeNewsNet dataset. It employs a hierarchical co-attention

[1]https://www.snopes.com/
[2]https://www.politifact.com/

mechanism that highlights both article sentences and comments that strongly support or contradict the article. These highlighted spans offer interpretable signals about where the model finds evidence for or against the article's credibility. XFake [10] targets news articles with rich attribute information and generates both veracity predictions and explanations through three sub-frameworks built using GloVe, Word2Vec, CNNs, attention mechanisms, and XGBoost. GCAN [11] extends attention-based explanation to social graphs by constructing graphs from tweets, users, and their interactions, and applying co-attention mechanisms over both the textual content and the graph structure. Wu et al. [14] proposed the Decision Tree–based Co-Attention (DTCA) model, which integrates a decision-tree structure with co-attention networks to capture deeper semantic interactions between evidence and claims.

2.1.2. Rule-based Explanation

Rule-based explanations focus on the reasoning process itself by exposing symbolic inference chains, logical rules, or knowledge-graph paths that connect a claim to supporting or refuting evidence. ExFaKT [16] advances this direction by deriving more human-understandable evidence through background-knowledge rules in the form of Horn clauses, a restricted form of first-order logic rules consisting of a conjunction of positive literals implying a single head literal. These rules are used to gather supporting information from both knowledge graphs and textual sources. Ahmadi et al. [19] enriched knowledge-graph information by employing logical rule discovery modeled as an inference problem using probabilistic answer set programming, which allows the system to effectively integrate and reason with both uncertain rules and uncertain facts extracted from the Web.

2.1.3. Explanation as Summarization

Summarization-based explanations formulate explanation generation as a post-verification task, where the goal is to produce a natural-language justification that summarizes the evidence supporting the predicted claim label. This line of work primarily leverages BERT-based architectures. Summarization models trained on datasets such as CNN/DailyMail, NYT, and XSum have been adapted to generate natural-language explanations for fact-checking decisions. Atanasova et al. [24] use BERT-based encoders and decoders to generate explanations on LIAR-style data, while Kotonya and Toni [26] introduce PUBHEALTH, where each public-health claim is paired with both a veracity label and an expert-written justification, and train models to mimic these rationales. Biased TextRank [27] introduced an unsupervised graph-based algorithm employed for generating extractive natural language explanations.

These conventional approaches mainly rely on existing knowledge bases, previously fact-checked claims, and human-annotated examples, which limits their ability to handle unseen claims or incorporate newly emerging knowledge [29]. Besides, these studies mainly focus on text-only, limiting their capacities in leveraging multimodal content which provide broader contexts for explainable fact-checking. Consequently, researchers have begun to explore more advanced techniques such as Large Language Models (LLMs) and Vision Language Models (VLMs) for explainable automated fact-checking to overcome these limitations.

2.2. LLM-based approaches

With the rapid development of large language models (LLMs), numerous studies explore how to integrate them into automated fact-checking pipelines and generate explanations for verdict predictions. In contrast to conventional architectures, LLM-based approaches can directly produce free-form natural language rationales, handle a wide range of topics without task-specific feature engineering, and be adapted to new domains through prompting or fine-tuning as well as retrieving new knowledge.

In this survey, we group LLM-based methods according to how the model is integrated into the fact-checking pipeline: (i) prompting-based approaches; (ii) retrieval-augmented generation (RAG) approaches; and (iii) fine-tuning approaches. These approaches will be detailed in the following sections.

2.2.1. Prompting-based approaches

Prompting-based approaches use pretrained LLMs in a zero- or few-shot manner. Given a claim and optionally some retrieved evidence, the model is instructed through a carefully designed prompt to output both a veracity label and a natural language rationale. In this approach, the form and style of explanations are mainly controlled by the prompt design. Several recent studies adopt this paradigm to explore the capabilities of LLMs for explainable fact-checking demonstrating the capacities of zero- and few-shot prompts while producing natural language justifications [30, 31].

However, prompting-based approaches face several challenges when applied to more complex automated fact-checking tasks. Explanations can be highly sensitive to prompt structures, ordering of examples, and the context provided, which leads to instability and makes it difficult to guarantee faithfulness [32, 33]. Moreover, long or compositional claims often require multi-step reasoning that is not explicitly represented in standard prompt formats, limiting the transparency of the underlying decision process.

To better handle such complex cases, some works extend prompting with explicit claim decomposition, where the original claim is first broken down into simpler sub-claims that are then verified separately and aggregated into a final verdict [34, 35, 36, 37]. Chain of Thoughts (CoT), on the other hand, leverages LLMs' reasoning abilities by encouraging the model to present intermediate reasoning steps before producing the final decision, making part of the decision process explicit in the generated rationale [38, 39, 40].

Although these approaches work well in controlled research settings, they are usually evaluated on limited datasets and often show weaker performance on real-world content, while also requiring high computational cost for LLMs [34, 40].

2.2.2. RAG approaches

With the increasing of information in the media ecosystem, retrieval-augmented approaches explicitly combine LLMs with external knowledge sources to generate predictions and explanations at inference time. Instead of relying solely on pretrained knowledge, these systems first retrieve multimodal data such as articles, documents, or web pages for a given claim, and then condition the LLM on both the claim and the retrieved evidence to produce a verdict and a natural language rationale [41, 42]. To further enhance the performance of the RAG, [43] introduces Chain of RAG (CoRAG) and Tree of RAG (ToRAG) reasoning techniques retrieving both textual and visual representations. LEAF [44] proposes leveraging the fact-checking to guide retrieval. In parallel, PACAR [45] integrates RAG with multi-agent LLMs that dynamically select reasoning tools and external evidence to tackle diverse multi-hop verification tasks.

This approach provides a better understanding and transparency since users are able to inspect the underlying documents and evidence and verify them. However, the drawbacks of this approach comes from the problem of the RAG methods such as handling retrieval errors, noisy or conflicting evidence, and increased computational cost due to longer contexts and multi-step interactions [44] and a persistent gap between scientific research and real-world applications [42].

2.2.3. Fine-tuning approaches

In another line of work, LLMs are adapted to specific tasks and application scenarios, leading to a demand for fine-tuning in order to achieve better performance and more tailored behavior [46, 47, 48, 49]. However, these studies mainly focus on single-task fact-checking, and may struggle to handle combined settings where the model is expected both to predict veracity labels and to generate corresponding rationales [50]. Subsequent work introduces multi-task fine-tuning, with a focus on exploiting knowledge transfer across related tasks [51].

Fine-tuning can enhance performance of LLMs in the fact-checking, but it is usually more practical for smaller LLMs due to the computational costs. In contrast, larger LLMs already handle multi-tasks, therefore additional fine-tuning is often both computationally expensive and not strictly needed.

2.2.4. Agent-based approaches

Furthermore, LLMs also could be treated as multi-agents. Multi-agent approaches use several cooperating or competing LLMs that assume different roles in the fact-checking process, such as debaters, critics, judges, or planners. Primary studies such as [40, 45] leverage LLMs for complex claim verification, structured planning, and reasoning tasks. MADR [52] proposed to apply multiple LLM agents with different roles in an iterative refinement process to improve the faithfulness of generated explanations. This process reduces unfaithful content and aligns the final output more closely with the underlying evidence. Furthermore, MAD-Fact [53] introduce a multi-agent debate system for factual verification that reduces single-model bias and improves reasoning reliability through structured interactions among three modules including clerk, jury, and judge. Ma et al. [54] introduce a knowledgeable debate mechanism that improves multi-agent efficiency by incorporating external knowledge.

Recent work also integrates structured knowledge sources, advocating the combination of LLMs with knowledge graphs to incorporate structured factual information [55, 56, 57].

Although these approaches open a promising direction for explainable fact-checking, they introduce substantial deployment challenges, including high computational cost and large token usage due to multi-agent interactions [53]. Moreover, uncertainty introduced at early debate stages may propagate through subsequent agent interactions, affecting the stability and reliability of final explanations [58].

2.3. VLM-based approaches

While most explainable AFC systems still operate primarily on English text, misinformation in the wild is often conveyed through multimodal data, especially images and videos, and explanations should ideally align evidence across modalities. Vision–Language Models (VLMs) extend LLM-based approaches by jointly modeling textual and visual inputs, enabling systems to reason about the consistency between claims and multimodal contents and to generate explanations that explicitly refer to visual cues.

In the image–text setting, Yao et al. [59] introduce MOCHEG, an end-to-end multimodal fact-checking and explanation benchmark where each claim is paired with ruling statements and both textual and visual evidence gathered from fact-checking websites. Their baseline models combine visual and textual encoders to perform multimodal evidence retrieval and claim verification, while also generating natural language rationales that summarise the supporting or refuting evidence. These explanations are grounded not only in retrieved documents but also in associated images, encouraging models to capture cross-modal inconsistencies such as mismatched locations, dates, or entities.

For short-video fact-checking task, Niu et al. [60] propose TRUE, a dataset specifically designed for explainable video fact-checking, together with 3MFact, a multi-role multimodal model. Given a video and an accompanying claim or caption, 3MFact jointly processes video frames, audio, and textual metadata to determine whether the video supports or contradicts the claim, and generates summarized rationales that highlight which temporal segments and modalities are most informative. TRUE thus provides fine-grained annotations about the roles of different modalities, making it possible to evaluate whether VLMs attend to the right visual or audio evidence when explaining their decisions.

Qi et al. [61] introduce SNIFFER, a multimodal large language model for explainable out-of-context misinformation detection. SNIFFER builds on InstructBLIP and applies a two-stage instruction-tuning procedure: the first stage aligns the model's understanding of generic objects in images with news-domain entities, while the second stage fine-tunes the model using out-of-context–specific instructions generated by GPT-4. The system produces both an OOC prediction and a textual explanation that indicates why the accompanying media does or does not match the claim context, thereby illustrating how instruction-tuned VLMs can support explainable AFC beyond purely textual inputs.

Despite these advances, explainable fact-checking in multimodal and non-English contexts remains in its early stages. Existing VLM-based approaches are often evaluated on relatively small benchmarks, and rely heavily on English-language fact-checking sources. There is a need for larger and more diverse multimodal datasets with fine-grained explanation annotations, better methods for grounding explanations in specific visual or temporal regions, and systematic studies of how multimodal explanations

affect user trust and decision-making in real-world fact-checking systems.

3. Datasets and Evaluation Metrics

3.1. Datasets

Recently, automated fact-checking has benefited from a growing number of datasets that include not only single-modal data but also multimodal content. Table 3 presents the main resources considered in this survey.

At the single-modal level, LIARPLUS [28] primarily extends the LIAR dataset with human-written justifications extracted from Politifact[3]. PUBHEALTH [26] targets the public-health domain, providing veracity labels accompanied by concise expert-written explanations. More recently, Ma et al. [62] introduced EX-FEVER, a large-scale multi-hop benchmark comprising around 60,000 claims that require two- or three-hop reasoning. ChartCheck, on the other hand, collects real-world charts consisting of 1,683 images along with 10,480 claims.

With the rapid development of automated fact-checking methods, a new line of multimodal datasets has emerged to support richer forms of evidence and explanation. MOCHEG [59] includes 15,601 annotated claims along with 33,880 textual evidence paragraphs and 12,112 images from fact-checking websites to support subtasks such as multimodal evidence retrieval, claim verification, and explanation generation.

The growing prevalence of misinformation in short videos has spurred further efforts to construct multimodal datasets that jointly incorporate video, audio, text, and social context. VMH [63] focuses on misleading videos related to the 2016 U.S. presidential election on the Meta platform, containing 2,247 annotated articles. FakeSV [64] offers the largest Chinese short video dataset leveraging different perspectives ranging from video, audio to social reactions, and publisher profiles. TRUE [60] represents a recent effort specifically aimed at explainable video fact-checking containing 1,097 true videos and 1,828 false videos, with a particular emphasis on summarized rationales.

Dataset	Year	Modality				No. classes	Claims	Lang.	Source
		Text	Image	Audio	Video				
LIARPLUS [28]	2018	✓	–	–	–	6	12,836	En	Extended LIAR
PUBHEALTH [26]	2020	✓	–	–	–	4	11,832	En	FC webs
AVeriTeC [65]	2023	✓	–	–	–	3	4,568	En	FC organizations
EX-FEVER [62]	2024	✓	–	–	–	3	>60,000	En	Wikipedia
ChartCheck [66]	2024	–	✓	–	–	4	10,480	En	Wikimedia
MOCHEG [59]	2022	✓	✓	–	–	3	15,601	En	FC webs
FakeSV [64]	2023	✓	–	✓	✓	4	3,654	Zh	TikTok/Kuai
VMH [63]	2023	✓	–	✓	✓	2	2,247	En	Meta
TRUE [60]	2025	✓	–	✓	✓	2	2,925	En	Snopes

Table 3
Overview of automated fact-checking datasets.

While these datasets have substantially advanced AFC across textual and multimodal settings, they remain largely English-centric and focus solely on truthfulness labels without providing the underlying reasoning, which directly impacts how explanation methods can be trained and evaluated.

3.2. Evaluation Metrics

In this section, we present several well-known metrics for explainable automated fact-checking task, including similarity-based and LLM-based metrics. One of the key steps in explanation is producing the

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

prediction for the claim in which several popular metrics are computed, such as accuracy, precision, recall, and F_1-score [67]. To evaluate the rationale, similarity-based metrics refer to conventional natural language processing (NLP) such as BLEU [68], ROUGE-n [69], METEOR [70], CIDEr [71], and BERTScore [72] are applied. While LLM-based metrics evaluate the performance of LLM prompts, including G-Eval [73]. Each metric is detailed below:

- BLEU measures the overlap between the generated and reference texts by using n-gram precision. BLEU is computed as Equation 1 follows:

$$\text{BLEU} = BP \cdot \exp\Big(\sum_{n=1}^{N} w_n \log p_n\Big) \tag{1}$$

  where BP is the brevity penalty, w_n are n-gram weights, and p_n are modified n-gram precisions.
- ROUGE similarly computes the n-gram recall overlap between the generated text and the reference text. Mathematically, it can be depicted as in Equation 2.

$$\text{ROUGE-n} = \frac{\text{overlap}_n(\text{gen}, \text{ref})}{\text{overlap}_n(\text{ref})} \tag{2}$$

- METEOR is an automatic metric originally designed for machine translation evaluation but now also used to assess other forms of generated text, and it is based on a generalized concept of unigram matching between the generated and reference text.

$$\text{METEOR} = F_\alpha \times (1 - Penalty) \tag{3}$$

  where F_α is a weighted harmonic mean of precision and recall, and $Penalty$ is the penalty derived from how fragmented the matched word sequences are.
- CIDEr is a metric designed to measure how well a generated sentence matches human-written sentences, focusing on consensus among multiple human references.

$$\text{CIDEr} = \frac{1}{4}\sum_{n=1}^{4} cosine(TFIDF_{gen}, TFIDF_{ref}) \tag{4}$$

- BERTScore leverages contextual embeddings from pretrained BERT to compute semantic similarity between generated and reference texts, making it more robust to paraphrasing and word-order variations. The BERTScore could be depicted as an F1_{score} as in Equation. 5:

$$BERTScore = 2 \times \frac{P_{BERT} \times R_{BERT}}{P_{BERT} + R_{BERT}} \tag{5}$$

- G-Eval is a reference-free evaluation framework that treats an LLM as a judge of generated text quality. Instead of comparing system outputs to human references with n-gram metrics, it asks the LLM to analyze the explanations step by step and then output a structured score.

However, these metrics are hardly aligned with explanation purposes. Metrics such as accuracy and F1-score measure whether the system predicts the correct label, but they do not show whether the explanation truly supports that decision. Similarity-based metrics like BLEU, ROUGE, and BERTScore mainly examine how close the generated explanation is to a reference text, focusing on wording rather than reasoning. LLM-based metrics such as G-Eval assess coherence and informativeness, but they still do not guarantee that the explanation is grounded in the actual evidence. As a matter of fact, standard classification and similarity-based metrics often miss important aspects such as factual consistency, robustness, and user impact, and many of them were originally developed for other tasks rather than claim–evidence verification. In addition, faithfulness metrics and LLM-based evaluators, although promising, are not yet fully reliable and may introduce their own biases. Therefore, these metrics cover

only part of explainability and cannot ensure that explanations reflect the real reasoning behind the verdict.

Besides these metrics, researchers and citizens are increasingly focusing on human-centered evaluations, which can offer a clearer and more meaningful fact-checking context. Human-centered evaluations typically consider two main aspects: the quality of the explanation itself and the contribution of that explanation to users' overall experience with the AI system [74]. The quality of the explanation concerns how easy it is for users to understand and interpret the provided rationale, while the contribution aspect focuses on how the explanation supports human–AI interaction, helps users better understand how the system works, and affects their perception of its performance [74].

While existing automatic and LLM-based metrics provide useful proxies for evaluating verification performance and explanation quality, they fail to fully capture key aspects such as faithfulness, evidence grounding, and human usefulness. These limitations in current evaluation practices directly motivate several of the open challenges and research directions discussed in Section 4, particularly the need for human-centered and process-oriented evaluation frameworks.

4. Research directions

4.1. Limitations of automated/explainable fact-checking

Despite rapid progress, current automated fact-checking (AFC) and explainable AFC (xAFC) systems still exhibit several limitations that restrict their usefulness in practice.

A key limitation is the predominantly English-centric orientation of the field. Many datasets, benchmarks, and pretrained models are built with English as primary language and sources, often reducing the performance and quality of the explanation for non-English languages, local contexts, and underrepresented accents [29].

Beyond language coverage, most approaches still focus on text-only or bi-modal settings, typically image–text, despite the increasing prevalence of misinformation conveyed through audio, short videos, and complex multimodal contents. While recent approaches and datasets begin to address multimodal fact-checking, robust end-to-end explanation methods for these modalities remain underexplored, and evaluation protocols for multimodal explanations are still in their early stages.

In addition, evaluation practices in AFC and xAFC often fail to align with end-user needs, as they typically rely on technical metrics that are difficult for end-users to interpret, such as accuracy, ROUGE/BLEU, or similarity-based scores. This leads to a misalignment between what researchers optimize and what fact-checkers and citizens actually need, including clear uncertainty estimates, inspectable evidence, and actionable next steps. As a result, system improvements do not necessarily translate into greater usability or increased trust [75, 76].

Finally, although LLM- and VLM-based methods show promise for automated fact-checking, they still exhibit significant limitations that are particularly problematic in fact-checking contexts. These include hallucinated statements, data-driven biases, and potential censorship, where outputs can appear fluent yet remain unsupported or systematically skewed [77]. For instance, an LLM may produce a convincing explanation while relying on weak or irrelevant evidence, or may fail disproportionately on region-specific claims due to insufficient coverage in its training data.

4.2. Research perspectives for AFC and xAFC

Addressing the above limitations requires research directions that explicitly prioritize faithfulness, transparency, and real-world usability, rather than purely benchmark-driven performance.

Future studies should develop evaluation protocols that go beyond accuracy or surface-level NLP similarity metrics toward hybrid assessments combining (i) correctness of the verdict, (ii) evidence grounding and attribution quality, and (iii) human-centered utility, such as whether explanations improve users' trust and decision-making. In practice, this should be achieved by considering user

studies involving fact-checkers and citizens, as well as conceiving structured ensemble strategies (e.g., multi-agent voting or consensus) to reduce single-model brittleness when evidence is ambiguous.

Furthermore, xAFC systems should emphasize process-level explanations, enabling end-users to understand the reasoning path that leads from a claim to the final verdict and explanation, rather than receiving only a final label or a short natural-language rationale [29]. For example, systems may expose retrieval queries, evidence ranking criteria, and intermediate sub-claims verified during the decision process.

In parallel, future systems should improve model transparency by clearly documenting training data sources, language and domain coverage, and known failure modes. Such documentation is critical to reduce distrust and to help practitioners anticipate biases or missing-coverage issues [29].

Moreover, progress in xAFC will depend on stronger real-world validation through sustained collaboration with professional fact-checkers on realistic cases and workflows. This includes handling incomplete or conflicting evidence and supporting abstention when verification is not possible [75].

Finally, for LLM- and VLM-based pipelines, hallucination and bias mitigation should be treated as core design requirements. Promising directions include constraining explanations to retrieved evidence, adding verification steps for generated statements, and monitoring systematic errors across languages and regions [77].

These future research perspectives encourage a shift from benchmark-oriented optimization toward reliable, auditable, and inclusive xAFC systems that align with the practical requirements of both fact-checkers and citizens.

5. Conclusion

In this survey, we have reviewed the emerging landscape of explainable automated fact-checking, with a particular focus on methods that explicitly expose the evidence and reasoning behind their verdicts. We first outlined the general AFC pipeline and clarified the role of interpretability and explainability within it, before categorising existing approaches into conventional models and more recent LLM- and VLM-based methods. We then discussed datasets that provide ground-truth explanations in textual and multimodal settings, and summarised commonly used automatic metrics and LLM-based evaluators for assessing explanation quality.

Our overview highlights that, despite rapid progress, explainable AFC is still in its early stages. Conventional approaches offer more controlled and often more faithful mechanisms for surfacing evidence, but they typically require task-specific engineering and are limited in their ability to generate rich, human-oriented rationales. LLM- and VLM-based methods, in contrast, are highly flexible and can natively handle complex, multimodal inputs, yet they raise serious concerns about faithfulness, hallucination, and robustness. Current datasets and evaluation protocols only partially capture these dimensions, especially for multimodal and multilingual scenarios.

Future perspectives should emphasize on faithful and robust explanations, multimodal and multilingual coverage, and human-centred evaluation and deployment. Progress in these directions will require closer collaboration between the researchers and professional fact-checkers, as well as high-quality annotated resources. By developing AFC systems that are not only about verification, but also clearly communicate why, the field can contribute more effectively to supporting fact-checkers and helping end-users navigate increasingly complex information ecosystems.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT in order to: Translation, grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policymaking, volume 27, Council of Europe Strasbourg, 2017.

[2] M. Akhtar, M. Schlichtkrull, Z. Guo, O. Cocarascu, E. Simperl, A. Vlachos, Multimodal automated fact-checking: A survey, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023, pp. 5430–5448.

[3] M. Fernández-Pichel, M. Petrocchi, K. Roitero, M. Viviani, Romcir 2026: Overview of the 6th workshop on reducing online misinformation through credible information retrieval, in: European Conference on Information Retrieval, Springer, 2026.

[4] N. Kotonya, F. Toni, Explainable automated fact-checking: A survey, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2020, pp. 5430–5443.

[5] K. Popat, S. Mukherjee, A. Yates, G. Weikum, DeClarE: Debunking fake news and false claims using evidence-aware deep learning, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 22–32.

[6] S. Mukherjee, G. Weikum, Leveraging joint interactions for credibility analysis in news communities, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, Association for Computing Machinery, 2015, p. 353–362.

[7] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, A. Zubiaga, SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, 2017, pp. 69–76.

[8] K. Shu, L. Cui, S. Wang, D. Lee, H. Liu, defend: Explainable fake news detection, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, 2019, p. 395–405.

[9] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media, Big data 8 (2020) 171–188.

[10] F. Yang, S. K. Pentyala, S. Mohseni, M. Du, H. Yuan, R. Linder, E. D. Ragan, S. Ji, X. Hu, Xfake: Explainable fake news detector with visualizations, in: The world wide web conference, 2019, pp. 3600–3604.

[11] Y.-J. Lu, C.-T. Li, GCAN: Graph-aware co-attention networks for explainable fake news detection on social media, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 505–514.

[12] C. Boididou, K. Andreadou, S. Papadopoulos, D. T. Dang Nguyen, G. Boato, M. Riegler, Y. Kompatsiaris, et al., Verifying multimedia use at mediaeval 2015, in: MediaEval 2015, volume 1436, CEUR-WS, 2015.

[13] C. Boididou, S. Papadopoulos, D. T. Dang Nguyen, G. Boato, M. Riegler, A. Petlund, I. Kompatsiaris, Verifying multimedia use at mediaeval 2016, 2016.

[14] L. Wu, Y. Rao, Y. Zhao, H. Liang, A. Nazir, DTCA: Decision tree-based co-attention networks for explainable claim verification, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 1024–1035.

[15] A. Zubiaga, M. Liakata, R. Procter, Exploiting context for rumour detection in social media, in: International conference on social informatics, Springer, 2017, pp. 109–123.

[16] M. H. Gad-Elrab, D. Stepanova, J. Urbani, G. Weikum, Exfakt: A framework for explaining facts over knowledge graphs and text, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, 2019, p. 87–95.

[17] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: Proceedings of the 16th International Conference on World Wide Web, Association for Computing Machinery, 2007, p. 697–706.

[18] P. Shiralkar, A. Flammini, F. Menczer, G. L. Ciampaglia, Finding streams in knowledge graphs

to support fact checking, in: 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 2017, pp. 859–864.
[19] N. Ahmadi, P. Papotti, M. Saeed, Explainable fact checking with probabilistic answer set programming, in: TTO 2019, Conference for truth and trust Online, 4-5 October 2019, London, UK, 2019.
[20] Y. Liu, M. Lapata, Text summarization with pretrained encoders, 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), Association for Computational Linguistics, 2019, pp. 3730–3740.
[21] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, Advances in neural information processing systems 28 (2015).
[22] E. Sandhaus, The New York Times Annotated Corpus, 2008.
[23] S. Narayan, S. B. Cohen, M. Lapata, Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 1797–1807.
[24] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, Generating fact checking explanations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 7352–7364.
[25] W. Y. Wang, "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), Association for Computational Linguistics, 2017, pp. 422–426.
[26] N. Kotonya, F. Toni, Explainable automated fact-checking for public health claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 7740–7754.
[27] A. Kazemi, Z. Li, V. Pérez-Rosas, R. Mihalcea, Extractive and abstractive explanations for fact-checking and evaluation of news, in: Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, Association for Computational Linguistics, 2021, pp. 45–50.
[28] T. Alhindi, S. Petridis, S. Muresan, Where is your evidence: Improving fact-checking by justification modeling, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, 2018, pp. 85–90.
[29] G. Warren, I. Shklovski, I. Augenstein, Show me the work: Fact-checkers' requirements for explainable automated fact-checking, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–21.
[30] M. G. Mohammadkhani, A. G. Mohammadkhani, H. Beigy, Zero-shot learning and key points are all you need for automated fact-checking, in: Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), Association for Computational Linguistics, 2024, pp. 86–90.
[31] D. Pisarevskaya, A. Zubiaga, Zero-shot and few-shot learning with instruction-following LLMs for claim matching in automated fact-checking, in: Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, 2025, pp. 9721–9736.
[32] R. Kamoi, S. S. S. Das, R. Lou, J. J. Ahn, Y. Zhao, X. Lu, N. Zhang, Y. Zhang, R. H. Zhang, S. R. Vummanthala, et al., Evaluating llms at detecting errors in llm responses, arXiv preprint arXiv:2404.03602 (2024).
[33] J. A. Leite, O. Razuvayevskaya, K. Bontcheva, C. Scarton, Weakly supervised veracity classification with llm-predicted credibility signals, EPJ Data Science 14 (2025) 16.
[34] L. Pan, X. Wu, X. Lu, A. T. Luu, W. Y. Wang, M.-Y. Kan, P. Nakov, Fact-checking complex claims with program-guided reasoning, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2023, pp. 6981–7004.
[35] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, FActScore: Fine-grained atomic evaluation of factual precision in long form text generation,

in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, pp. 12076–12100.

[36] Y. Lu, N. Ziems, H. Dang, M. Jiang, Optimizing decomposition for optimal claim verification, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2025, pp. 5095–5114.

[37] Q. Hu, Q. Long, W. Wang, Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance?, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, 2025, pp. 6313–6336.

[38] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, in: The eleventh international conference on learning representations, 2022.

[39] X. Li, R. Zhao, Y. K. Chia, B. Ding, S. Joty, S. Poria, L. Bing, Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources, in: ICLR, 2024.

[40] X. Zhang, W. Gao, Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method, 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), Association for Computational Linguistics, 2023, pp. 996–1011.

[41] J. Chen, G. Kim, A. Sriram, G. Durrett, E. Choi, Complex claim verification with evidence retrieved in the wild, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, 2024, pp. 3569–3587.

[42] X. Zhang, W. Gao, Reinforcement retrieval leveraging fine-grained feedback for fact checking news claims with black-box LLM, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, 2024, pp. 13861–13873.

[43] M. A. Khaliq, P. Y.-C. Chang, M. Ma, B. Pflugfelder, F. Miletić, RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models, in: Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), Association for Computational Linguistics, 2024, pp. 280–296.

[44] H. Tran, J. Wang, Y. Ting, H. Yu, W. Huang, T. Chen, LEAF: Learning and evaluation augmented by fact-checking to improve factualness in large language models, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, 2025, pp. 338–363.

[45] X. Zhao, L. Wang, Z. Wang, H. Cheng, R. Zhang, K.-F. Wong, PACAR: Automated fact-checking with planning and customized action reasoning using large language models, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, 2024, pp. 12564–12573.

[46] S. Bhatia, J. H. Lau, T. Baldwin, Automatic claim review for climate science via explanation generation, arXiv preprint arXiv:2107.14740 (2021).

[47] M. Hyben, S. Kula, I. Srba, R. Moro, J. Simko, Multilingual and multi-topical benchmark of fine-tuned language models and large language models for check-worthy claim detection, arXiv preprint arXiv:2311.06121 (2023).

[48] S. Althabiti, M. A. Alsalka, E. Atwell, Generative ai for explainable automated fact checking on the factex: A new benchmark dataset, in: Multidisciplinary International Symposium on Disinformation in Open Online Media, Springer, 2023, pp. 1–13.

[49] X. Zeng, A. Zubiaga, MAPLE: Micro analysis of pairwise language evolution for few-shot claim verification, in: Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, 2024, pp. 1177–1196.

[50] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, Y. Zhang, An empirical study of catastrophic forgetting in large language models during continual fine-tuning, IEEE Transactions on Audio, Speech and

Language Processing (2025).
[51] M. Du, S. D. Gollapalli, S.-K. Ng, Nus-ids at checkthat! 2022: Identifying check-worthiness of tweets using checkthat5., in: CLEF (Working Notes), 2022, pp. 468–477.
[52] K. Kim, S. Lee, K.-H. Huang, H. P. Chan, M. Li, H. Ji, Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate, arXiv preprint arXiv:2402.07401 (2024).
[53] Y. Ning, X. Lin, F. Fang, Y. Cao, Mad-fact: A multi-agent debate framework for long-form factuality evaluation in llms, arXiv preprint arXiv:2510.22967 (2025).
[54] X. Ma, G. Rao, L. Xu, X. Wang, Z. Fan, Z. Zhang, Guided and knowledgeable multi-agent debate for fact verification, Expert Systems with Applications (2025) 130103.
[55] N. Giarelis, C. Mastrokostas, N. Karacapilidis, A unified LLM-KG framework to assist fact-checking in public deliberation, in: Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE) @ LREC-COLING 2024, ELRA and ICCL, 2024, pp. 13–19.
[56] B. Ghosh, S. Hasan, N. A. Arafat, A. Khan, Logical consistency of large language models in fact-checking, arXiv preprint arXiv:2412.16100 (2024).
[57] X. Jing, S. Billa, D. Godbout, On a scale from 1 to 5: Quantifying hallucination in faithfulness evaluation, in: Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, 2025, pp. 7765–7780.
[58] L. Yoffe, A. Amayuelas, W. Y. Wang, DebUnc: Improving large language model agent communication with uncertainty metrics, in: Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, 2025, pp. 23299–23315.
[59] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, L. Huang, End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2733–2743.
[60] K. Niu, D. Xu, B. Yang, W. Liu, Z. Wang, Pioneering explainable video fact-checking with a new dataset and multi-role multimodal model approach, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 28276–28283.
[61] P. Qi, Z. Yan, W. Hsu, M. L. Lee, Sniffer: Multimodal large language model for explainable out-of-context misinformation detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 13052–13062.
[62] H. Ma, W. Xu, Y. Wei, L. Chen, L. Wang, Q. Liu, S. Wu, L. Wang, EX-FEVER: A dataset for multi-hop explainable fact verification, in: Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, 2024, pp. 9340–9353.
[63] Y. Y. Sung, J. Boyd-Graber, N. Hassan, Not all fake news is written: A dataset and analysis of misleading video headlines, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, pp. 16241–16258.
[64] P. Qi, Y. Bu, J. Cao, W. Ji, R. Shui, J. Xiao, D. Wang, T.-S. Chua, Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 14444–14452.
[65] M. Schlichtkrull, Z. Guo, A. Vlachos, Averitec: A dataset for real-world claim verification with evidence from the web, Advances in Neural Information Processing Systems 36 (2023) 65128–65167.
[66] M. Akhtar, N. Subedi, V. Gupta, S. Tahmasebi, O. Cocarascu, E. Simperl, Chartcheck: Explainable fact-checking over real-world chart images, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 13921–13937.
[67] S. Tufchi, A. Yadav, T. Ahmed, A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities, International Journal of Multimedia Information Retrieval 12 (2023) 28.
[68] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 311–318.
[69] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization

Branches Out, Association for Computational Linguistics, 2004, pp. 74–81.
[70] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, 2005, pp. 65–72.
[71] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.
[72] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020.
[73] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: NLG evaluation using gpt-4 with better human alignment, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, pp. 2511–2522.
[74] J. Kim, H. Maathuis, D. Sent, Human-centered evaluation of explainable ai applications: a systematic review, Frontiers in Artificial Intelligence 7 (2024) 1456486.
[75] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for human-ai interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2019, p. 1–13.
[76] R. Tomsett, A. Preece, D. Braines, F. Cerutti, S. Chakraborty, M. Srivastava, G. Pearson, L. Kaplan, Rapid trust calibration through interpretable and uncertainty-aware ai, Patterns 1 (2020).
[77] S. S. Rahman, M. A. Islam, M. M. Alam, M. Zeba, M. A. Rahman, S. S. Chowa, M. A. K. Raiaan, S. Azam, Hallucination to truth: A review of fact-checking and factuality evaluation in large language models, Artificial Intelligence Review (2025).